

Volume 8 Issue 8

August 2017



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 8 Issue 8 August 2017

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

- **Aamir Shaikh**
- **Abbas Al-Ghaili**
Mendeley
- **Abbas Karimi**
Islamic Azad University Arak Branch
- **Abdelghni Lakehal**
Université Abdelmalek Essaadi Faculté
Polydisciplinaire de Larache Route de Rabat, Km 2 -
Larache BP. 745 - Larache 92004. Maroc.
- **Abdul Razak**
- **Abdul Karim ABED**
- **Abdur Rashid Khan**
Gomal University
- **Abeer ELkorany**
Faculty of computers and information, Cairo
- **ADEMOLA ADESINA**
University of the Western Cape
- **Aderemi A. Atayero**
Covenant University
- **Adi Maaita**
ISRA UNIVERSITY
- **Adnan Ahmad**
- **Adrian Branga**
Department of Mathematics and Informatics,
Lucian Blaga University of Sibiu
- **agana Becejski-Vujaklija**
University of Belgrade, Faculty of organizational
- **Ahmad Saifan**
yarmouk university
- **Ahmed Boutejdar**
- **Ahmed AL-Jumaily**
Ahlia University
- **Ahmed Nabih Zaki Rashed**
Menoufia University
- **Ajantha Herath**
Stockton University Galloway
- **Akbar Hossain**
- **Akram Belghith**
University Of California, San Diego
- **Albert S**
Kongu Engineering College
- **Alcinia Zita Sampaio**
Technical University of Lisbon
- **Alexane Bouënard**
Sensopia
- **ALI ALWAN**
International Islamic University Malaysia
- **Ali Ismail Awad**
Luleå University of Technology
- **Alicia Valdez**
- **Amin Shaqrah**
Taibah University
- **Amirrudin Kamsin**
- **Amitava Biswas**
Cisco Systems
- **Anand Nayyar**
KCL Institute of Management and Technology,
Jalandhar
- **Andi Wahyu Rahardjo Emanuel**
Maranatha Christian University
- **Anews Samraj**
Mahendra Engineering College
- **Anirban Sarkar**
National Institute of Technology, Durgapur
- **Anthony Isizoh**
Nnamdi Azikiwe University, Awka, Nigeria
- **Antonio Formisano**
University of Naples Federico II
- **Anuj Gupta**
IKG Punjab Technical University
- **Anuranjan misra**
Bhagwant Institute of Technology, Ghaziabad, India
- **Appasami Govindasamy**
- **Arash Habibi Lashkari**
University Technology Malaysia(UTM)
- **Aree Mohammed**
Directorate of IT/ University of Sulaimani
- **ARINDAM SARKAR**
University of Kalyani, DST INSPIRE Fellow
- **Aris Skander**
Constantine 1 University
- **Ashok Matani**
Government College of Engg, Amravati
- **Ashraf Owis**
Cairo University
- **Asoke Nath**

St. Xaviers College(Autonomous), 30 Park Street,
Kolkata-700 016

- **Athanasios Koutras**
- **Ayad Ismaeel**
Department of Information Systems Engineering-
Technical Engineering College-Erbil Polytechnic
University, Erbil-Kurdistan Region- IRAQ
- **Ayman Shehata**
Department of Mathematics, Faculty of Science,
Assiut University, Assiut 71516, Egypt.
- **Ayman EL-SAYED**
Computer Science and Eng. Dept., Faculty of
Electronic Engineering, Menofia University
- **Babatunde Opeoluwa Akinkunmi**
University of Ibadan
- **Bae Bossoufi**
University of Liege
- **BALAMURUGAN RAJAMANICKAM**
Anna university
- **Balasubramanie Palanisamy**
- **BASANT VERMA**
RAJEEV GANDHI MEMORIAL COLLEGE, HYDERABAD
- **Basil Hamed**
Islamic University of Gaza
- **Basil Hamed**
Islamic University of Gaza
- **Bhanu Prasad Pinnamaneni**
Rajalakshmi Engineering College; Matrix Vision
GmbH
- **Bharti Waman Gawali**
Department of Computer Science & information T
- **Bilian Song**
LinkedIn
- **Binod Kumar**
JSPM's Jayawant Technical Campus, Pune, India
- **Bogdan Belean**
- **Bohumil Brtnik**
University of Pardubice, Department of Electrical
Engineering
- **Bouchaib CHERRADI**
CRMEF
- **Brahim Raouyane**
FSAC
- **Branko Karan**
- **Bright Keswani**
Department of Computer Applications, Suresh Gyan
Vihar University, Jaipur (Rajasthan) INDIA
- **Brij Gupta**

University of New Brunswick

- **C Venkateswarlu Sonagiri**
JNTU
- **Chanashekhhar Meshram**
Chhattisgarh Swami Vivekananda Technical
University
- **Chao Wang**
- **Chao-Tung Yang**
Department of Computer Science, Tunghai
University
- **Charlie Obimbo**
University of Guelph
- **Chee Hon Lew**
- **Chien-Peng Ho**
Information and Communications Research
Laboratories, Industrial Technology Research
Institute of Taiwan
- **Chun-Kit (Ben) Ngan**
The Pennsylvania State University
- **Ciprian Dobre**
University Politehnica of Bucharest
- **Constantin POPESCU**
Department of Mathematics and Computer
Science, University of Oradea
- **Constantin Filote**
Stefan cel Mare University of Suceava
- **CORNELIA AURORA Gyorödi**
University of Oradea
- **Cosmina Ivan**
- **Cristina Turcu**
- **Dana PETCU**
West University of Timisoara
- **Daniel Albuquerque**
- **Dariusz Jakóbczak**
Technical University of Koszalin
- **Deepak Garg**
Thapar University
- **Devena Prasad**
- **DHAYA R**
- **Dheyaa Kadhim**
University of Baghdad
- **Djilali IDOUGH**
University A.. Mira of Bejaia
- **Dong-Han Ham**
Chonnam National University
- **Dr. Arvind Sharma**

- Aryan College of Technology, Rajasthan Technology University, Kota
- **Duck Hee Lee**
Medical Engineering R&D Center/Asan Institute for Life Sciences/Asan Medical Center
 - **Elena SCUTELNICU**
"Dunarea de Jos" University of Galati
 - **Elena Camossi**
Joint Research Centre
 - **Eui Lee**
Sangmyung University
 - **Evgeny Nikulchev**
Moscow Technological Institute
 - **Ezekiel OKIKE**
UNIVERSITY OF BOTSWANA, GABORONE
 - **Fahim Akhter**
King Saud University
 - **FANGYONG HOU**
School of IT, Deakin University
 - **Faris Al-Salem**
GCET
 - **Firkhan Ali Hamid Ali**
UTHM
 - **Fokrul Alom Mazarbhuiya**
King Khalid University
 - **Frank Ibikunle**
Botswana Int'l University of Science & Technology (BIUST), Botswana
 - **Fu-Chien Kao**
Da-Y eh University
 - **Gamil Abdel Azim**
Suez Canal University
 - **Ganesh Sahoo**
RMRIMS
 - **Gaurav Kumar**
Manav Bharti University, Solan Himachal Pradesh
 - **George Pecherle**
University of Oradea
 - **George Mastorakis**
Technological Educational Institute of Crete
 - **Georgios Galatas**
The University of Texas at Arlington
 - **Gerard Dumancas**
Oklahoma Baptist University
 - **Ghalem Belalem**
University of Oran 1, Ahmed Ben Bella
 - **gherabi noreddine**
 - **Giacomo Veneri**
University of Siena
 - **Giri Babu**
Indian Space Research Organisation
 - **Govindarajulu Salendra**
 - **Grebenisan Gavril**
University of Oradea
 - **Gufran Ahmad Ansari**
Qassim University
 - **Gunaseelan Devaraj**
Jazan University, Kingdom of Saudi Arabia
 - **GYÖRÖDI ROBERT STEFAN**
University of Oradea
 - **Hadj Tadjine**
IAV GmbH
 - **Haewon Byeon**
Nambu University
 - **Haiguang Chen**
ShangHai Normal University
 - **Hamid Alinejad-Rokny**
The University of New South Wales
 - **Hamid AL-Asadi**
Department of Computer Science, Faculty of Education for Pure Science, Basra University
 - **Hamid Mukhtar**
National University of Sciences and Technology
 - **Hany Hassan**
EPF
 - **Harco Leslie Henic SPITS WARNARS**
Bina Nusantara University
 - **Hariharan Shanmugasundaram**
Associate Professor, SRM
 - **Harish Garg**
Thapar University Patiala
 - **Hazem I. El Shekh Ahmed**
Pure mathematics
 - **Hemalatha SenthilMahesh**
 - **Hesham Ibrahim**
Faculty of Marine Resources, Al-Mergheb University
 - **Himanshu Aggarwal**
Department of Computer Engineering
 - **Hongda Mao**
Hossam Faris
 - **Huda K. AL-Jobori**
Ahlia University
 - **Imed JABRI**

- **iss EL OUADGHIRI**
- **Iwan Setyawan**
Satya Wacana Christian University
- **Jacek M. Czerniak**
Casimir the Great University in Bydgoszcz
- **Jai Singh W**
- **JAMAIAH HAJI YAHAYA**
NORTHERN UNIVERSITY OF MALAYSIA (UUM)
- **James Coleman**
Edge Hill University
- **Jatinderkumar Saini**
Narmada College of Computer Application, Bharuch
- **Javed Sheikh**
University of Lahore, Pakistan
- **Jayaram A**
Siddaganga Institute of Technology
- **Ji Zhu**
University of Illinois at Urbana Champaign
- **Jia Uddin Jia**
Assistant Professor
- **Jim Wang**
The State University of New York at Buffalo,
Buffalo, NY
- **John Sahlin**
George Washington University
- **JOHN MANOHAR**
VTU, Belgaum
- **JOSE PASTRANA**
University of Malaga
- **Jui-Pin Yang**
Shih Chien University
- **Jyoti Chaudhary**
high performance computing research lab
- **K V.L.N.Acharyulu**
Bapatla Engineering college
- **Ka-Chun Wong**
- **Kamatchi R**
- **Kamran Kowsari**
The George Washington University
- **KANNADHASAN SURIYAN**
- **Kashif Nisar**
Universiti Utara Malaysia
- **Kato Mivule**
- **Kayhan Zrar Ghafoor**
University Technology Malaysia
- **Kennedy Okafor**
Federal University of Technology, Owerri
- **Khalid Mahmood**
IEEE
- **Khalid Sattar Abdul**
Assistant Professor
- **Khin Wee Lai**
Biomedical Engineering Department, University
Malaya
- **Khurram Khurshid**
Institute of Space Technology
- **KIRAN SREE POKKULURI**
Professor, Sri Vishnu Engineering College for
Women
- **KITIMAPORN CHOOCHOTE**
Prince of Songkla University, Phuket Campus
- **Krasimir Yordzhev**
South-West University, Faculty of Mathematics and
Natural Sciences, Blagoevgrad, Bulgaria
- **Krassen Stefanov**
Professor at Sofia University St. Kliment Ohridski
- **Labib Gergis**
Misr Academy for Engineering and Technology
- **LATHA RAJAGOPAL**
- **Lazar Stošić**
College for professional studies educators
Aleksinac, Serbia
- **Leanos Maglaras**
De Montfort University
- **Leon Abdillah**
Bina Darma University
- **Lijian Sun**
Chinese Academy of Surveying and
- **Ljubomir Jerinic**
University of Novi Sad, Faculty of Sciences,
Department of Mathematics and Computer Science
- **Lokesh Sharma**
Indian Council of Medical Research
- **Long Chen**
Qualcomm Incorporated
- **M. Reza Mashinchi**
Research Fellow
- **M. Tariq Banday**
University of Kashmir
- **madjid khalilian**
- **majzoob omer**
- **Mallikarjuna Doodipala**
Department of Engineering Mathematics, GITAM
University, Hyderabad Campus, Telangana, INDIA

- **Manas deep**
Masters in Cyber Law & Information Security
- **Manju Kaushik**
- **Manoharan P.S.**
Associate Professor
- **Manoj Wadhwa**
Echelon Institute of Technology Faridabad
- **Manpreet Manna**
Director, All India Council for Technical Education,
Ministry of HRD, Govt. of India
- **Manuj Darbari**
BBD University
- **Marcellin Julius Nkenlifack**
University of Dschang
- **Maria-Angeles Grado-Caffaro**
Scientific Consultant
- **Marwan Alseid**
Applied Science Private University
- **Mazin Al-Hakeem**
LFU (Lebanese French University) - Erbil, IRAQ
- **Md Islam**
sikkim manipal university
- **Md. Bhuiyan**
King Faisal University
- **Md. Zia Ur Rahman**
Narasaraopeta Engg. College, Narasaraopeta
- **Mehdi Bahrami**
University of California, Merced
- **Messaouda AZZOUZI**
Ziane Achour University of Djelfa
- **Milena Bogdanovic**
University of Nis, Teacher Training Faculty in Vranje
- **Miriampally Venkata Raghavendra**
Adama Science & Technology University, Ethiopia
- **Mirjana Popovic**
School of Electrical Engineering, Belgrade University
- **Miroslav Baca**
University of Zagreb, Faculty of organization and
informatics / Center for biometrics
- **Moeiz Miraoui**
University of Gafsa
- **Mohamed Eldosoky**
- **Mohamed Ali Mahjoub**
Preparatory Institute of Engineer of Monastir
- **Mohamed Kaloup**
- **Mohamed El-Sayed**
Faculty of Science, Fayoum University, Egypt
- **Mohamed Najeh LAKHOUA**
ESTI, University of Carthage
- **Mohammad Ali Badamchizadeh**
University of Tabriz
- **Mohammad Jannati**
- **Mohammad Alomari**
Applied Science University
- **Mohammad Haghighat**
University of Miami
- **Mohammad Azzeh**
Applied Science university
- **Mohammed Akour**
Yarmouk University
- **Mohammed Sadgal**
Cadi Ayyad University
- **Mohammed Al-shabi**
Associate Professor
- **Mohammed Hussein**
- **Mohammed Kaiser**
Institute of Information Technology
- **Mohammed Ali Hussain**
Sri Sai Madhavi Institute of Science & Technology
- **Mohd Helmy Abd Wahab**
University Tun Hussein Onn Malaysia
- **Mokhtar Beldjehem**
University of Ottawa
- **Mona Elshinawy**
Howard University
- **Mostafa Ezziyani**
FSTT
- **Mouhammd sharari alkasassbeh**
- **Mourad Amad**
Laboratory LAMOS, Bejaia University
- **Mueen Uddin**
University Malaysia Pahang
- **MUNTASIR AL-ASFOOR**
University of Al-Qadisiyah
- **Murphy Choy**
- **Murthy Dasika**
Geethanjali College of Engineering & Technology
- **Mustapha OUJAOURA**
Faculty of Science and Technology Béni-Mellal
- **MUTHUKUMAR SUBRAMANYAM**
DGCT, ANNA UNIVERSITY
- **N.Ch. Iyengar**
VIT University
- **Nagy Darwish**

Department of Computer and Information Sciences,
Institute of Statistical Studies and Researches, Cairo
University

- **Najib Kofahi**
Yarmouk University
- **Nan Wang**
LinkedIn
- **Natarajan Subramanyam**
PES Institute of Technology
- **Natheer Gharaibeh**
College of Computer Science & Engineering at
Yanbu - Taibah University
- **Nazeeh Ghatasheh**
The University of Jordan
- **Nazeeruddin Mohammad**
Prince Mohammad Bin Fahd University
- **NEERAJ SHUKLA**
ITM University, Gurgaon, (Haryana) India
- **Neeraj Tiwari**
- **Nestor Velasco-Bermeo**
UPFIM, Mexican Society of Artificial Intelligence
- **Nidhi Arora**
M.C.A. Institute, Ganpat University
- **Nilanjan Dey**
- **Ning Cai**
Northwest University for Nationalities
- **Nithyanandam Subramanian**
Professor & Dean
- **Noura Aknin**
University Abdelamlek Essaadi
- **Obaida Al-Hazaimeh**
Al- Balqa' Applied University (BAU)
- **Oliviu Matei**
Technical University of Cluj-Napoca
- **Om Sangwan**
- **Omaima Al-Allaf**
Asesstant Professor
- **Osama Omer**
Aswan University
- **Ouchtati Salim**
- **Ousmane THIARE**
Associate Professor University Gaston Berger of
Saint-Louis SENEGAL
- **Paresh V Virparia**
Sardar Patel University
- **Peng Xia**
Microsoft

- **Ping Zhang**
IBM
- **Poonam Garg**
Institute of Management Technology, Ghaziabad
- **Prabhat K Mahanti**
UNIVERSITY OF NEW BRUNSWICK
- **PROF DURGA SHARMA (PHD)**
AMUIT, MOEFDRE & External Consultant (IT) &
Technology Transfer Research under ILO & UNDP,
Academic Ambassador for Cloud Offering IBM-USA
- **Purwanto Purwanto**
Faculty of Computer Science, Dian Nuswantoro
University
- **Qifeng Qiao**
University of Virginia
- **Rachid Saadane**
EE departement EHTP
- **Radwan Tahboub**
Palestine Polytechnic University
- **raed Kanaan**
Amman Arab University
- **Raghuraj Singh**
Harcourt Butler Technological Institute
- **Rahul Malik**
- **raja boddu**
LENORA COLLEGE OF ENGINEERING
- **Raja Ramachandran**
- **Rajesh Kumar**
National University of Singapore
- **Rakesh Dr.**
Madan Mohan Malviya University of Technology
- **Rakesh Balabantaray**
IIIT Bhubaneswar
- **Ramani Kannan**
Universiti Teknologi PETRONAS, Bandar Seri
Iskandar, 31750, Tronoh, Perak, Malaysia
- **Rashad Al-Jawfi**
Ibb university
- **Rashid Sheikh**
Shri Aurobindo Institute of Technology, Indore
- **Ravi Prakash**
University of Mumbai
- **RAVINA CHANGALA**
- **Ravisankar Hari**
CENTRAL TOBACCO RESEARCH INSTITUTE
- **Rawya Rizk**
Port Said University

- **Reshmy Krishnan**
Muscat College affiliated to Stirling University.U
- **Ricardo Vardasca**
Faculty of Engineering of University of Porto
- **Ritaban Dutta**
ISSL, CSIRO, Tasmania, Australia
- **Rowayda Sadek**
- **Ruchika Malhotra**
Delhi Technological University
- **Rutvij Jhaveri**
Gujarat
- **SAADI Slami**
University of Djelfa
- **Sachin Kumar Agrawal**
University of Limerick
- **Sagarmay Deb**
Central Queensland University, Australia
- **Said Ghoniemy**
Taif University
- **Sandeep Reddivari**
University of North Florida
- **Sanskriti Patel**
Charotar University of Science & Technology,
Changa, Gujarat, India
- **Santosh Kumar**
Graphic Era University, Dehradun (UK)
- **Sasan Adibi**
Research In Motion (RIM)
- **Satyena Singh**
Professor
- **Sebastian Marius Rosu**
Special Telecommunications Service
- **Seema Shah**
Vidyalankar Institute of Technology Mumbai
- **Seifedine Kadry**
American University of the Middle East
- **Selem Charfi**
HD Technology
- **SENGOTTUVELAN P**
Anna University, Chennai
- **Senol Piskin**
Istanbul Technical University, Informatics Institute
- **Sérgio Ferreira**
School of Education and Psychology, Portuguese
Catholic University
- **Seyed Hamidreza Mohades Kasaei**
University of Isfahan
- **Shafiqul Abidin**
HMR Institute of Technology & Management
(Affiliated to GGS Indraprastha University), Hamidpur, Delhi -
110036
- **Shahanawaj Ahamad**
The University of Al-Kharj
- **Shaidah Jusoh**
- **Shaiful Bakri Ismail**
- **Shakir Khan**
Al-Imam Muhammad Ibn Saud Islamic University
- **Shawki Al-Dubae**
Assistant Professor
- **Sherif Hussein**
Mansoura University
- **Shriram Vasudevan**
Amrita University
- **Siddhartha Jonnalagadda**
Mayo Clinic
- **Sim-Hui Tee**
Multimedia University
- **Simon Ewedafe**
The University of the West Indies
- **Siniša Opic**
University of Zagreb, Faculty of Teacher Education
- **Sivakumar Poruran**
SKP ENGINEERING COLLEGE
- **Slim BEN SAOUD**
National Institute of Applied Sciences and
Technology
- **Sofien Mhatli**
- **sofyan Hayajneh**
- **Sohail Jabbar**
Bahria University
- **Sri Devi Ravana**
University of Malaya
- **Sudarson Jena**
GITAM University, Hyderabad
- **Suhail Sami Owais Owais**
- **Suhas J Manangi**
Microsoft
- **SUKUMAR SENTHILKUMAR**
Universiti Sains Malaysia
- **Süleyman Eken**
Kocaeli University
- **Sumazly Sulaiman**
Institute of Space Science (ANGKASA), Universiti
Kebangsaan Malaysia

- **Sumit Goyal**
National Dairy Research Institute
- **Suparek Janjarasjitt**
Ubon Ratchathani University
- **Suresh Sankaranarayanan**
Institut Teknologi Brunei
- **Susarla Sastry**
JNTUK, Kakinada
- **Suseendran G**
Vels University, Chennai
- **Suxing Liu**
Arkansas State University
- **Syed Ali**
SMI University Karachi Pakistan
- **T C.Manjunath**
HKBK College of Engg
- **T V Narayana rao Rao**
SNIST
- **T. V. Prasad**
Lingaya's University
- **Taiwo Ayodele**
Infonetmedia/University of Portsmouth
- **Talal Bonny**
Department of Electrical and Computer Engineering, Sharjah University, UAE
- **Tamara Zhukabayeva**
- **Tarek Gharib**
Ain Shams University
- **thabet slimani**
College of Computer Science and Information Technology
- **Totok Biyanto**
Engineering Physics, ITS Surabaya
- **Touati Youcef**
Computer sce Lab LIASD - University of Paris 8
- **Tran Sang**
IT Faculty - Vinh University - Vietnam
- **Tsvetanka Georgieva-Trifonova**
University of Veliko Tarnovo
- **Uchechukwu Awada**
Dalian University of Technology
- **Udai Pratap Rao**
- **Urmila Shrawankar**
GHRCE, Nagpur, India
- **Vaka MOHAN**
TRR COLLEGE OF ENGINEERING
- **VENKATESH JAGANATHAN**
- **ANNA UNIVERSITY**
- **Vinayak Bairagi**
AISSMS Institute of Information Technology, Pune
- **Vishnu Mishra**
SVNIT, Surat
- **Vitus Lam**
The University of Hong Kong
- **VUDA SREENIVASARAO**
PROFESSOR AND DEAN, St.Mary's Integrated Campus, Hyderabad
- **Wali Mashwani**
Kohat University of Science & Technology (KUST)
- **Wei Wei**
Xi'an Univ. of Tech.
- **Wenbin Chen**
360Fly
- **Xi Zhang**
illinois Institute of Technology
- **Xiaojing Xiang**
AT&T Labs
- **Xiaolong Wang**
University of Delaware
- **Yanping Huang**
- **Yao-Chin Wang**
- **Yasser Albagory**
College of Computers and Information Technology, Taif University, Saudi Arabia
- **Yasser Alginahi**
- **Yi Fei Wang**
The University of British Columbia
- **Yihong Yuan**
University of California Santa Barbara
- **Yilun Shang**
Tongji University
- **Yu Qi**
Mesh Capital LLC
- **Zacchaeus Omogbadegun**
Covenant University
- **Zairi Rizman**
Universiti Teknologi MARA
- **Zarul Zaaba**
Universiti Sains Malaysia
- **Zenzo Ncube**
North West University
- **Zhao Zhang**
Deptment of EE, City University of Hong Kong
- **Zhihan Lv**

Chinese Academy of Science

- **Zhixin Chen**
ILX Lightwave Corporation
- **Ziyue Xu**
National Institutes of Health, Bethesda, MD

- **Zlatko Stacic**
University of Zagreb, Faculty of Organization and
Informatics Varazdin
- **Zuraini Ismail**
Universiti Teknologi Malaysia

CONTENTS

Paper 1: *HappyMeter: An Automated System for Real-Time Twitter Sentiment Analysis*

Authors: Joaquim Perotti Canela, Tina Tian

PAGE 1 – 7

Paper 2: *A Comparison between Chemical Reaction Optimization and Genetic Algorithms for Max Flow Problem*

Authors: Mohammad Y. Khanafseh, Ola M. Surakhi, Ahmad Sharieh, Azzam Sleit

PAGE 8 – 15

Paper 3: *Meteonowcasting using Deep Learning Architecture*

Authors: Sanam Narejo, Eros Pasero

PAGE 16 – 23

Paper 4: *A Review of Towered Big-Data Service Model for Biomedical Text-Mining Databases*

Authors: Alshreef Abed, Jingling Yuan, Lin Li

PAGE 24 – 35

Paper 5: *A Non-Linear Regression Modeling is used for Asymmetry Co-Integration and Managerial Economics in Iraqi Firms*

Authors: Karrar Abdulellah Azeez, Han DongPing, Marwah Abdulkareem Mahmood

PAGE 36 – 41

Paper 6: *DDoS Attacks Classification using Numeric Attribute-based Gaussian Naive Bayes*

Authors: Abdul Fadlil, Imam Riadi, Sukma Aji

PAGE 42 – 50

Paper 7: *A Features-based Comparative Study of the State-of-the-art Cloud Computing Simulators and Future Directions*

Authors: Ahmad Waqas, M. Abdul Rehman, Abdul Rehman Gilal, Mohammad Asif Khan, Javed Ahmed, Zulkefli Muhammed Yusof

PAGE 51 – 59

Paper 8: *An Innovative Cognitive Architecture for Humanoid Robot*

Authors: Muhammad Faheem Mushtaq, Urooj Akram, Adeel Tariq, Irfan Khan, Muhammad Zulqarnain, Umer Iqbal

PAGE 60 – 67

Paper 9: *Shadow Identification in Food Images using Extreme Learning Machine*

Authors: Salwa Khalid Abdulateef, Massudi Mahmuddin, Nor Hazlyna Harun

PAGE 68 – 74

Paper 10: *PCA based Optimization using Conjugate Gradient Descent Algorithm*

Authors: Subhas A. Meti, V.G. Sangam

PAGE 75 – 82

Paper 11: *Improvement of Radial basis Function Interpolation Performance on Cranial Implant Design*

Authors: Ferhat Atasoy, Baha Sen, Fatih Nar, Ismail Bozkurt

PAGE 83 – 88

Paper 12: Performance Evaluation of Transmission Line Protection Characteristics with DSTATCOM Implementation

Authors: Yasar Khan, Khalid Mahmood, Sanaullah Ahmad

PAGE 89 – 99

Paper 13: Synchronous Authentication Key Management Scheme for Inter-eNB Handover over LTE Networks

Authors: Shadi Nashwan

PAGE 100 – 107

Paper 14: A New Cryptosystem using Vigenere and Metaheuristics for RGB Pixel Shuffling

Authors: Zakaria KADDOURI, Mohamed Amine Hyaya, Mohamed KADDOURI

PAGE 108 – 113

Paper 15: Improved Hybrid Model in Vehicular Clouds based on Data Types (IHVCDT)

Authors: Saleh A. Khawatreh, Enas N. Al-Zubi

PAGE 114 – 118

Paper 16: FPGA Implementation of SVM for Nonlinear Systems Regression

Authors: Intissar SAYEHI, Mohsen MACHHOUT, Rached TOURKI

PAGE 119 – 129

Paper 17: A Synthesis on SWOT Analysis of Public Sector Healthcare Knowledge Management Information Systems in Pakistan

Authors: Arfan Arshad, Mohamad Fauzan Noordin, Roslina Bint Othman

PAGE 130 – 136

Paper 18: Multi-Agent based Functional Testing in the Distributed Environment

Authors: Muhammad Fraz Malik, M. N. A. Khan, Uzma Bibi, Muhammad Ayaz Malik

PAGE 137 – 143

Paper 19: Modeling and Implementing Ontology for Managing Learners' Profiles

Authors: Korchi Adil, El Amrani El Idrissi Najiba, Oughdir Lahcen

PAGE 144 – 152

Paper 20: Suitable Personality Traits for Learning Programming Subjects: A Rough-Fuzzy Model

Authors: Abdul Rehman Gilal, Jafreezal Jaafar, Mazni Omar, Shuib Basri, Izzatdin Abdul Aziz, Qamar Uddin Khand, Mohd Hilmi Hasan

PAGE 153 – 162

Paper 21: Usability of Government Websites

Authors: Mahmood Ashraf, Faiza Shabbir Cheema, Tanzila Saba, Abdul Mateen

PAGE 163 – 167

Paper 22: InstDroid: A Light Weight Instant Malware Detector for Android Operating Systems

Authors: Saba Arshad, Rabia Chaudhary, Munam Ali Shah, Neshmia Hafeez, Muhammad Kamran Abbasi

PAGE 168 – 175

Paper 23: A 7-Layered E-Government Framework Consolidating Tehnical, Social and Managerial Aspects

Authors: Mohammed Hitham M.H, Dr. Hatem Elkadi H.K, Dr. Sherine Ghoneim S.G

PAGE 176 – 184

Paper 24: Validating A Novel Conflict Resolution Strategy Selection Method (Confrssm) Via Multi-Agent Simulation

Authors: Alicia Y.C. Tang, Ghusoon Salim Basheer

PAGE 185 – 194

Paper 25: Modeling and Verification of Payment System in E-Banking

Authors: Iqra Obaid, Syed Asad Raza Kazmi, Awais Qasim

PAGE 195 – 201

Paper 26: ReCSDN: Resilient Controller for Software Defined Networks

Authors: Soomaiya Hamid, Narmeen Zakaria Bawany, Jawwad Ahmed Shamsi

PAGE 202 – 208

Paper 27: Detection and Prevention of SQL Injection Attack by Dynamic Analyzer and Testing Model

Authors: Rana Muhammad Nadeem, Rana Muhammad Saleem, Rabnawaz Bashir, Sidra Habib

PAGE 209 – 214

Paper 28: Normalisation of Technology Use in a Developing Country Higher Education Institution

Authors: Ibrahim Osman Adam, Osman Issah

PAGE 215 – 222

Paper 29: Design and Simulation of a Novel Dual Band Microstrip Antenna for LTE-3 and LTE-7 Bands

Authors: Abdullah Al Hasan, Mohammad Shahriar Siraj, Muhammad Mostafa Amir Faisal

PAGE 223 – 228

Paper 30: Mobile Learning Application Development for Improvement of English Listening Comprehension

Authors: Zahida Parveen Laghari, Hameedullah Kazi, Muhammad Ali Nizamani

PAGE 229 – 237

Paper 31: Toward a New Massively Distributed Virtual Machine based Cloud Micro-Services Team Model for HPC: SPMD Applications

Authors: Fatéma Zahra Benchara, Mohamed Youssfi, Omar Bouattane, Ouafae Serrar, Hassan Ouajji

PAGE 238 – 249

Paper 32: Energy Management Strategy of a PV/Fuel Cell/Supercapacitor Hybrid Source Feeding an off-Grid Pumping Station

Authors: Housseem CHAOUALI, Hichem OTHMANI, Mohamed Selméne BEN YAHIA, Dhafer MEZGHANI, Abdelkader MAMI

PAGE 250 – 257

Paper 33: Object's Shape Recognition using Local Binary Patterns

Authors: Muhammad Wasim, Adnan Ahmed Siddiqui, Abdul Aziz, Lubaid Ahmed, Syed Faisal Ali, Fauzan Saeed

PAGE 258 – 262

Paper 34: Feature Extraction and Classification Methods for a Motor Task Brain Computer Interface: A Comparative Evaluation for Two Databases

Authors: Oana Diana Eva, Anca Mihaela Lazar

PAGE 263 – 269

Paper 35: Creating and Protecting Password: A User Intention

Authors: Ari Kusyanti, Yustiyana April Lia Sari

PAGE 270 – 275

Paper 36: *Analyzing the Social Awareness in Autistic Children Trained Through Multimedia Intervention Tool using Data Mining*

Authors: Richa Mishra, Divya Bhatnagar

PAGE 276 – 280

Paper 37: *Context Aware Fuel Monitoring System for Cellular Sites*

Authors: Mohammad Asif Khan, Ahmad Waqas, Qamar Uddin Khand, Sajid Khan

PAGE 281 – 285

Paper 38: *Text Steganography using Extensions Kashida based on the Moon and Sun Letters Concept*

Authors: Anes.A.Shaker, Farida Ridzuan, Sakinah Ali Pitchay

PAGE 286 – 290

Paper 39: *Studying the Influence of Static Converters' Current Harmonics on a PEM Fuel Cell using Bond Graph Modeling Technique*

Authors: Wafa BEN SALEM, Housseem CHAOUALI, Dhia MZOUGH, Abdelkader MAMI

PAGE 291 – 301

Paper 40: *The Effect of Religious Beliefs, Participation and Values on Corruption: Survey Evidence from Iraq*

Authors: Marwah Abdulkareem Mahmood Zuhaira, Tian Ye-zhuang

PAGE 302 – 305

Paper 41: *Detection of Distributed Denial of Service Attacks Using Artificial Neural Networks*

Authors: Abdullah Aljumah

PAGE 306 – 318

Paper 42: *Artificial Intelligence in Bio-Medical Domain*

Authors: Muhammad Salman, Abdul Wahab Ahmed, Omair Ahmad Khan, Basit Raza, Khalid Latif

PAGE 319 – 327

Paper 43: *A Hybrid Curvelet Transform and Genetic Algorithm for Image Steganography*

Authors: Heba Mostafa Mohamed, Ahmed Fouad Ali, Ghada Sami Altaweel

PAGE 328 – 336

Paper 44: *Automatic Music Genres Classification using Machine Learning*

Authors: Muhammad Asim Ali, Zain Ahmed Siddiqui

PAGE 337 – 344

Paper 45: *The Identification of Randles Impedance Model Parameters of a PEM Fuel Cell by the Least Square Method*

Authors: Med Selméne Ben Yahia, Hatem Allagui, Abdelkader Mami

PAGE 345 – 354

Paper 46: *Using the Facebook Iframe as an Effective Tool for Collaborative Learning in Higher Education*

Authors: Mohamed A. Amasha, Salem Alkhalaf

PAGE 355 – 360

Paper 47: *AES-Route Server Model for Location based Services in Road Networks*

Authors: Mohamad Shady Alrahal, Muhammad Usman Ashraf, Adnan Abesen, Sabah Arif

PAGE 361 – 368

Paper 48: Visualizing Computer Programming in a Computer-based Simulated Environment

Authors: Belsam Attallah

PAGE 369 – 378

Paper 49: Hybrid Technique for Java Code Complexity Analysis

Authors: Nouh Alhindawi, Mohammad Subhi Al-Batah, Rami Malkawi, Ahmad Al-Zuraiqi

PAGE 379 – 385

Paper 50: An Unsupervised Local Outlier Detection Method for Wireless Sensor Networks

Authors: Tianyu Zhang, Qian Zhao, Yoshihiro Shin, Yukikazu Nakamoto

PAGE 386 – 393

Paper 51: Adaptive e-learning using Genetic Algorithm and Sentiments Analysis in a Big Data System

Authors: Youness MADANI, Jamaa BENGOURRAM, Mohammed ERRITALI, Badr HSSINA, Marouane Birjali

PAGE 394 – 403

Paper 52: Solving the Free Clustered TSP Using a Memetic Algorithm

Authors: Abdullah Alsheddy

PAGE 404 – 408

Paper 53: Lung Cancer Detection and Classification with 3D Convolutional Neural Network (3D-CNN)

Authors: Wafaa Alakwaa, Mohammad Nassef, Amr Badr

PAGE 409 – 417

Paper 54: Multiple Vehicles Semi-Self-driving System Using GNSS Coordinate Tracking under Relative Position with Correction Algorithm

Authors: Heejin Lee, Hiroshi Suzuki, Takahiro Kitajima, Akinobu Kuwahara, Takashi Yasuno

PAGE 418 – 426

Paper 55: An Efficient Scheme for Real-time Information Storage and Retrieval Systems: A Hybrid Approach

Authors: Syed Ali Hassan, Imran Ul Haq, Muhammad Asif, Maaz Bin Ahmad, Moeen Tayyab

PAGE 427 – 431

Paper 56: Exploiting Temporal Information in Documents and Query to Improve the Information Retrieval Process: Application to Medical Articles

Authors: Jihen MAJDOUBI, Ahlam Nabli

PAGE 432 – 440

Paper 57: A Comparison of Predictive Parameter Estimation using Kalman Filter and Analysis of Variance

Authors: Asim ur Rehman Khan, Haider Mehdi, Syed Muhammad Atif Saleem, Muhammad Junaid Rabbani

PAGE 441 – 446

Paper 58: Fine Grained Accelerometer based Smartphone Carrying States Recognition during Walking

Authors: Kaori Fujinami, Tsubasa Saeki, Yinghuan Li, Tsuyoshi Ishikawa, Takuya Jimbo, Daigo Nagase, Koji Sato

PAGE 447 – 456

Paper 59: Automated Player Selection for a Sports Team using Competitive Neural Networks

Authors: Rabah Al-Shboul, Tahir Syed, Jamshed Memon, Furqan Khan

PAGE 457 – 460

HappyMeter: An Automated System for Real-Time Twitter Sentiment Analysis

Joaquim Perotti Canela
Department of Computer Science
Manhattan College
New York, USA

Tina Tian
Department of Computer Science
Manhattan College
New York, USA

Abstract—The paper presents HappyMeter, an automated system for real-time Twitter sentiment analysis. More than 380 million tweets consisting of nearly 30,000 words, almost 6,000 hashtags and over 5,000 user mentioned have been studied. A sentiment model is used to measure the sentiment level of each term in the contiguous United States. The system automatically mines real-time Twitter data and reveals the changing patterns of the public sentiment over an extended period of time. It is possible to compare the public opinions regarding a subject, hashtag or a Twitter user between different states in the U.S. Users may choose to see the overall sentiment level of a term, as well as its sentiment value on a specific day. Real-time results are delivered continuously and visualized through a web-based graphical user interface.

Keywords—Twitter; social networks; data mining; sentiment analysis

I. INTRODUCTION

Twitter has become an increasingly popular microblogging service that allows users to publish messages, a.k.a. tweets [1]. It functions as a platform for people to express themselves, which often carries opinions on different subjects. Twitter usage is growing exponentially. There are 328 million monthly active users on Twitter and over 500 million tweets are created per day [2].

The rapid growth of Twitter and the public access of tweets have made Twitter a popular research subject. For example, researchers have examined the use of Twitter in promoting products and sharing consumer opinions [3]. Enterprises have studied the usefulness of Twitter in organizational communication and information-gathering [4]. Furthermore, tweets have been monitored to detect earthquakes [5].

In this paper, we present HappyMeter, a sentiment analysis tool that measures happiness on Twitter. Sentiment analysis is to computationally categorize opinions expressed in a given text. It is essentially important in social media monitoring as it provides an overview of the public sentiment regarding certain topics.

Unlike other online articles, Twitter messages share several unique features. Firstly, the vernacular on Twitter is informal [6]. There could be misspelled words, slang and acronyms in a tweet due to Twitter's informal language style [7]. Secondly, every tweet has a length constraint of maximum 140 characters [8]. Moreover, Twitter covers an exceedingly broad range of topics [6]. Lastly, due to the wide usage of mobile devices and

the rapid flow of tweets, this user-generated data reflect instant reactions as events evolve. Therefore, we built our system intended for providing real-time insights of the public sentiment and showing changes over time.

Our paper presents an automated sentiment analyzer based on the Twitter traffic. The system streams all the tweets published in the contiguous United States in real time. For each tweet, a sentiment score is computed using a statistical sentiment model and the geographical data associated with the tweet are stored. We developed a web-based graphical user interface to deliver results instantly and continuously.

The rest of the paper is organized as follows. Section II reviews the related work. In Section III, we describe the data set used to build the system and introduce the methods and algorithms adapted in measuring the data. Section IV presents the results of the study and the visualization we have built. Section V concludes the paper and proposes future directions.

II. RELATED WORK

Applications of sentiment analysis are broad and powerful. As a subfield of Machine Learning and Natural Language Processing, research has been conducted ranging from document level classification [9] to determining the polarity (positive, negative or neutral) of sentences [10] and terms [11]. In recent years, sentiment analysis on Twitter, specifically, has attracted increasing attention from many research communities. For instance, Bollen et al. investigated whether public mood on Twitter is correlated with shifts in the stock market [12]. Vegas et al. modeled the 2016 U.S. presidential campaign in the context of Twitter [13]. Zeitzoff used data on Twitter to measure social movements [14].

To determine the sentiment of a tweet, many past studies have focused on supervised learning where the training data are collected based on emoticons, hashtags or both [15], [16]. Experiments show, however, that they contain biased information in sentiment analysis [7]. Another common practice is to manually annotate the data in order to build a pool of training data. The apparent disadvantage of this method is the intensive labor and time involved in the process.

Our work is inspired by Dodds et al.'s study on temporal patterns of happiness on Twitter, in which they used a corpus mapped with happiness scores to examine the sentiment variations on different expressions over time [17]. The expressions, however, consist of mainly words. Other crucial

elements of a tweet, such as hashtags and user mentions, have not been much studied. Moreover, there was no geographical comparison; for example, exploring the public sentiment of a term between different states in the U.S. A later study by Mitchell et al. considered the geographical factor [18], but the results did not reveal the changing patterns over time, for example, observing the sentiment of a term in a particular state over an extended period of time.

In this paper, we present a system that shows the public sentiment of every term on Twitter, including unigrams, hashtags and user mentions. The system computes the public sentiment of every term in each contiguous U.S. state. The process is repeated daily. Results are visualized to reveal the sentiment alternation over time, as well as the comparisons between different states.

III. THE SYSTEM

The system performs a sentiment analysis on a Twitter tweet corpus collected since June 2016. In this section, we discuss the data set used in the study and how we define and calculate sentiment.

A. Data Set

The Twitter Streaming API [19] allows us to crawl real-time tweets and receive instant updates. Currently, we have gathered over 380 million tweets with geographical annotation enabled from the contiguous United States. This number keeps growing in the rate of 1.4 million tweets per day on average. Table 1 shows the basic statistics of the data set in the study. The highest volume of tweets was received on November 8, 2016, when the United States presidential election took place. The system collected more than 2 million tweets (2,292,345 to be precise) from the contiguous U.S. in a day.

TABLE I. STATISTICS OF THE TWEETS COLLECTED

	Number of Tweets
Total Tweets Collected	388,664,640
Average Daily Tweets Collected	1,423,680
Standard Deviation	199,943

As mentioned in Section II, one of the major contributions of this work is the sentiment mining of some key components on Twitter, such as hashtags and user mentions. Thus far, the extracted tweets consist of a massive corpus of more than 29,000 unique unigrams, almost 6,000 distinctive hashtags and over 5,000 different user mentions. Table 2 provides an up-to-date summary of the individual terms collected in the study. Our system performs a sentiment analysis on each of these terms in the context of Twitter.

TABLE II. STATISTICS OF THE TERMS EXTRACTED

	Number of Terms
Words	29,455
Hashtags (#)	5,991
User Mentions (@)	5,074
Total	40,520

B. Defining Sentiment

The sentiment of a term in a Twitter message is determined by an existing sentiment lexicon, the dictionary of Language Assessment by Mechanical Turk [17]. The list contains over

10,000 most popular words with their average sentiment score, ranging from 1 to 9. In general, happy words have a high sentiment value with a score close to 9, while sad words are usually associated with a low score. Table 3 shows a sample of the lexicon.

TABLE III. SAMPLE OF THE SENTIMENT LEXICON

Word	Sentiment Score
laugh	8.22
dancing	7.08
torch	5.11
tension	2.94
suicide	1.30

C. Processing

The system collects real-time tweets through the streaming API and saves them on a server for further processing. As seen in Fig. 1, the process includes data manipulations, such as data cleaning and location identification, and sentiment computation. Results are thereafter stored in a database. The rest of this section elaborates the processing procedure.

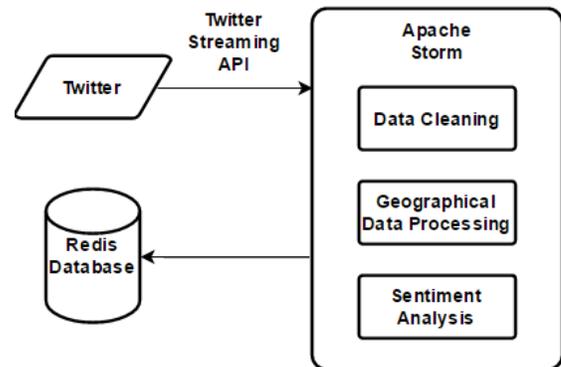


Fig. 1. System architectural overview.

Twitter has the geotagging feature (Tweeting with Location) which allows users to publish a tweet with their location [20]. This feature helps to make tweets more contextual. In the meantime, it provides valuable data for research. One thing to note is that users must give explicit permission for their exact location to be displayed with their tweets, due to Twitter’s user privacy policy. Thus, not all the tweets collected come with geographical data. In this study, we keep only the geo-tagged Twitter messages. For each tweet, we store the state where the tweet was issued.

As our work targets tweets in English, tweets written in other languages are discarded. Non-English characters in a tweet are also erased. Due to the informal language model on Twitter (mentioned in Section I), misspelled words can often occur. The system cleans the data by removing words that do not exist in the sentiment lexicon. Hashtags and user mentions, however, are kept in the data set.

To compute the sentiment of a tweet, the system performs a simple average on the sentiment score of each word. Hashtags and user mentions are excluded in this process. Moreover, stop words with a neutral sentiment value falling between 4 and 6 are also excluded in the calculation, following Dodds et al. [17].

Let us take the following tweet for an example, “@missnemmanuel is so gorgeous! #GoTS7e2 #GoT #newcrush”.

Among the terms, “is” has a sentiment score of 5.18, while “so” and “gorgeous” have a sentiment value being 5.08 and 7.42, respectively. Discarding hashtags, user mentions and neutral stop words, only word “gorgeous” is kept in calculating the sentiment of the tweet. Therefore, the average of 7.42, which is 7.42 itself, is assigned to the example tweet.

The system associates this computed sentiment value with every term in the tweet, including hashtags, user mentions and even stop words. Thus, in the previous example, each of the following terms receives a sentiment score of 7.42 along with the tweet: they are @missnemmanuel, is, so, gorgeous, #GoTS7e2, #GoT and #newcrush.

Each of the terms in the example is highly likely to appear in other tweets as well. The system collects the sentiment scores of a term in all occasions in a day and concludes a mean value. For instance, if @missnemmanuel is mentioned 10,000 times in one day, the system would gather the sentiment values from the 10,000 tweets and compute the average. In this work, we examine the daily sentiment of a term in the contiguous United States as a whole, as well as in each state.

One may question the need of computing sentiment of a neutral word. It may not seem necessary from the previously given tweet. But let us consider word “governor” as another example. According to the lexicon, it has a sentiment value of 5.14, which falls in the range of a neutral stop word. However, it would be interesting to see that some states share a higher sentiment value towards “governor” than others do. Moreover, it would be especially interesting to observe the changing pattern over time.

Another concern one may raise is the capacity and scalability of the system. After all, there are millions of potential user mentions and hashtags on Twitter. Plus, new ones are emerging in every second. Keeping a daily record of sentiment for all of them would require tremendous spatial resources and computing power. To tackle this issue, we set a threshold of 3 to be the minimum daily occurrences of a term. Hashtags or usernames mentioned less than three times in a day in the contiguous U.S. are discarded from the database. We believe that this method can help filtering only the active hashtags and popular user mentions.

IV. RESULTS

In this section, we demonstrate our system and present results from the analysis. We first studied the amount of daily tweets issued by each contiguous state. To justify the different populations in each state, we calculated the average number of tweets published per day per 10,000 capita in a state. Population estimates are retrieved from the United States Census Bureau [21]. Fig. 2 shows the results after applying Jenks natural breaks optimization [22]. Results ranging from 18 to 80 have been divided into five groups. Among the contiguous U.S. states, Louisiana delivers the most tweets per day per capita, while Wyoming has the least number of daily tweets per capita, as seen in Table 4.

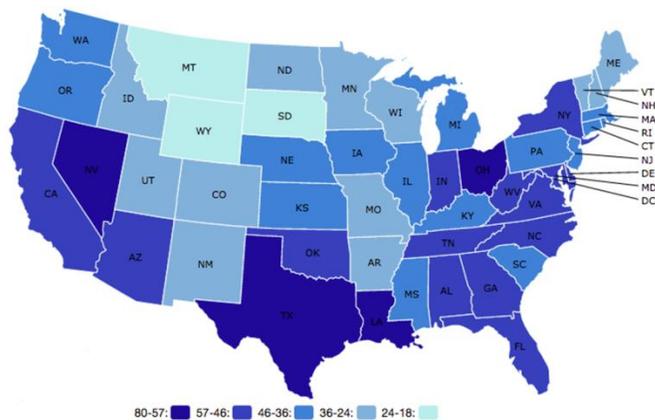


Fig. 2. Average number of daily tweets per 10,000 people in the contiguous United States.

TABLE IV. NUMBER OF DAILY TWEETS PER 10,000 CAPITA

	Daily Tweets per 10,000 Capita	States
Highest	80	Louisiana
Lowest	18	Wyoming
Average	44	N/A
Standard Deviation	13	N/A

Using the methodology introduced in Section III, we investigated the overall sentiment value of each U.S. state based on tweets collected from that region. Results range from 5.92 to 6.03 with a small standard deviation. The average sentiment value for the contiguous United States overall is 5.96. In our study, West Virginia and Wisconsin have the highest average sentiment value, while Alabama, Arkansas, Vermont and Virginia share the lowest sentiment score. Table 5 shows a summary of the statistics.

TABLE V. AVERAGE SENTIMENT SCORES

	Sentiment Score	States
Highest	6.04	Minnesota, Iowa, Nebraska and Utah
Lowest	5.92	Delaware
Average	5.99	N/A
Standard Deviation	0.03	N/A

Similar to Fig. 2, Jenks classification algorithm was utilized to visualize the variation of average sentiment values among U.S. states. Fig. 3 shows the results with five splits.

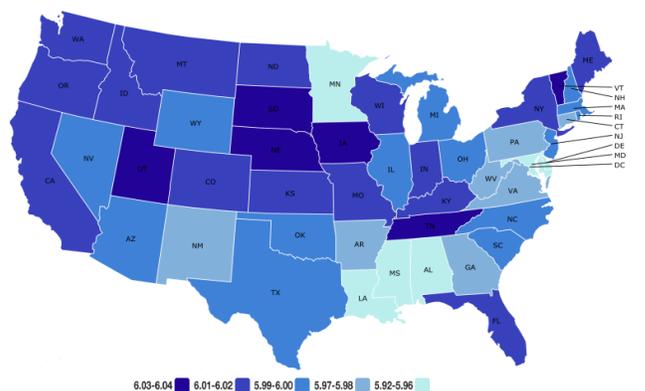


Fig. 3. Average sentiment scores of states in the contiguous United States

This paper also examines the overall sentiment level of each word tweeted in the network. As one can see from the histogram in Fig. 4, most of the words (nearly 75%) have an average sentiment score between 6 and 7, which is considered positive. More than 21% of the words have an overall sentiment value falling between 5 and 6. The highest sentiment score of a word is 8.01 and the lowest sentiment value calculated is 3.73.

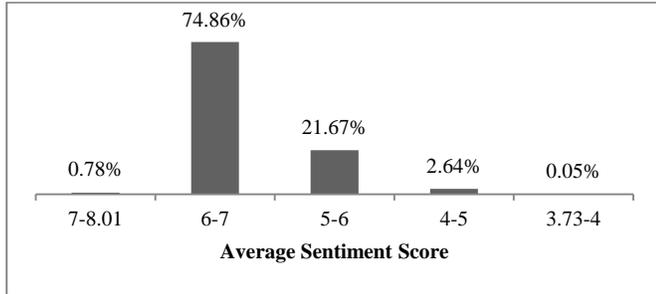


Fig. 4. Histogram of average sentiment scores.

Recall in Section III, a sentiment lexicon built with Language Assessment by Mechanical Turk (LabMT) was used to determine the initial sentiment of a standalone word. Each term was then processed by the HappyMeter system for the overall sentiment in the context of tweets. Table 6 shows the comparisons between the sentiment values before and after the processing of our system. As shown in the table, averagely, the overall sentiment has increased in the Twitter context. Moreover, the standard deviation has significantly dropped, meaning there are fewer extreme ratings. In general, contextual sentiment has become higher and milder. The two sentiment values have a strong Pearson’s correlation of 0.73.

TABLE VI. CHANGE OF SENTIMENT SCORES BEFORE AND AFTER HAPPY METER

	LabMT Sentiment Score	HappyMeter Sentiment Score
Average	5.38	5.97
Standard Deviation	1.08	0.64

TABLE VII. TOP 10 WORDS, HASHTAGS AND USER MENTIONS WITH THE HIGHEST SENTIMENT SCORES

Rank	Word	Score	Hashtag	Score	User Mention	Score
1	birthdayyy	8.01	#beats	7.42	@wwwbigbalthead	6.91
2	brotherly	7.71	#sweepstakes	7.17	@taylorswift13	6.88
3	gday	7.60	#birthday	7.16	@samheughan	6.86
4	belated	7.50	#happybirthday	7.14	@blakeshelton	6.85
5	birthday	7.46	#sweeps	7.13	@jlo	6.85
6	happy	7.41	#shoutout	7.08	@iheartradio	6.81
7	happyy	7.38	#flowers	7.07	@britneyspears	6.80
8	splendid	7.35	#beach	7.06	@shawnmendes	6.79
9	congratulations	7.34	#karaoke	7.06	@applebees	6.79
10	unconditional	7.34	#puppylove	7.05	@ethandolan	6.79

TABLE VIII. TOP 10 WORDS, HASHTAGS AND USER MENTIONS WITH THE LOWEST SENTIMENT SCORES

Rank	Word	Score	Hashtag	Score	User Mention	Score
1	headache	3.73	#atxtraffic	3.75	@cnn	5.56
2	killling	3.93	#atltraffic	3.89	@foxnews	5.61
3	kill	3.96	#dfwtraffic	3.89	@thehill	5.62
4	murder	3.98	#traffic	4.08	@msnbc	5.67
5	dead	3.99	#sfltraffic	4.39	@cnnpolitics	5.69
6	accident	4.00	#fail	4.94	@senatemajldr	5.70
7	crying	4.02	#weather	5.16	@nytimes	5.72
8	ouch	4.02	#tampabay	5.36	@nbcnews	5.72
9	jail	4.02	#breaking	5.45	@politico	5.72
10	sick	4.03	#audible	5.69	@senategop	5.72

Table 7 shows the top 10 words, hashtags and user mentions with the highest sentiment level. As we can see from the top words and hashtags, most people feel happy when they tweet about birthdays, flowers, beaches, karaoke and puppies. The best rated Twitter users include mostly movie and television stars, singers and comedians. Table 8, on the other hand, shows a list of the top 10 words, hashtags and user mentions with the lowest sentiment values. As shown in the table, besides the extreme words, such as murder, kill and dead, traffic is the number one problem in people’s common life. User mentions that are associated with low sentiment values are mainly Twitter accounts belonging to the news media.

In this work, we also examined the frequencies of each term on the Twitter network. Table 9 gives a glance of the most often appeared words during the observation, along with their sentiment score. Note that neutral stop words (mentioned in Section III) have been excluded from the list. For example, “just” is the most used word appearing over 12 million times in our data set. However, it is not collected in the list because it can tell little about the public state of mind. We are happy to report that all of the top 10 popular words have a positive sentiment polarity with a sentiment value greater than 6.

TABLE IX. TOP 10 MOST FREQUENTLY APPEARED WORDS (EXCLUDING STOP WORDS)

Rank	Word	Occurrence	Score
1	like	23,880,374	6.38
2	love	16,934,940	7.29
3	see	13,304,971	6.14
4	good	12,727,193	6.59
5	day	12,204,366	6.41
6	lol	11,388,765	6.27
7	great	10,811,384	7.02
8	will	10,507,779	6.02
9	happy	10,177,887	7.41
10	today	9,759,280	6.19

To further investigate the language model on the Twitter network, we built a histogram of usage frequencies of words in the data set, as seen in Fig. 5. Among the 380 million tweets we have received, 35% of the English words appeared between 10,000 and 100,000 times. Interestingly, 26% of the words occurred only less than 100 times. This again proved the casual language style and the rapid change of vocabulary on Twitter.

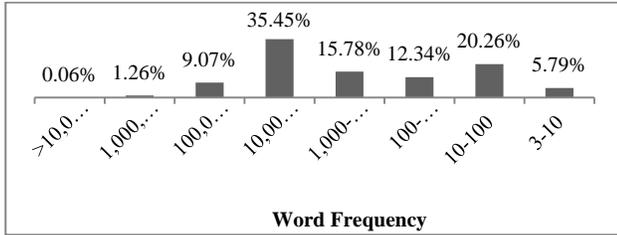


Fig. 5. Histogram of word frequencies.

We studied the most frequently occurred hashtags, shown in Table X along with their average sentiment value. Job related subjects, such as #job and #hiring, rank at the top of the list leaving the rest far behind. Hashtag #traffic ranks at number 15 with 426,309 mentions and #trump rank at number 18 with 377,559 appearances (not shown in Table 10). Hashtag #education appears later in the list, with 206,122 tags ranking at number 34. Geographical hashtags wise, New York attracts the most attention with 253,696 times mentioning #newyork and 228,345 references of #nyc. Followed after it are #houston with 369,117 occurrences and #chicago with 260,625 tags. The complete list of rankings is available upon request.

TABLE X. TOP 10 MOST FREQUENTLY APPEARED HASHTAGS

Rank	Hashtag	Occurrence	Score
1	#job	17,983,887	6.54
2	#hiring	15,404,646	6.55
3	#careerarc	7,903,993	6.59
4	#jobs	3,537,381	6.51
5	#hospitality	2,568,649	6.57
6	#nursing	1,623,456	6.48
7	#veterans	1,491,045	6.60
8	#retail	1,410,858	6.61
9	#healthcare	1,209,446	6.50
10	#sales	845,894	6.59

Table 11 lists the top 10 most mentioned Twitter users and their overall sentiment score. As we can see, politicians dominate the list. The United States president Donald Trump (@realdonaldtrump) has been quoted more than 2.6 million times during our observation, which is nearly three times more than the second place, Hillary Clinton, his formal presidential campaign competitor. User @potus (President of the United States) ranks at number 4 and Kellyanne Conway (@kellyannepolls) holds the 9th place in the list. The rest of the list consists of mainly television news channels, such as

Fox News (@foxnews), CNN (@cnn), New York Times (@nytimes) and MSNBC (@msnbc). Sean Hannity (@seanhannity), the radio and television host from Fox News, also has been frequently mentioned by the Twitter community, ranking number 10 in the list. The only Twitter account appearing in the top 10 list that is not politics-related is YouTube (@youtube), which holds the 6th place with over 250,000 mentions. The second popular non-political account is @nfl (National Football League), who received less than 100,000 quotes with a rank of 20.

TABLE XI. TOP 10 MOST FREQUENTLY APPEARED USER MENTIONS

Rank	User Mention	Occurrence	Score
1	@realdonaldtrump	2,636,189	5.74
2	@hillaryclinton	707,982	5.75
3	@foxnews	594,875	5.61
4	@potus	533,544	5.83
5	@cnn	511,286	5.56
6	@youtube	256,558	6.17
7	@nytimes	215,899	5.72
8	@msnbc	206,649	5.67
9	@kellyannepolls	190,334	5.80
10	@seanhannity	162,287	5.84

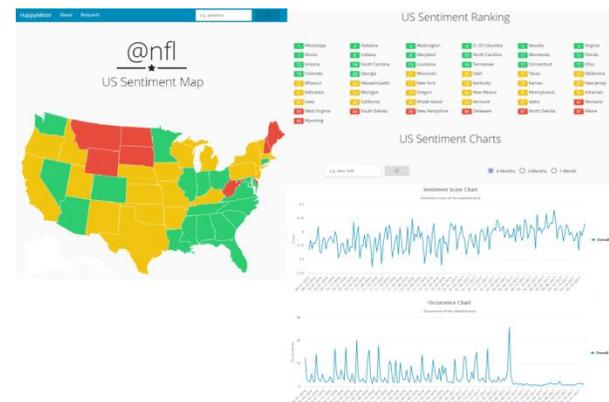


Fig. 6. HappyMeter dashboards of @nfl.

The analysis results of the system are visualized through a web-based graphical user interfaces available at www.happymeter.us. The dashboards display the Twitter sentiment map of a given term, sentiment rankings from the highest to the lowest among states in the contiguous U.S. and charts to reveal the temporal patterns. An example of the dashboards for Twitter user @nfl is shown in Fig. 6.

The sentiment map shows the average sentiment score of a selected term in each contiguous state. To better understand the geography of the public opinions regarding a subject, we applied Jenks natural breaks optimization to cluster the states into three classes. States with higher sentiment scores are classified as the (relatively) positive group. States with lower sentiment values are categorized as the (relatively) negative groups and the remaining states are part of the neutral class. On

the sentiment map, the green color is used to mark the positive group, while states belonging to the neutral and negative clusters are colored with yellow and red, respectively. The map is made interactively to display the sentiment score and polarity of a state when the mouse is hovered over. Fig. 7 shows an example of the sentiment map of Twitter user @realdonaldtrump.

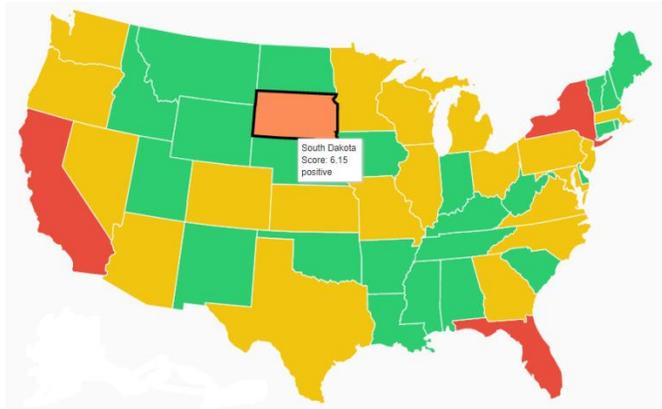


Fig. 7. Twitter sentiment map of @realdonaldtrump.

Besides the sentiment map and rankings, there are two charts on the dashboards. One is the sentiment score chart and the other is the occurrence chart. The sentiment score chart reveals the temporal patterns of a chosen term over an extended period of time. Users are able to choose if they want to see the overall pattern in the contiguous United States as a whole or the sentiment trend in a particular state. Fig. 8 shows an example sentiment score chart of hashtag #job in the state of New York. As one can see, the interactive chart displays the date and sentiment value on that specific day when a data point is selected.

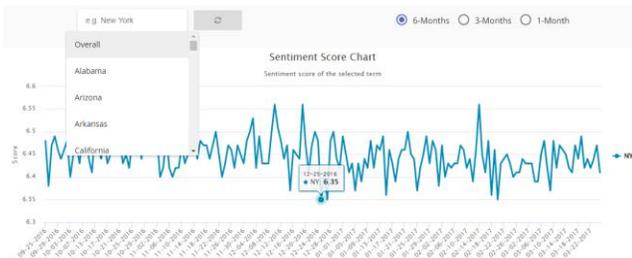


Fig. 8. Twitter sentiment score chart of #job in New York State.

The last component of the dashboards to introduce is the occurrence chart. In addition to sentiment values, the system also keeps track of the daily appearances of a Twitter term. Similar to the sentiment score chart, users are able to project the diagram on the overall contiguous U.S. as well as each state. Fig. 9 shows the Twitter occurrence chart of term “trump” in the whole contiguous U.S. The highest point showing in the figure represents a burst of tweet volume mentioning trump. It occurred on November 8, 2016, the day of the U.S presidential election.

The system uses Apache Storm to gather real-time tweets and manipulate data. Records of sentiment values are stored in a Redis database. The graphical web interface is developed in Python and JavaScript.

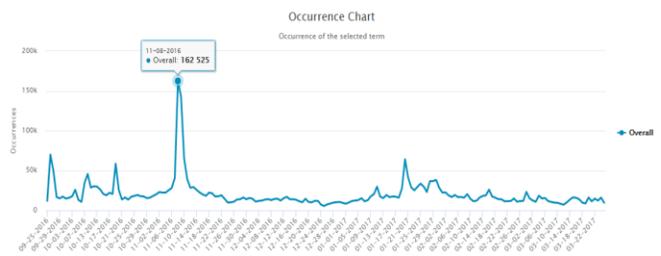


Fig. 9. Twitter occurrence chart of trump.

V. CONCLUSIONS AND FUTURE WORK

This paper presents HappyMeter, a real-time data processing infrastructure to evaluate public sentiment changes in the context of Twitter. The system examines every term tweeted in the contiguous United States and computes their sentiment scores in the range of 1 through 9. Daily analysis has been conducted throughout the contiguous U.S. as well as in each state. Over 40,000 terms extracted from more than 380 million tweets have been studied. These terms include words, hashtags and user mentions. The system shows the sentiment map and state rankings for each given term. Sentiment charts are automatically generated to reveal the changing pattern of the public sentiment towards a term in the nation or a selected state.

The study also investigates the amount of daily tweets published in a state, as well as its overall sentiment. Interesting findings have been conducted regarding word frequencies, terms with the highest and the lowest sentiment values and the most frequently tweeted words, hashtags and users. The complete analysis results can be made available upon request.

One limitation of the study is that the sentiment lexicon used in the experiment does not cover all the terms. Due to the informal language model on Twitter, new slangs, abbreviations and acronyms are created each day, many of which are Twitter-specific. For example, “twitterati” is a popular term in the Twitter community, which stands for popular users on Twitter. Future work includes designing a mechanism to regularly update the lexicon in order to expand the vocabulary of the dictionary.

Another limitation of the presented work is that context was not taken into account while calculating the sentiment value of a tweet. Our system determines the sentiment by averaging the sentiment score of each unigram. This method performs well in most cases, especially when the data set is at large. But there are times that an average is not able to reflect the true sentiment of a tweet. This is particularly the situation when a sentence is stated as double negative or laid out ironically. In the future, we plan to investigate sentiment scores of n-grams, specifically phrases, in order to achieve results with higher accuracy.

REFERENCES

- [1] C. C. Aggarwal, Social Network Data Analytics. Springer, 2011.
- [2] Twitter, <https://about.twitter.com/company>, retrieved on 07/28/2017.
- [3] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury, “Twitter power: tweets as electronic word of mouth,” Journal of the American Society for Information Science and Technology, vol. 60, pp. 1-20, 2009.

- [4] A. Archambault and J. Grudin, "A longitudinal study of Facebook, LinkedIn, & Twitter use," Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 2741-2750, 2012.
- [5] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes Twitter users: real-time event detection by social sensors," Proceedings of the 19th International Conference on World Wide Web, pp. 851-860, 2010.
- [6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy E. Kouloumpis, T. Wilson and J. D. Moore, "Twitter sentiment analysis: the good the bad and the OMG!" Proceedings of the Fifth International Conference on Weblogs and Social Media, pp. 164, 2011.
- [7] K. L. Liu, W. J. Li and M. Guo, "Emoticon smoothed language models for Twitter sentiment analysis," Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, pp. 1678-1684. 2012.
- [8] Twitter tweet length limit, <https://dev.twitter.com/basics/counting-characters>, retrieved on 07/28/2017.
- [9] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Foundations and Trends in Information Retrieval, volume 2, issue 1-2, pp. 1-135, 2008.
- [10] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Foundations and Trends in Information Retrieval, volume 2, issue 1-2, pp. 1-135, 2008.
- [11] A. Esuli and F. Sebastiani, "SentiWordNet: A publicly available lexical resource for opinion mining," In Proceedings of the 5th Conference on Language Resources and Evaluation, vol. 10, pp. 2200-2204, 2006.
- [12] J. Bollen, H. Mao and X. Zeng, "Twitter mood predicts the stock market," Journal of Computational Science, vol. 2, issue 1, pp. 1-8, 2011.
- [13] I. Vegas, T. Tian, and W. Xiong, "Characterizing the 2016 U.S. presidential campaign using Twitter data," Journal of Advanced Computer Science and Applications, vol. 7, no. 10, 2016.
- [14] T. Zeitzoff, "Using social media to measure conflict dynamics," Journal of Conflict Resolution, vol. 55, issue 6, pp. 938-969, 2011.
- [15] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," Technical report, 2009.
- [16] D. Davidov, O. Tsur and A. Rappoport, "Enhanced sentiment learning using twitter hashtags and smileys," Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pp. 241-249, 2010.
- [17] P. S. Dodds, K. D. Harris, I. M. Kloumann, C. A. Bliss and C. M. Danforth, "Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter," PLoS ONE, vol. 6, issue 12, pp. e26752, 2011.
- [18] L. Mitchell, M. R. Frank, K. D. Harris, P. S. Dodds and C. M. Danforth, "The geography of happiness: connecting Twitter sentiment and expression, demographics, and objective characteristics of place," PLoS ONE, vol. 8, issue 5, pp. e64417, 2013.
- [19] Twitter streaming API, <https://dev.twitter.com/streaming/overview>, retrieved on 07/28/2017.
- [20] Tweeting with Location, <https://dev.twitter.com/overview/terms/geo-developer-guidelines>, retrieved on 07/28/2017.
- [21] Population estimates from United States Census Bureau, <https://www.census.gov/data/tables/2016/demo/popest/state-total.html>, retrieved on 07/28/2017.
- [22] G. F. Jenks, "The data model concept in statistical mapping", International Yearbook of Cartography, vol. 7, pp. 186-190, 1967.

A Comparison between Chemical Reaction Optimization and Genetic Algorithms for Max Flow Problem

Mohammad Y. Khanafseh
King Abdulla II School for
Information and Technology
University of Jordan
Amman-Jordan

Ola M. Surakhi
King Abdulla II School for
Information and Technology
University of Jordan
Amman-Jordan

Ahmad Sharieh
King Abdulla II School for
Information and Technology
University of Jordan
Amman-Jordan

Azzam Sleit
King Abdulla II School for
Information and Technology
University of Jordan
Amman-Jordan

Abstract—This paper presents a comparison between the performance of Chemical Reaction Optimization algorithm and Genetic algorithm in solving maximum flow problem with the performance of Ford-Fulkerson algorithm in that. The algorithms have been implemented sequentially using JAVA programming language, and executed to find maximum flow problem using different network size. Ford-Fulkerson algorithm which is based on the idea of finding augmenting path is the most popular algorithm used to find maximum flow value but its time complexity is high. The main aim of this study is to determine which algorithm will give results closer to the Ford-Fulkerson results in less time and with the same degree of accuracy. The results showed that both algorithms can solve Max Flow problem with accuracy results close to Ford Fulkerson results, with a better performance achieved when using the genetic algorithm in term of time and accuracy.

Keywords—Chemical reaction optimization; Ford-Fulkerson algorithm; genetic algorithm; maximum flow problem

I. INTRODUCTION

A flow network is a weighted directed graph where each edge has a capacity and receives a flow [17]. The amount of flow on an edge cannot exceed the capacity of the edge. A flow must satisfy the restriction that the amount of flow into a node equals the amount of flow out of it, except when it is a source or sink. The maximum flow problem is to determine an optimal solution for the directed graph by finding the maximum flow from the source to the sink node [17].

Flow network can represent many real-life situations like a traffic in a road system, fluids in pipes, currents in an electrical circuit, or anything similar in which something travels through a network of nodes [15]. Due to its importance in many areas of applications such as computer science, engineering and operations research, the maximum flow problem has been extensively studied by many researchers using a variety of

methods [14], [15]. They include: a classic approach [8], maximal flow problem in layered network [3], the shortest augmenting path algorithm [10], and more [2], [4]-[6], [9], [11]-[13].

In this study, Chemical Reaction Optimization (CRO) algorithm and Genetic algorithm (GA) will be implemented and tested on the maximum flow problem. The goal is to determine which algorithm could give a better performance on finding a solution to the maximum flow problem near to the Ford-Fulkerson (FF) solution with less running time duration and same accuracy.

The rest of paper is organized as follows: Section 1 introduces the maximum flow problem. Section 2 presents some related works. Sections 3 and 4 explains the review the CRO and Genetic algorithms, respectively for solving the maximum flow problem. Section 5 shows the experimental and comparison results and Section 6 presents the conclusion and future works.

II. RELATED WORKS

A. Maximum Flow Problem

The flow network is a directed graph with two special vertices; the source and the sink [17]. Each edge in the graph connect two vertices and has a capacity and receives a flow that should be less than or equal to its capacity. In the operation research, a directed graph is called a network, the vertices are called nodes and the edges are called arcs [17].

A network is a directed graph $G = (V, E)$, with two special kinds of vertices are distinguished: a source S and a sink T , and every edge $e = (u,v) \in E$ has a non-negative, real-valued capacity $c(u,v)$. A flow network is an integer valued function f defined on the edges of G and satisfying that $0 \leq f(u,v) \leq c(u,v)$, for every Edge (u,v) in E .

For each edge (u,v) in E , the flow $f(u,v)$ is a real valued function that must satisfy the following three properties for all nodes u and v :

- 1) Capacity constraints: $f(u,v) \leq c(u,v)$. The flow along an edge cannot exceed its capacity.
- 2) Skew symmetry: $f(u,v) = -f(v,u)$. The flow from u to v must be the opposite of the net flow from v to u .
- 3) Flow conservation: $\sum_{v \in V} f(s, v) = 0$,
 $\sum_{v \in V} f(v, t) = 0$

unless $u = s$ or $u = t$. The flow to a node is zero, except for the source, which “produces” flow, and the sink, which “consumes” flow.

To achieve flow conservation, the flow into the node should be equal to the flow going out from the node. Also, the total amount of flow going from source s equals total amount of flow going into the sink t . The value of the flow is given by (1):

$$|f| = \sum_{v \in V} f(s, v) = \sum_{v \in V} f(v, t) \quad (1)$$

An example of the flow network with a source node s , sink node t and four additional nodes is shown in Fig. 1. The flow and the capacity is denoted by f/c . The network upholds skew symmetry and capacity constraints. The total amount of flow from s is 5, which is also the incoming flow to t .

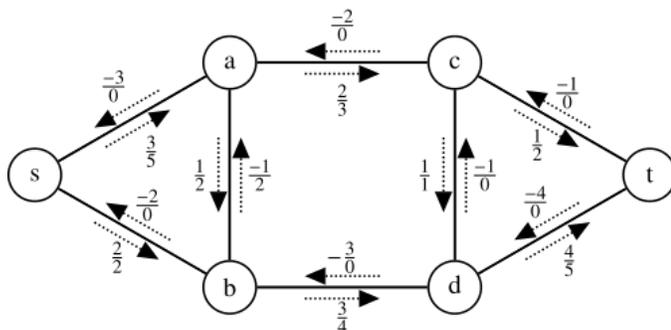


Fig. 1. A flow network with the flow and capacity.

The maximum flow problem involves finding a maximum flow through a single-source, single-sink flow network.

B. Ford-Fulkerson Algorithm

The Ford-Fulkerson method [1] (named for L. R. Ford, Jr. and D. R. Fulkerson) is the most popular algorithm used to compute the maximum flow in a flow network. The main idea of the algorithm is to find an augmenting path from the source to the sink with available capacity on all edges in the path to send flow along it. While there exist an augmenting path, you send a flow along it.

The Ford-Fulkerson algorithm has two main steps as shown in Fig. 2. The first is a labeling process that searches for a flow augmenting path i.e., a path from the source s to the sink t where the flow is less than the capacity along all forward arcs and the flow > 0 along all backward arcs. If this step finds a flow augmenting path, the second step changes the flow accordingly. Otherwise, no augmenting path exists then you get the maximum flow.

The runtime of Ford-Fulkerson is bounded by $O(Ef)$, where E is the number of edges in the graph and f is the maximum flow in the graph. We run a loop as long as there exists an augmenting path, each iteration of the loop takes $O(E)$ time to find an augmenting path, and increases the flow by at least 1 and an upper bound f , so the time complexity of the algorithm might not be a polynomial.

To decrease the computational time and get a better performance, many researches gave different algorithms.

```
1. Ford-Fulkerson algorithm:
2. initialize flow to 0
3. path = find Augmenting Path(G, s, t)
4. while path exists:
5. augment flow along path
6. G_f = create Residual Graph()
7. path = find Augmenting Path(G_f, s, t)
8. return flow
9. end algorithm
```

Fig. 2. Ford-Fulkerson algorithm.

Because of the importance of the maximum flow problem in many applications such as computer science, engineering researches, it has been extensively studied by many researchers using a variety of methods and techniques. A recent research in [15] applied to solve maximum flow problem using Chemical Reaction Optimization algorithm. The results showed a better performance with a complexity of $O(I E^2)$, for I iterations and E edges. Genetic algorithm was also used to solve maximum flow problem in [13]. The algorithm was implemented sequentially, and the fitness function is defined to reflect two characteristics: balancing vertices and the saturation rate of the flow. The performance of the algorithm depends on the population size and the number of generations needed to find the solution. In order to reduce running time of the algorithm, a parallel implementation was proposed in [18], the results showed a good enhancement in terms of the running time and system performance.

III. CHEMICAL REACTION OPTIMIZATION

Chemical reaction optimization (CRO), proposed in [6], is a chemical-reaction-inspired general-purpose meta-heuristic established for optimization and inspired by the nature of chemical reactions.

CRO refers to multi-agent algorithm which consists of different molecules, where each molecule has different attributes. Some of these attributes are important to CRO operations like molecular structure, kinetic energy (KE) and potential energy (PE) which refers to flow at the graph.

There are four main elementary reactions in CRO operation that take place at the CRO iteration, and are employed to manipulate the solution and distribute the energy through the molecule.

The molecule, here, can be described as container where these molecules are interacting with each other through this container with different forms as follows:

On-wall effective collision which refers to situation when different molecules collide with the wall of the container that contains different molecules. This collide converts the structure of molecule when collision happens new structure like this $\omega \rightarrow \omega'$.

- Decomposition interaction, this refers to situation when a molecule was collided with the wall of the container and then a molecule was divided into two parts $\omega \rightarrow \omega_1 + \omega_2$.
- Inter-molecular ineffective collision: this situation of collision between molecules happens when two molecules collide with each other and they bounce away, like this example when there are ω_1 and ω_2 where both collide with each other, then two new molecules ω'_1 and ω'_2 were produced from those two molecules which interact or collide with each other. This can be presented as: $\omega_1 + \omega_2 \rightarrow \omega'_1 + \omega'_2$.
- Synthesis: This makes an opposite of decomposition. Through this kind of interaction between molecules, two molecules hit with each other to produce new molecule. It can be implemented as $\omega_1 + \omega_2 \rightarrow \omega'$.

The CRO can be implemented to solve Maxflow problem. This needs to explore search space and to generate number of solutions and molecules to achieve optimal solution. Different solutions will be happening due to reaction between different selected molecules. Some of these solutions are near to desired solution and others were far away from it. After a selected number of iterations, the best solution will be taken from the list of these generated solutions.

In this paper, the CRO was applied to generate a possible solution for the Maxflow problem.

A. Cro-Maxflow Algorithm

The CRO-Maxflow implementation has three main phases, initialization, iteration and final phase.

- The initialization phase. In this phase, we define the graph as a source, sink node and a number of graph nodes. The nodes on graph are connected by edges where each edge has a weight value or capacity. From the source to the sink node, there are different flows that can be found, these flows refer to parent size and number of generated parent which depends on the value of parent size that had been specified through this step.

The first population will make the reaction with each other or with the wall of the container to generate other molecule or populations.

Some other basic CRO parameters like KE and molecule random number used as stopping criteria beside the use of the number of iteration that had been defined. The Maxflow value can be found in CRO using objective function, which can be computed using shortest augmenting path from source to sink. This value determines the Maxflow value which will be improved by the number of iterations. The objective function was used here as potential energy, other values were defined in the initialization step, such as α which refers to

decomposition threshold and β which refers to synthesis threshold.

- The iteration step, the goal of this step is to improve solution or objective function value. Most of the heuristic algorithms depend on the number of iteration to get a better solution. Through iteration step, potential energy or objective function was calculated for each iteration until reaching to iteration number, which was specified at previous step. Other collision happens based on the value of β which refers to the value generated randomly. This value is compared with molecule value. If β value is greater than molecule value, then one parent will be selected. Parent selection is important to know what kind of collision will happen when one parent is selected and this will give the ability for the decomposition reaction or on-wall-effective collision to occur; otherwise the other type of collision will occur.
- After selecting different molecule and calculating Potential energy for different iteration and the number of iterations reach max, the last step will start that refers to selection step. Through this step molecules with best value or largest value for Potential energy will be selected, this value present Maxflow result for the graph.

The pseudo code for the CRO-Maxflow algorithm is shown in Fig. 3 [15].

```
1. CRO-Maxflow algorithm:
2. Initialization phase
3. Set flow_network_size, C[i][j]: maximum capacity
4. ParentSize, iterationNumber,s: source node, t:
5. Sink node
6. HIT= 0
7. B =parentSize/2
8. A =parentSize/2
9. KE = parentSize/ 1.5
10. Generate molecule  $\epsilon \in [0,1]$ 
11. parentGeneratins (C[i][j], parentSize)
12. for (int i=1 to iterationNumber )//iteration phase
13.     Generate b  $\epsilon \in [0,1]$ 
14.     If b> Molecule then
15.         Randomly select one parent
16.         If (HIT >  $\alpha$ ) then
17.             Decomposition()
18.         Else
19.             OnWallIneffectiveCollision ()
20.         End if
21.
22.     Else
23.         Randomly select two molecules
24.         If (KE <= $\beta$  && parentSize >=2) then
25.             Synthesis ()
26.         Else if (parentSize >=2)
27.             IntermolecularIneffectiveCollision ()
28.         End if
29.     End if
30. HIT ++
31. KE --
32. Check for any new maximum solution
33. End for-loop //final stage
34. Return best solution found
35. End algorithm
```

Fig. 3. Pseudo-code for CRO-Maxflow algorithm.

IV. GENETIC ALGORITHM

A genetic algorithm (GA) is a method for solving complex optimization problems based on a natural selection process that mimics biological evolution. It can be used to design computer algorithms, to schedule tasks, and to solve other optimization problems.

Population can be viewed as binary bit strings. The initial values of this population are usually randomly generated and evaluated. The relation between the combination of ones or zeros in the population is found by an evaluation function that return a 'fitness' value for some bit string [7].

The three main operations of the genetic algorithm are: Reproduction (or Selection), Crossover and Mutation.

- **Reproduction:** use a fitness value to selects the best individuals and discards the bad ones from the population. The best individuals are those having more chances to survive in the next generation.
- **Crossover:** includes two steps. First, pairs of bit strings will be mated randomly to become the parents of two new bit strings. The second part consists of choosing a place (crossover site) in the bit string and exchanges all characters of the parents after that point. The process tries to artificially reproduce the mating process where the DNA of two parents determines the DNA for the newly born.
- **Mutation:** changes a 0 for a 1 and vice versa for the bits that can't be changed by the previous operations due to its absence from the generation, either by a random chance or because it has been discarded.
- Repeat the above steps until reaching the termination condition. The pseudo code of the Genetic algorithm is shown in Fig. 4 [16].

```
1. Genetic algorithm:
2. Initialize population
3. Evaluate population
4. While (!stopCondition) do
5.   Select the best-fit individuals for reproduction
6.   Breed new individuals through crossover and mutation
   operations
7.   Evaluate the individuals fitness of new individuals
8.   Replace least-fit population with new individuals
9. End algorithm
```

Fig. 4. Pseudo-code of Genetic algorithm [16].

A sequential implementation for Genetic algorithm has been applied to solve max flow optimization problems [14]. The number of iterations have been determined according to previous GA applications to achieve optimal or near optimal solutions.

This study aims to apply a sequential version of genetic algorithm on the max flow problem again. In order to get better results and performance, one of the possible solutions here is to reduce the number of generations required to determine the solutions.

A. Genetic-Maxflow ALGORITHM

Maxflow problem consist of graph with number of nodes and edges between these nodes. Each edge has specific weight or capacity which is saved in matrix called $C[i,j]$. Based on this value, different optimistic path was selected for each iteration and solution will be build based on this paths between source and sink. These solutions in the Genetic algorithm refer to population and will be saved in population matrix.

For a graph, G , with n vertices and m edges; G is represented by the flow capacity matrix, $C = [c_{ij}]$, $i, j = 1, n$. Each solution is represented by a flow matrix $F = [f_{ij}]$, $i, j = 1, n$. The initial flow was generated randomly.

The first step of Genetic-Maxflow refers to initialization step. In this step, number of iteration, population size and mutation ratio values must be defined. After this step, the selection step must be implemented to select best solution from different solutions. This is implemented based on the fitness function to find path between source and sink. Based on the value of Fitness Function, it will select some ratio for all matrix. Next the crossover will be implemented between different solutions. The first population or solution have best Maxflow result, then generate a new population in new matrix Pop1 with good results for Maxflow. After that step, mutation will be implemented to change some of solution to different one from parent solution which depends on doing a crossover process. This mutation based on specific ratio. This ration must be small between 0.01 and 0.025 of all population to change population. New solutions for Maxflow problem was generated in new Matrix Pop2. This solution for one iteration. Genetic algorithm is heuristic algorithm where number of iterations was important to achieve enhancement of solution at each iteration. These steps were repeated at each iteration in the proposed solution and best population will be selected based on population size which specified at the beginning of the algorithm.

V. EXPERIMENTAL RESULTS AND COMPARISONS

In this study, a sequential implementation for FF, CRO and GA is applied to find maximum flow problem with a different network size. The algorithms FF, CRO and GA were tested using Intel core I7-3632QM CPU2.20GHz, 8GB of RAM and windows 7 64 bits. The application programs were written using java and executed on Net-Beans IDE 8.1. The implementations were done over different network size started from 50 nodes until 6030 nodes, with different number of parents and different number of iterations, in order to achieve the best solution which is near to Ford-Fulkerson one, as we will see from results. Each experiment was repeated 10 times, and the average results were calculated.

The first comparison was made between the Ford Fulkerson algorithm and the CRO algorithm on the Maxflow problem based on the time needs to calculate Maxflow and the accuracy level of the proposed solution for a graph when using different number of nodes, and different number of iterations. Each experiment was executed for 10 times and the average result was calculated.

Different results for CRO-Maxflow were conducted. These results were optimized until reaching to accuracy level near to the Ford-Fulkerson with less time than the Ford-Fulkerson algorithm running time and using different network size. A part of comparison between the program results with optimal Ford Fulkerson results are shown in Table 1. Runtime in Table 1 is in milliseconds. Table 1 shows the results for the implementation of CRO-Maxflow and Ford Fulkerson Maxflow.

The number of nodes plays an important role in Ford-Fulkerson solution. The CRO algorithm took less time to find maxflow than Ford-Fulkerson for large number of nodes. The four main steps for CRO were repeated based on the number of iterations needed to reach to a solution near the Ford-Fulkerson according to the accuracy when finding Maxflow. The number of iterations have some limitations when increasing it to be more than some specific value it will consumes the memory efficiency.

TABLE. I. RESULTS FOR IMPLEMENTING CRO-MAXFLOW AND FF MAXFLOW, WITH NUMBER OF NODES= 50 TO 6030

Size	CRO Time	FF Time	FF result	CRO result	Quality
3550	8.3	14.3	14.2	9.7	0.683099
3602	14.3	15.1	13.4	13.4	1
3654	9.2	15.2	14.7	13.2	0.897959
3706	6.3	15.7	37.2	33.2	0.892473
3758	13.9	16.3	22.5	22.3	0.991111
3810	16.5	16.5	24.3	19.7	0.8107
3862	11.5	17	22.8	21.8	0.95614
3914	17.1	17.5	17.1	14.7	0.859649
3966	15.4	18.1	22.8	16	0.701754
4018	10.7	18.3	17.4	16.7	0.95977
4070	8.2	19	19.1	15.4	0.806283
4122	19	20.1	30.6	27.5	0.898693
4226	22.6	20.1	22.9	22.5	0.982533
4278	15.1	20.8	22.2	19.4	0.873874
4330	20.3	21.3	16.5	15.3	0.927273
4486	13.5	22.9	23.2	14.4	0.62069
4590	20.6	24.1	13.6	12.6	0.926471
4746	25.5	25.6	20.9	14.3	0.684211
5162	11.7	29.9	16.4	12.4	0.756098
5214	26.4	30.6	11.6	6.7	0.577586
5422	20.2	33.4	22.2	18.8	0.846847
5474	13	34.3	20.3	20.3	1
5942	31.2	39.8	19.6	18.6	0.94898
5990	18.8	40.6	18.8	18.4	0.978723
6098	25.5	41.7	21.3	19	0.892019
6306	46	49	29	28.5	0.982759

From the result of 1000 nodes and 10 iterations, we can say that the CRO-Maxflow gives less time for execution than Optimal of Ford-Fulkerson Maxflow with accuracy rate near to 90% to Ford-Fulkerson results. This gives a good enhancement when use a heuristic algorithm to solve Maxflow problem with less time and same level of accuracy, which we implemented through the proposed solution. To compare results for Maxflow problem using CRO and Ford-Fulkerson solution, Fig. 5 presents the relation between time needed for CRO and Ford-Fulkerson Solution at same data sets.

Fig. 5 shows that run time needed to solve Maxflow using CRO is less than that needed to solve the same problem on the same dataset and using same machine and environment for optimal Ford Fulkerson algorithm.

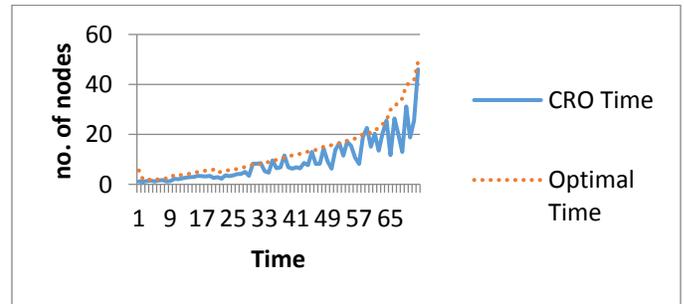


Fig. 5. Execution time of CRO and FF algorithm to solve max flow problem.

Fig. 6 shows the accuracy of CRO-Maxflow problem and Ford-Fulkerson Maxflow. This results for different number of nodes, started from 50 nodes to 1000 nodes. The proposed solution to solve Maxflow has accuracy near to Ford-Fulkerson accuracy, which is a good achievement.

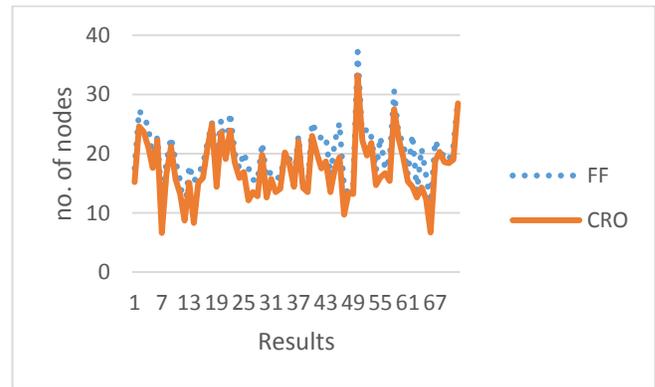


Fig. 6. Accuracy results for calculating Maxflow using CRO and FF.

The second comparison was between Genetic-Maxflow and FF Maxflow.

In [14], the authors applied GA to find the maximum flow from the source to sink in a weighted directed graph. The experiment was run for various graphs with different number of vertices. The results showed that Genetic algorithm found an optimal or near optimal solution for the maximum flow problem, with a reasonable number of iterations compared to other previous GA applications [14].

TABLE. II. RESULTS FOR IMPLEMENTING GENETIC-MAXFLOW AND FF MAXFLOW, FOR DIFFERENT NUMBER OF NODES STARTED FROM 50 NOD UNTIL 1986

Size	GA Time	FF Time	FF result	GA result
4590	12.9	23.5	13.6	12.6
4642	12.8	23.8	20.2	18.7
4694	13.5	24.4	15.7	15.6
4746	12.9	24.7	20.9	20.7
4798	13.8	25.4	19.3	17.5
4850	14.8	26.9	24.6	20.6
4902	15.3	26.8	17.1	14
4954	14.7	27.2	14.4	13.1
5006	14.3	27.7	27.4	27.4
5058	14.7	28.2	17.9	16.8
5110	14.6	28.4	17.1	16.3
5162	14.5	29.4	16.4	15.1
5214	14.8	29.9	11.6	8.1
5266	14.7	30.3	19.1	18.9
5318	14.1	30.8	26.1	25.5
5370	16.1	31.7	19.3	14.8
5422	16.9	32.6	22.2	20.5
5474	17.2	33.1	20.3	19.9
5526	17.9	33.8	20.9	17.8
5578	17.2	34.2	28.1	26
5630	17.6	34.7	22	22
5682	17.5	35.3	15.7	15.3
5734	17.7	36.2	16.6	14
5786	18.3	37.9	12.7	11.7
5838	18.1	37.7	14.6	13.3
5890	19.7	38	12.7	10.7
5942	18.3	38.9	19.6	19.6
5994	18.9	39.4	22	12.4
6046	20.5	39.8	21	20
6098	19.5	40.7	21.3	19.7
6150	21.7	41.6	20.3	18.2
6202	20.2	41.7	25.1	20.9
6254	21	42.6	19.9	18.4
6306	21.4	43.5	29	26.9
6358	20.9	44	19.7	19.3

We implemented the GA to solve Maxflow problem and compare results. The implementation of the three algorithms FF, CRO and GA were tested on the same data and the same environment. Both objective functions which we used to calculate Maxflow in CRO, are the same as Fitness functions which we used in GA.

When implementing GA on Maxflow problem, the population is selected first then we start with genetic steps from selection of best population of all populations to cross over process. Through this process, we did cross over for two selected solutions from selection process then we implemented the mutation step in order to make changes for the generated population. After the cross over step, the mutation step was checked to make sure it does not exceed the range from 0.01- 0.025 for all populations. The process was done randomly.

These steps of GA were repeated based on the number of iterations, which specified to reach solutions near to FF one in accuracy level to find Maxflow. But generating number has some limitation. If you increase this number more than specific value, it will affect the memory space.

The GA was implemented at the same environment that used to implement CRO-Maxflow. We compared results of GA with results of optimal FF for the same experiment. Each experiment with specific size of network was repeated 10 times and average result for this repeated time was calculated with specific number of iterations similar to process which is done in CRO-Maxflow experiment. Table 2 shows part of the results for the time needed for Genetic-Maxflow less than time which needs to solve Maxflow problem using FF solution, at level of accuracy near to FF level.

The results show that GA reach to accuracy near to FF with time less than time needs to solve maxflow by FF algorithm as shown in Fig. 7 and 8, respectively.

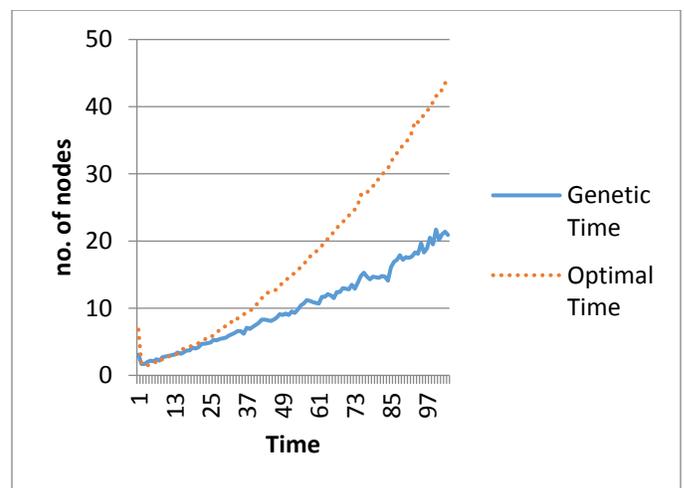


Fig. 7. Relation between time needs for Maxflow Problem using GA and FF, at same node.

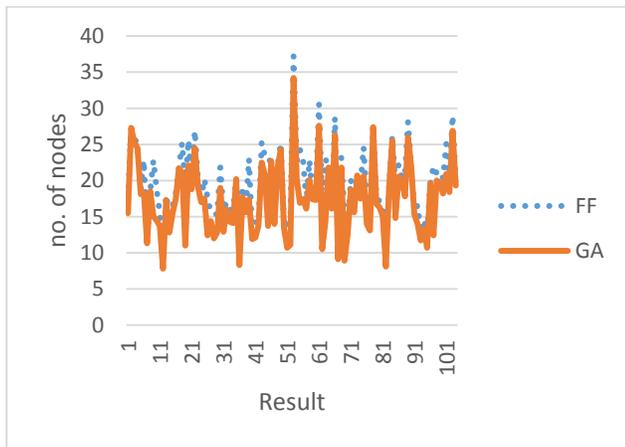


Fig. 8. Accuracy results for calculate Maxflow using GA and FF.

The results from using CRO-Maxflow were compared with the Genetic-Maxflow as shown in Table 3 and Fig. 9.

TABLE III. CRO-MAXFLOW VS GENETIC-MAXFLOW

Size	CRO Time	FF Time	GA Time	CRO Result	GA Result
3550	8.3	14.3	14.7	9.7	16.8
3602	14.3	15.1	14.6	13.4	16.3
3654	9.2	15.2	14.5	13.2	15.1
3706	6.3	15.7	14.8	33.2	8.1
3758	13.9	16.3	14.7	22.3	18.9
3810	16.5	16.5	14.1	19.7	25.5
3862	11.5	17	16.1	21.8	14.8
3914	17.1	17.5	16.9	14.7	20.5
3966	15.4	18.1	17.2	16	19.9
4018	10.7	18.3	17.9	16.7	17.8
4070	8.2	19	17.2	15.4	26
4122	19	20.1	17.6	27.5	22
4226	22.6	20.1	17.5	22.5	15.3
4278	15.1	20.8	17.7	19.4	14
4330	20.3	21.3	18.3	15.3	11.7
4486	13.5	22.9	18.1	14.4	13.3
4590	20.6	24.1	19.7	12.6	10.7
4746	25.5	25.6	18.3	14.3	19.6
5162	11.7	29.9	18.9	12.4	12.4
5214	26.4	30.6	20.5	6.7	20
5422	20.2	33.4	19.5	18.8	19.7
5474	13	34.3	21.7	20.3	18.2
5942	31.2	39.8	20.2	18.6	20.9
5990	18.8	40.6	21	18.4	18.4
6098	25.5	41.7	21.4	19	26.9
6306	46	49	20.9	28.5	19.3

The results show that GA took less time with same level of accuracy as CRO algorithm for the same network size and same number of iterations. As the performance of the GA depends on doing the cross over and mutation steps by each iteration. While CRO algorithm depends on different number of collisions that can be happened between different molecules which compared with molecule value to determine number of molecules that will be selected on interaction or collision process. This will need more time to achieve it. Through the experiment of implementation, both algorithms were implemented using java programming language. Based on the results for each step, calculate the time needs for each step for both experiments GA and CRO to decide which step consume most of the time. According to that optimize that step which spent most of execution time in CRO and GA to achieve same level of enhancement on both algorithms.

The copy matrix step consumes most of the time, some optimization steps was done to enhance time results like allocation and de allocation for matrix when one matrix for a graph was deleted, new matrix for a new graph was build and this consumes time and memory through the implementation, because of that we reuse the matrix by using a java object pool feature that allow to use the same matrix with replacing the nodes for the old matrix with a new matrix results.

Other technical enhancement which has been done for both the CRO and GA deals with the connected edges. As the matrix presents a graph with nodes and edges, we worked with the submatrix that contains ones instead of dealing with the whole matrix with its connected and disconnected edges.

Both algorithms, CRO and GA gave better execution time than FF algorithm for large network size. When this number becomes very large, the CRO and GA keep on the same level of efficiency in terms of accuracy and velocity.

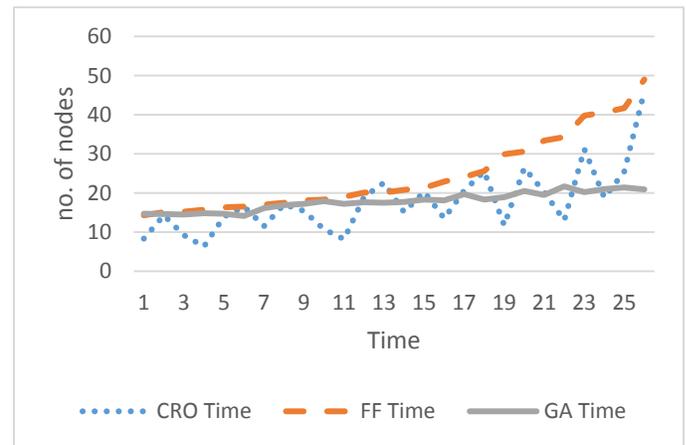


Fig. 9. Relation between time needs for Maxflow problem using CRO, GA and FF, at same node.

VI. CONCLUSIONS AND FUTURE WORKS

In this paper, CRO and Genetic algorithms were implemented sequentially on Intel core I7-3632QM CPU2.20GHz, 8GB of RAM and windows 7 64 bits to solve the Maxflow problem. The application programs were written using java and executed on Net-Beans IDE 8.1. The implementations were done over different network size started

from 50 nodes until 6030 nodes, with different number of parents and different number of iterations, in order to achieve the best solution which is near to Ford-Fulkerson one.

The results show that GA and CRO both can solve Max Flow problem with accuracy result near to FF results, with better performance achieved when using the genetic algorithm in term of time and accuracy.

For future work, we need to implement both Genetic and CRO on parallel to solve max flow problem by using a super computer to test the amount of enhancement on time with large number of network size.

REFERENCES

- [1] FORD.L.R. AND D. R. FULKERSON 1956. Maximal Flow Through a Network. Can. J. Math. 8,399-404
- [2] Edmonds, Jack; Karp, Richard M. (1972). "Theoretical improvements in algorithmic efficiency for network flow problems". *Journal of the ACM. Association for Computing Machinery*. 19 (2): 248–264. doi:10.1145/321694.321699.
- [3] Dinic, E. A. (1970). "Algorithm for solution of a problem of maximum flow in a network with power estimation". *Soviet Math. Doklady*. Doklady. 11: 1277–1280.
- [4] KARZANOVA.V. 1974. Determining the Maximal Row in a Network by the Method of Preflows. *Soviet Math.Dokl*.15,434-437.
- [5] Baumann N, Skutella M (2006) Solving evacuation problems efficiently—earliest arrival flows with multiple sources. In: 47th annual IEEE symposium on foundations of computer science (FOCS'06), pp 399–410
- [6] A. Y. S. Lam and V. O. K. Li, "Chemical-reaction-inspired metaheuristic for optimization," *IEEE Trans. Evol. Comput.*, vol. 14, no. 3, pp. 381– 399, 2010
- [7] D. Arjona, "A hybrid artificial neural network/genetic algorithm approach to on-line switching operations for the optimization of electrical power systems", appears in "Energy Conversion Engineering Conference", pp 2286 – 2290, vol.4, Aug 1996.
- [8] L.R. Ford, Jr. and D.R. Fulkerson, *Flows in Networks*, Princeton, NJ: Princeton University Press, 1962
- [9] Goldberg, A.V., Tarjan, R.E., "A new approach to the maximum flow problem". *Proc. 18th ACM STOC*, 1986, pp. 136-146
- [10] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin., *Network Flows: Theory, Algorithms, and Applications*, Prentice Hall, 1993.
- [11] G. Mazzoni, S. Pallottino and M.G. Scutella, "The Maximum Flow Problem: A Max-Preflow Approach," *European Journal of Operations Research*, vol. 53, pp. 257-278, 1991
- [12] J. B. Orlin. Max flows in $o(nm)$ time, or better. In *STOC'13: Proceedings of the 45th Annual ACM Symposium on the Theory of Computing*, 2013, pp.765–774.
- [13] V. King, S. Rao, and R. Tarjan, "A faster deterministic maximum flow algorithm", In *Proceedings of the 8th Annual ACM–SIAM Symposium on Discrete Algorithms*, 1992, pp. 157–164.
- [14] Munakata, T. and Hashier, D.J. "A genetic algorithm applied to the maximum flow problem", *Proc. 5thInt. Conf. Genetic Algorithms*, 1993, pp. 488-493
- [15] R.Barham, A.Sharieh, A.Sliet. "Chemical Reaction Optimization for Max Flow Problem", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 8, 2016
- [16] MahmoodA.Rashid, M.A.Hakim Newton,Md. Tamjidul Hoque, Abdul Sattar," Mixing Energy Models in Genetic Algorithms for On-Lattice Protein Structure Prediction", *Hindawi Publishing Corporation Bio Med Research International*, Volume2013,ArticleID924137,15pages
- [17] Zhipeng Jiang, Xiaodong Hu, and Suixiang Gao, "A Parallel Ford-Fulkerson Algorithm For Maximum Flow Problem", *The Allen Institute for Artificial Intelligence, SemanticScholar*.
- [18] Ola M.Surakhi, Mohammad Qatawneh, Hussein A.. "A Parallel Genetic Algorithm for Maximum Flow Problem", *International Journal of Advanced Computer Science and Applications*, 2017.

Meteonowcasting using Deep Learning Architecture

Sanam Narejo and Eros Pasero

Department of Electronics and Telecommunication
Politecnico Di Torino
Torino, Italy

Abstract—The area of deep learning has enjoyed a resurgence on its peak, in almost every field of interest. Weather forecasting is a complicated and one of the most challenging tasks that includes observing and processing huge amount of data. The present paper proposes an effort to apply deep learning approach for the prediction of weather parameters such as temperature, pressure and humidity of a particular site. The implemented predictive models are based on Deep Belief Network (DBN) and Restricted Boltzmann Machine (RBM). Initially, each model is trained layer by layer in an unsupervised manner to learn the non-linear hierarchical features from the input distribution of dataset. Subsequently, each model is re-trained globally in supervised manner with an output layer to predict the appropriate output. The obtained results are encouraging. It is found that the feature based forecasting model can make predictions with high degree of accuracy. This implies that the model can be suitably adapted for making longer forecasts over larger geographical areas.

Keywords—Deep learning architectures; deep belief network; time series prediction; weather nowcasting

I. INTRODUCTION

In the few last decades, the use of machine learning has spread rapidly beyond the limitations of computer science field. Machine learning is extensive and so pervasive today that one probably uses it dozens of times a day without knowing it. Deep learning is a category of machine learning models. Recently, the area of Deep learning has enjoyed a resurgence on its peak, in almost every field of interest. Deep learning architectures include several models such as Deep Neural Networks (DNNs), Convolutional Neural networks (CNNs), Recurrent Neural Network (RNN), Deep Belief Network (DBN), Recursive NN and more [1]. Present literature suggests that these architectures are being applied widely and it has produced state of the art results on various problems in major fields like computer vision, automatic speech recognition, natural language processing, audio recognition and bioinformatics.

Weather forecasting is a complex time series forecasting problem. This is due to its dynamic and non-linear chaotic behaviour [2], [3]. It is an arduous skill that involves observing and processing vast amounts of data. In literature, several approaches have been proposed in order to deal with accurate time series forecasting problem. Artificial Neural Network (ANN) is known to be one of the successfully developed models widely used in solving many time series forecasting and prediction problem in diversity of applications [4]-[6]. ANN is general, flexible, non-linear tool capable of approximating any arbitrary function [7]. ANN is based on a

collection of connected units called artificial neurons. Neurons receive input, change their internal state (activation) according to that input, and produce output depending on the input and activation. The network forms by connecting the output of certain neurons to the input of other neurons forming a directed, weighted graph. The weights as well as the functions that compute the activation can be modified by a process called learning.

Deep learning is an application of ANN to learning tasks that contain more than one hidden layer. Deep ANNs contain numerous levels of non-linearities depending on the depth of hidden layers. The deep hierarchical architecture allows them to efficiently represent highly nonlinear patterns and highly varying functional abstractions. Although, it was not clear how to train such deep networks, as the random initialization of network parameters appears to often get stuck in poor solutions [8].

Nowcasting is defined as the prediction of the present, the very near future. The term is a contraction for now and forecasting. This term has been used for a long time in meteorology. In other words, nowcasting is a strategy to perform very short range forecasting. This procedure maps the current weather and then uses an estimate of its speed and direction of movement to forecast the weather a short period ahead. A critical aspect regarding the time series prediction problem is to capture the temporal relationship [9] and underlying structure residing in given input series data [10].

This research work is inspired by the recent advances in the realm of deep learning methods. In this work, a predictive ANN model based on the deep learning is obtained by firstly training the layers of Restricted Boltzmann Machine (RBM) in an unsupervised fashion. Subsequently, stacking those trained RBMs to create Deep Belief Network (DBN). Afterwards, the DBM is finally trained in supervised manner to predict the parameters of weather, i.e., temperature, pressure and humidity. For each parameter, a separate predictive model is implemented and trained. The accuracy of predictions confirms the promising performance of Deep learning algorithm specifically DBN. The data used in this study is sampled every 15 minutes by means of a traditional metrological station. The approach proposed here is basically a local level and it is restricted to a particular geographical area. However, it can be further extended and possible to apply on global level.

A literature review is presented in Section II. Section III deals with the illustration of adopted research methodology and experimental setup. Section IV presents the obtained results and discussion in detail. The paper ends with conclusions and

suggestions for possible future research as specified in Section V.

II. LITERATURE REVIEW

Keeping our focus particularly in the realm of time series forecasting and prediction, we shed some light by presenting some research related with time series forecasting with deep learning methods. As aforementioned in previous section, various different architectures come under the umbrella of “Deep Learning models”.

The initial research advocates that DBN models are efficient at the classification and prediction tasks but, it actually lacks the efficacy to model temporal sequences. The authors in [11] have reported that recurrent neural networks performed drastically better on energy load forecasting using dataset from kaggle competition. They further argued that the greedy layerwise trained feed forward neural networks with stacked AutoEncoders obtained discouraging results with no significant performance gain but added complexity. As it is well known fact that feed forward networks are deficient in capturing temporal dynamics. Thus, these models are unable to access the past terms while modelling the underlying structure.

On the contrary, the Stacked AutoEncoders are also deployed in [12] in order to learn feature representations for weather forecasting. The empirical evaluations are further compared between models using raw features and models using learned representations as features. The obtained results in the above mentioned study prove that Deep Neural Network (DNN) is capable enough to provide better feature space for highly varying and non-stationary data like weather data series of temperature, pressure and wind speed. Related study has been provided in [13] for short term wind prediction and in [14] for load forecasts.

A predictive model has been proposed for time series data in [15] by using a DBN with RBM. Additionally, the performance of the proposed model is evaluated on data of CATS benchmark [16], [17] and chaotic time series. According to the experimental outcomes, it was confirmed that the proposed prediction model, DBN with RBM using pre-training and fine-tuning learning algorithm and PSO structure decision, performed better than traditional models although it is unable to beat the best of IJCNN 2004 competition model. The authors in [10] proposed a novel hybrid approach for multi-step ahead time series forecasting by using deep learning and Nonlinear Autoregressive Neural Network.

The research in [18] presents another novel hybrid model with discriminative and generative components for spatio-temporal inference about weather. Furthermore, a data driven kernel is implemented that forms the predictions according to physical laws. A detailed review of unsupervised feature learning and deep learning for time series modelling has been conferred in [19]. The article presents the detailed review on time series analysis and temporal sequence modelling using deep architectures. However, according to our observation, the study was found to be deficit, as far as the overview regarding time series forecasting with deeper models is concerned. The CRBM was introduced in the family of Deep Learning by applying it to capture the activity related to human motion [20].

Similarly, the other variants of RBM, for example, Temporal RBM [21], [22] and Gated RBM [23], [24] have also been introduced. With the exception to these models, another deep architecture producing outstanding state-of-the-art research is Convolutional Neural network (CNN). These models are of high interest specifically for image data or high dimensional time series data. Apart from being stand alone, convolution has also been applied as Convolutional RBM [25], [26] and Convolutional AutoEncoders [27]-[29].

III. MATERIALS AND METHODS

A. Meteorological Nowcasting

The activity conducted in this current work is related with our earlier work done in [30]. In above mentioned research, a statistical neural system was used to “nowcast” meteorological data measured by a weather station deployed at Neuronica laboratory, Politecnico Di Torino, see Fig. 1. For further details please refer [30]-[32]. By utilizing the same resources of meteorological data, i.e., “NEMEFO”, we have performed weather parameter prediction by using deep learning algorithm.

In our previous work [33], we presented predictive models for internet traffic prediction by using DBN. We explored the useful strategies for topological architecture for deeper networks and we also did validation on standard benchmark time series. Keeping all those aspects in mind, which we earned for successful training of deep models, we were motivated to attempt some more case studies for real time data sets.



Fig. 1. METEO weather station.

Weather forecasting has been one of the most challenging problems around the world for more than a half century. However, nowcasting is weather forecasting on a very short

term. It makes difficult for traditional mathematical or statistical models to adapt irregular patterns of data which cannot be written in form of function, or deduced from a formula. In response to this, we developed and trained some more DBN models for nowcasting the air temperature, relative humidity, and air pressure for the next future value. The pictorial view of our contributed activity is presented in Fig. 2.

The standard training of deeper models through gradient back propagation appears to be difficult until Hinton gave a breakthrough in 2006. The standard training strategies attempts to allocate the parameters in the region of parameter space that generalize poorly. This has been shown practically in number of studies [34].

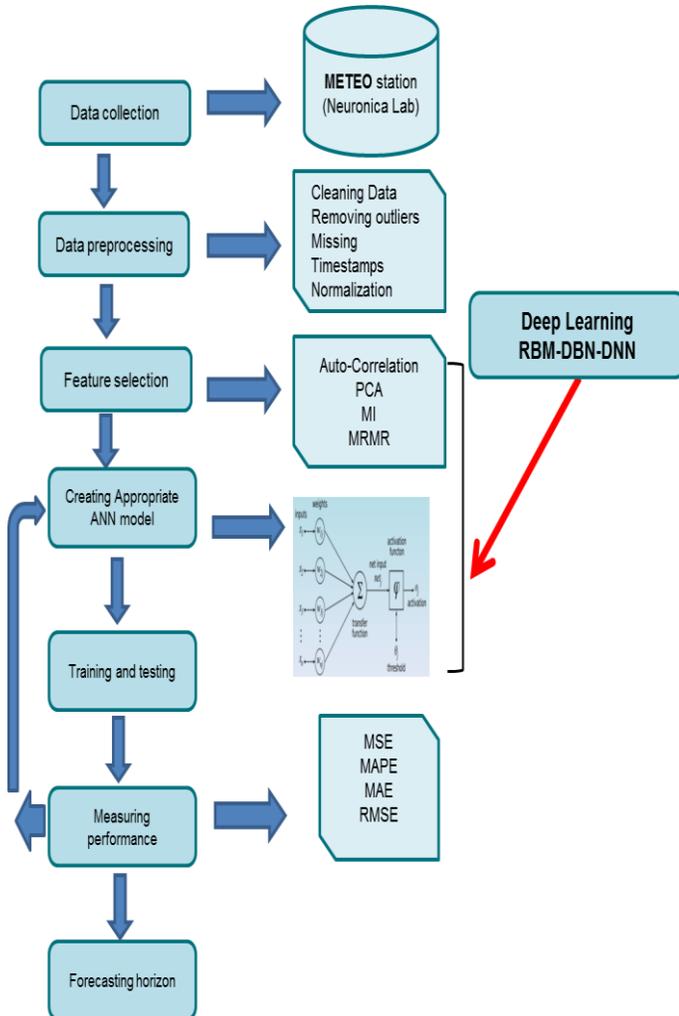


Fig. 2. Illustration of contributed activity for weather nowcasting.

B. METEO Weather Station and NEMEF0

The Weather forecasting is a complicated and one of the most challenging tasks that includes observing and processing huge amount of data. NEMEF0 stands for NEural METeological FOrecasts. It is basically a software tool connected to Meteo weather station at Neuronica laboratory, which samples meteo data after every 15 minutes. Meteo station contains following recorded weather data.

- Air temperature
- Relative Humidity
- Air Pressure
- Solar radiation
- Wind velocity
- Precipitation
- Wind Direction
- Corresponding Date and Time.

The sensors at Meteo station provide a new recording after every fifteen minutes. The dataset was downloaded from weather station. It contains the records from 4 October 2010 to 3 September 2015. However, the predictive models are only implemented for nowcasting of Air temperature, Relative humidity, and Air pressure as mentioned previously. The data recorded through sensors may have noise, some of missing samples and unwanted frequency fluctuations. In order to detect the outliers and to remove sensor noise, some of the pre-processing in the form of filtering has been done on the data prior to considering it as an input set. Subsequently, features are extracted individually for each case to be predicted for the next sample.

C. Air Temperature Prediction

Temperature is one of the most common parameters for an accurate weather forecast. The unit of recorded temperature at Meteo station is Celsius. It is one of the known facts that temperatures gets effected by season. For example, in extreme summer we face scorching heat by sun and in winter we experience freezing cold temperature. Consequently, the recorded temperature has maximum and minimum values. Apart from this the second effecting parameter could be the particular hours in a day; at that time the air temperature can possibly vary, i.e., the day time hours and the night time hours. Since temperature is clearly dependent on the season and hour, these two attributes have been taken into account in order to reach a right nowcasting. Month and Hour have been computed using the date of the record and they have been used as predictors. They have been preprocessed in order to transform them as sinusoidal features as shown in Fig. 3 and 4, respectively.

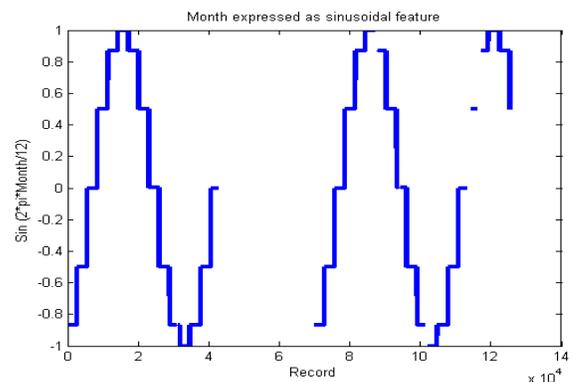


Fig. 3. Recorded months converted into sine waveform.

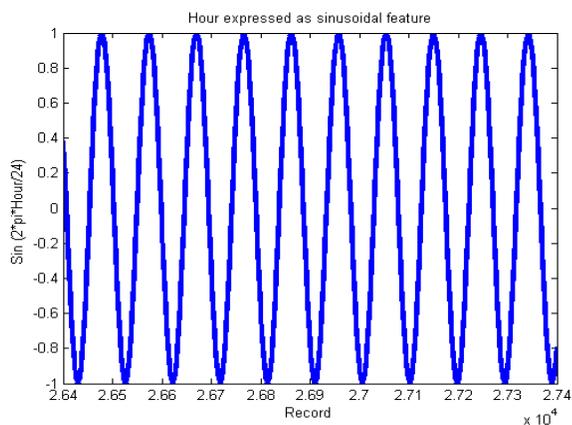


Fig. 4. Recorded hours converted into sine waveform.

In order to predict temperature at time $t+1$, the final input feature set contains particular values of month, hour, temperature at interval (t) and temperature at ($t-1$). Although, before taking the temperature as attribute, we have done pre-processing to reduce the noisy fluctuations from raw sensor data which includes Butterworth lowpass filter with order 2 and 0.11 Cutoff frequency in mHz. The difference between actual and filtered data can be seen in Fig. 5. Moreover, the identified outliers in series were replaced by NAN. Additionally, the interpolation method was applied to cover the missing samples where sensor was unable to record the samples.

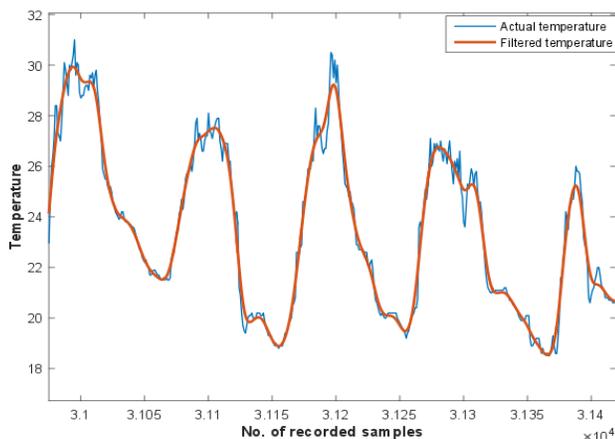


Fig. 5. Actual and filtered temperature series.

The most noticeable step is that the input data as well as labels were normalized in the range of (0,1). Apparently because, we have used RBM which deals with binary hidden and visible units. The detail explanation related to this has already been demonstrated [34]. The training data was selected from October 2010 to March 2014. As aforementioned, the input data set consist of five attributes. According to this, the input layer was based on five nodes, whereas, the output layer with one output neuron. In order to select the number of hidden layers and the size of hidden units in each layer, we preferred random search method. In response to this, we developed and trained several architectures. The selection of hyper-parameter for this model and the next upcoming models

presented in Section III-D and III-E was based on our earlier hypothesis which provided great support to select better model. The best predictive model for temperature prediction, which was initially pretrained layer by layer with total four hidden layers was with the dimension (500-200-100-10). After training each layer separately the model was trained globally by adding an output layer with temperature labels. The architecture of model for temperature prediction is illustrated in Fig. 6. The results are further discussed in Section IV.

D. Relative Humidity Prediction

Humidity is a quantity representing the amount of water vapour in the atmosphere. However, relative humidity depends on the temperature and the pressure of the system of interest. The variation of the temperature, which has a larger variability, depends on the hour and season. Apart from considering the above mentioned attributes, we applied Mutual Information Criteria (MIC) to find the best correlations in between of weather parameters. This further confirmed the attributes selection as mentioned below. The correlations between features computed via MIC are presented in Table 1.

To further explain the feature selection procedures through MIC assume a target class labelled as c . For selecting the features with the highest relevance of attributes to the target class c is crucial. Relevance is usually characterized in terms of correlation or mutual information, of which the latter is one of the widely used measures to define dependency of variables.

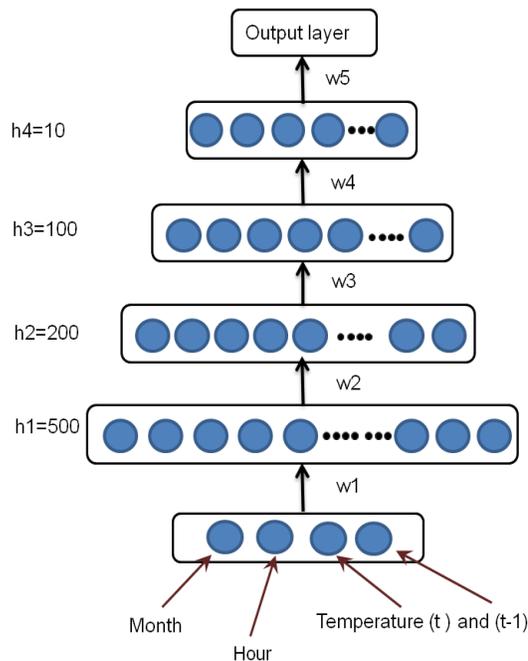


Fig. 6. DBN-RBM for temperature prediction.

Given two random variables x and y , their mutual information is defined in terms of their probabilistic density functions $p(x)$, $p(y)$ and $p(x,y)$:

$$I(x, y) = \iint p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dx dy \quad 1$$

The selected features x_i are required, individually, to have the largest mutual information, i.e., $I(x_i; c)$ with the target class

c, reflecting the largest dependency on the target class. In terms of sequential search, the m best individual features, i.e., the top m features in the descent ordering of $I(x_i;c)$, are often selected as the first m features [35].

Hence, features used as inputs for the training are corresponding temperature, previous pressure, previous humidity, corresponding Month and Hour. This feature set and labels were further normalized in the range of (0,1) prior to training. The humidity data was filtered with Butterworth filter corresponding same order of 2 and cutoff frequency at 0.11 mHz. The Actual and filtered humidity data is shown in Fig.7.

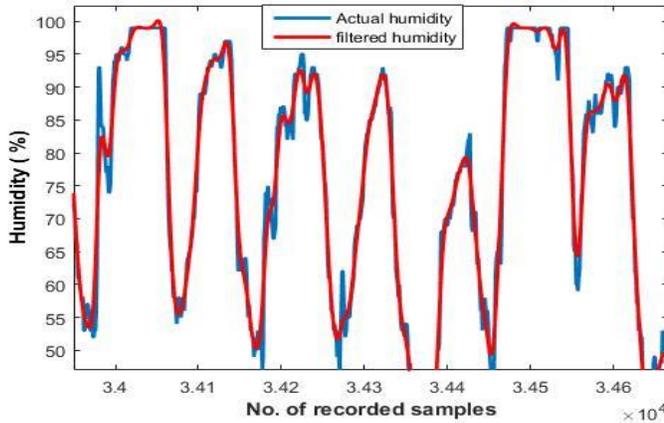


Fig. 7. Actual and filtered humidity series.

In order to construct and train a predictive model for humidity prediction, a DBN was developed with four hidden layers, one input layer consist of Six nodes and one output layer based on one output neuron. The hidden layers were RBM of size (300-200-100-10), which was trained one by one in a layer wise greedy way with contrastive divergence. The model is illustrated in Fig. 8. Initially, all weights and biases were assigned the value zero. The model was trained with total 120k samples and was further tested with rest of the data. The pretraining of RBM was performed using minibatches of size 10, with maximum one iteration for each layer pretraining. After training each layer separately the model was trained globally by adding an output layer with normalized humidity labels. However, fine tuning was performed with Maximum 800 iterations. The results are further discussed in Section IV.

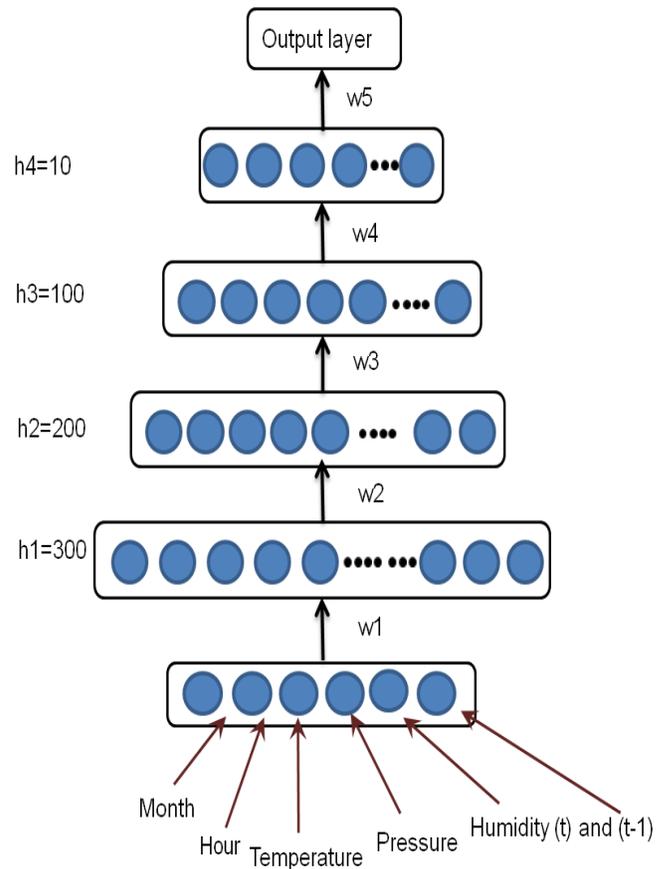


Fig. 8. DBN-RBM for humidity prediction.

E. Pressure Prediction

In general, pressure is a force applied perpendicular to the surface of an object per unit area over which that force is distributed. However, atmospheric pressure or air pressure, sometimes also called barometric pressure, is the pressure exerted by the weight of air in the atmosphere of Earth. The pressure data was smoothed with Butterworth low pass filter in same way as air temperature and relative humidity. However, a high pass filter was also applied on the data to detrend the linearly decreasing trend observed in recorded air pressure series. Fig. 9 presents the graph of actual and filtered pressure samples.

TABLE I. FEATURE SELECTION FOR HUMIDITY USING MUTUAL INFORMATION METHOD

Attributes	Time	Temperature	Pressure	Rain	Humidity	Wind direction	Wind velocity
Time	2.0748	9.06e-4	0.1261	0.0081	0.0246	0.0186	0.0281
Temperature	0.3748	1.933	0.0092	0.0189	0.0189	0.0421	0.0671
Pressure	0.1261	0.0920	1.9787	0.0223	0.0661	0.0227	0.0343
Rain	0.0081	0.0189	0.0223	0.308	0.05	0.01	0.005
Humidity	0.0246	0.0189	0.7256	0.0534	1.8134	0.0415	0.0920
Wind direction	0.0186	0.0421	0.0534	0.0154	0.0415	1.6601	0.0252
Wind velocity	0.0281	0.0671	0.0343	0.0058	0.0920	0.0252	1.7766

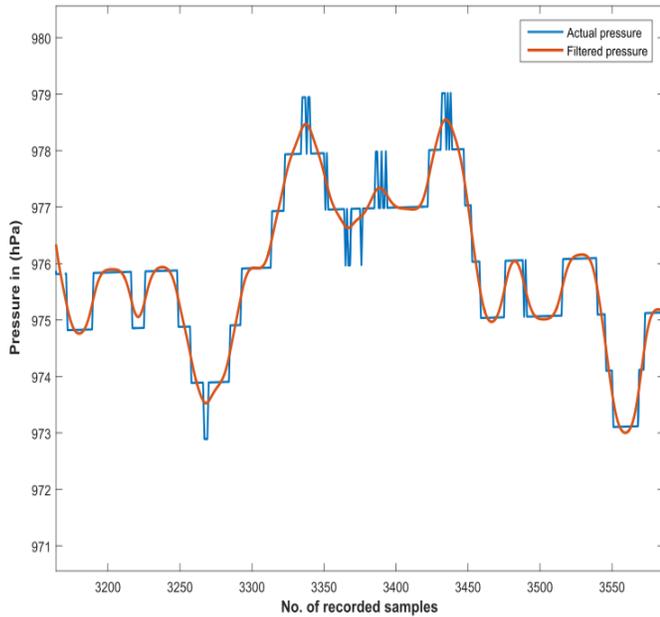


Fig. 9. Actual and filtered humidity series.

In order to extract valuable features for pressure prediction we explored some more aspects. The main factor that affects the air pressure at a given location is the altitude (or height above sea level) of that location. In order to select the meaningful features for air pressure prediction, we did little research. We came to know that the pressure depends on the density or mass of the air. Moreover, the density of air depends on its temperature and from our meteorological dataset the temperature depends on Season (categorized in months) and hour of the day. Thus we took the following parameters as input attributes for predicting the next pressure in series, the month, an hour, corresponding temperature, and pressure at (t) and (t-1).

In order to construct and train a predictive model for pressure prediction, a DBN was developed with three hidden layers, one input layer and one output layer. The hidden layers were RBM of size (300-200-5), which was trained one by one in a layer wise greedy way with contrastive divergence. The model is illustrated in Fig. 10.

The model was trained with total 120k samples and was further tested with the rest of the data. The pretraining of RBM was performed using minibatches of size 10, with maximum one iteration for each layer pretraining. After training each layer separately the model was trained globally by adding an output layer with normalized pressure labels. However, fine tuning was performed with maximum 800 iterations.

Initially, all weights and biases were assigned with the value zero in the training of each predictive model case. However, after pretraining the weights were found to be in form of normally distributed data. This identifies that weights are not randomly initialized in order to find the suitable solution. The weights were further transformed during the fine tuning phase.

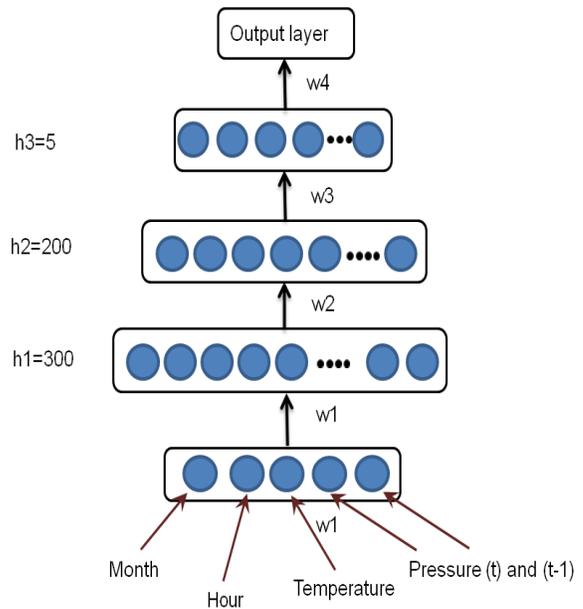


Fig. 10. DBN-RBM for humidity prediction.

Weights in the hidden layers of an each predictive model attempts to explore the nonlinear representations or features from the data. In this regard, the computed weights are also termed as feature detectors or receptive fields. These can be considered as a good way of visualizing which kind of features the hidden units have learned. There is a possibility that less meaningful or insignificant detectors may also be present. The results are further discussed in next section.

IV. RESULTS AND DISCUSSION

In this section, we describe and further discuss in detail the evaluation of trained DBN models for METEO nowcasting. According to our objective, in this work we attempted for the next sample prediction of non-stationary time series through DBN. As aforementioned, we were successful in deploying and training accurate models METEO nowcasting, weather parameter prediction. In our dataset, each weather parameter owns distinct trend and diverse behavior in its time series. For each parameter, we trained separate model and feature selection was performed accordingly. The performance for each predictive model is measured through three different performance metrics, i.e., Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Regression parameter R on Training and Test sets for prediction of Metrological Parameters. The statistics for measuring the performance of each predictive model is presented in the following Table 2.

TABLE II. MSE ON TRAINING AND TEST SETS FOR PREDICTION OF METROLOGICAL PARAMETERS

DBN Models	RMSE		MSE		R
	Training	Test	Training	Test	
Temperature prediction	8.57e-4	9.06e-4	7.35e-7	8.20e-7	0.99
Humidity prediction	1.3e-3	1.5e-3	1.58e-6	2.27e-6	0.99
Pressure prediction	9.07e-4	7.73e-4	8.24e-7	5.98e-7	0.99

The R parameter is linear regression, which relates targets to outputs estimated by network. If this number is equal to 1, then there is perfect correlation between targets and outputs. It is clearly obvious from the measures presented in Table 2 that the number for each model is very close to 1, which indicates a good fit. The MSE is a measure of the quality of predictive model. It is always non-negative, and values closer to zero are better. The MSE is computed as presented in (2), where \hat{y} is a vector of n predictions computed by model, and y is the vector of observed values. Taking the square root of MSE yields the RMSE.

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2 \quad 2$$

In Fig. 11, we present actual and predicted temperature samples taken from test set. It is visible from the figure that predicted samples are highly replicating the original temperature data.

It is clearly depicted from the Fig. 12 that, predicted humidity samples of test set are very close to the original recorded humidity. In the same way, strong correlations of predicted and recorded pressure can be seen from Fig. 13.

The results obtained from models are robust and shows considerably good predictions. Since, the models perform the forecasting task for only next one sample, however for meteo nowcasting our concerned objective is to predict for next three hours. This is considered as our next future target.

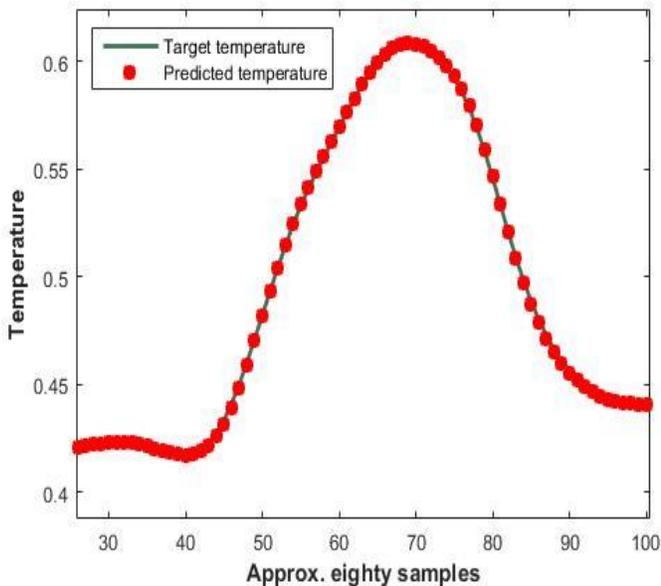


Fig. 11. Close view of target and predicted temperature with eighty samples from test dataset.

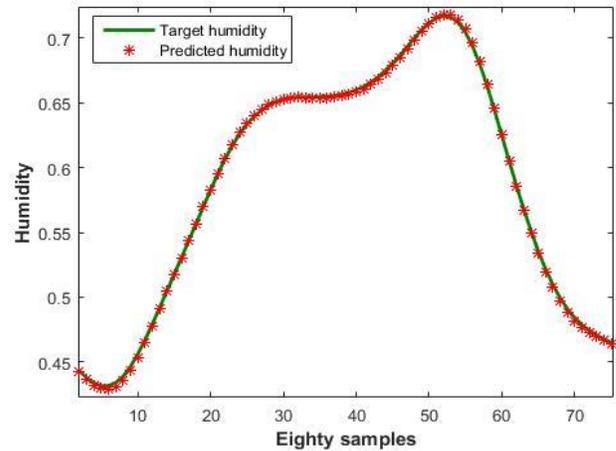


Fig. 12. Close view of target and predicted humidity with eighty samples from test dataset.

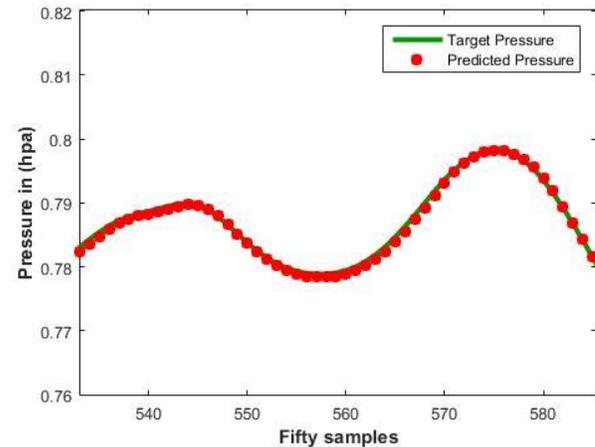


Fig. 13. Close view of target and predicted pressure with fifty samples from test dataset.

V. CONCLUSION AND FUTURE WORK

This research introduces a predictive ANN model based on deep learning hierarchical architecture, for the prediction of weather parameters such as temperature, pressure and humidity. The results shows outstanding performance of implemented DBN models while producing the accurate estimations.

The current research is limited to the implementation of DBN models, with the exception of providing any comparative evaluation with existing traditional neural network models. However, this can be taken as the direction for the future work.

REFERENCES

- [1] QT. Ain, M. Ali, A. Riaz, A. Noureen, M. Kamran, B. Hayat and A. Rehman, "Sentiment Analysis Using Deep Learning Techniques: A Review," *International Journal of Advanced Computer Science and Applications(IJACSA)*, vol. 8(6), pp. 424-433, 2017.
- [2] S. Narejo and E. Pasero, "Time Series Forecasting for Outdoor Temperature using Nonlinear Autoregressive Neural Network Models," *Journal of Theoretical and Applied Information Technology*, vol. 94(2) pp. 451-463, 2016.
- [3] M. Tektaş, "Weather forecasting using ANFIS and ARIMA models." *Environmental Research, Engineering and Management*. Vol. 51(1), pp. 5-10, 2010.
- [4] M. Khashei and M. Bijari, "An artificial neural network (p, d, q) model for timeseries forecasting," *Expert Systems with applications*, vol. 37(1), pp. 479-489, 2010.
- [5] R. E. Abdel-Aal, "Hourly temperature forecasting using abductive networks," *Engineering Applications of Artificial Intelligence*, vol. 17(5), pp. 543-556, 2004.
- [6] K. Abhishek, M. P. Singh, P., S. Ghosh and A. Anand, "Weather forecasting model using artificial neural network," *Procedia Technology*, vol. 4, pp. 311-318, 2012.
- [7] JJ Montaña Moreno, "Artificial neural networks applied to forecasting time series. *Psicothema*, vol. 23(2). 2011.
- [8] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, "Greedy layer-wise training of deep networks," *Advances in Neural Information Processing Systems*, vol. 19, pp. 153-160, 2007.
- [9] F. S. Albuquerque Filho, F. Madeiro, S.M. Fernandes, P.S. de Mattos Neto, T.A. Ferreira, "Time-series forecasting of pollutant concentration levels using particle swarm optimization and artificial neural networks," *Química Nova*, vol. 36(6), pp. 783-789, 2013.
- [10] S. Narejo and E. Pasero. "A Hybrid Approach for Time Series Forecasting Using Deep Learning and Nonlinear Autoregressive Neural Networks.," pp. 69-75, 2016
- [11] E. Busseti, I Osband, and S Wong, "Deep learning for time series modeling," *Technical report, Stanford University*, 2012.
- [12] J.N.K Liu, Y. Hu, J.J. You, and P.W. Chan, "Deep neural network based feature representation for weather forecasting." In *Proceedings on the International Conference on Artificial Intelligence (ICAI)*, p. 1. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2014.
- [13] M. Dalto, J. Matuško, and M. Vašak. "Deep neural networks for ultra-short-term wind forecasting." In *Industrial Technology (ICIT), 2015 IEEE International Conference on*, pp. 1657-1663. IEEE, 2015.
- [14] W. He, "Deep neural network based load forecast." *Comput. Model. New Technol*, vol. 18(3), pp. 258-262, 2014.
- [15] T. Kuremoto, S. Kimura, K. Kobayashi, and M. Obayashi. "Time series forecasting using a deep belief network with restricted Boltzmann machines." *Neurocomputing*, vol. 137, pp. 47-56, 2014.
- [16] Amaury Lendasse Erkki, Erkki Oja, and Olli Simula. *Time series prediction competition: the cats benchmark*. 2004.
- [17] Amaury Lendasse, Erkki Oja, Olli Simula, and Michel Verleysen. *Time series prediction competition: The cats benchmark*, 2007.
- [18] A. Grover, A. Kapoor, and E. Horvitz. "A deep hybrid model for weather forecasting." In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 379-386. ACM, 2015.
- [19] M. Längkvist, L. Karlsson, and A. Loutfi. "A review of unsupervised feature learning and deep learning for time-series modeling." *Pattern Recognition Letters* vol. 42, pp. 11-24, 2014.
- [20] G.W. Taylor, G.E. Hinton, and S.T. Roweis. "Modeling human motion using binary latent variables." In *Advances in neural information processing systems*, pp. 1345-1352, 2007.
- [21] I. Sutskever and G.E Hinton. "Learning multilevel distributed representations for high-dimensional sequences." In *Artificial Intelligence and Statistics*, pp. 548-555, 2007.
- [22] N. Garg, and J. Henderson. "Temporal restricted boltzmann machines for dependency parsing." In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers Vol. 2*, pp. 11-17. Association for Computational Linguistics, 2011.
- [23] I. Sorokin, "Classification Factored Gated Restricted Boltzmann Machine." In *AALTD@ PKDD/ECML*. 2015.
- [24] R. Memisevic, and G Hinton. "Unsupervised learning of image transformations." In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pp. 1-8. IEEE, 2007.
- [25] H. Lee, P. Pham, Y. Largman, A.Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," *Advances in neural information processing systems*, pp. 1096-1104, 2009.
- [26] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations." In *Proceedings of the 26th annual international conference on machine learning*, pp. 609-616. ACM, 2009.
- [27] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber. "Stacked convolutional auto-encoders for hierarchical feature extraction." *Artificial Neural Networks and Machine Learning-ICANN*, pp. 52-59, 2011.
- [28] K. Chen, M. Seuret, M. Liwicki, J. Hennebert, and R. Ingold. "Page segmentation of historical document images with convolutional autoencoders," In *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, pp. 1011-1015. IEEE, 2015.
- [29] A. Makhzani and B. Frey. "A winner-take-all method for training sparse convolutional autoencoders." In *NIPS Deep Learning Workshop*. 2014.
- [30] E. Pasero and W. Moniaci. "Artificial neural networks for meteorological nowcast," In *Computational Intelligence for Measurement Systems and Applications, CIMSA. 2004 IEEE International Conference*, pp. 36-39. IEEE, 2004.
- [31] M. Costa, W. Moniaci, and E. Pasero. "INFO: an artificial neural system to forecast ice formation on the road." In *Computational Intelligence for Measurement Systems and Applications, 2003. CIMSA'03. 2003 IEEE International Symposium*, pp. 216-221. IEEE, 2003.
- [32] E. Pasero, and W. Moniaci. "Learning and data driver methods for short term meteo forecast." *Lecture Notes in Computer Science 3931*, pp 105, 2006.
- [33] S. Narejo and E. Pasero, "An Application of Internet Traffic Prediction with Deep Neural Network", *Multidisciplinary Approaches to Neural Computing*. Springer , Cham, vol. 69, pp. 139-149, 2018.
- [34] D. Erhan, Y. Bengio, A. Courville, P.A. Manzagol, P. Vincent, and S. Bengio. "Why does unsupervised pre-training help deep learning?." *Journal of Machine Learning Research*, vol. 11(Feb), pp. 625-660, 2010.
- [35] H. Peng, F. Long, and C. Ding. "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy." *IEEE Transactions on pattern analysis and machine intelligence*, vol. 27(8), pp. 1226-1238, 2005.

A Review of Towered Big-Data Service Model for Biomedical Text-Mining Databases

Alshreef Abed

Department of Computer Science
and Technology, Wuhan University
of Technology Wuhan, China

Jingling Yuan

Department of Computer Science
and Technology, Wuhan University
of Technology Wuhan, China

Lin Li

Department of Computer Science
and Technology, Wuhan University
of Technology Wuhan, China

Abstract—The rapid growth of biomedical informatics has drawn increasing popularity and attention. The reason behind this are the advances in genomic, new molecular, biomedical approaches and various applications like protein identification, patient medical records, genome sequencing, medical imaging and a huge set of biomedical research data are being generated day to day. The increase of biomedical data consists of both structured and unstructured data. Subsequently, in a traditional database system (structured data), managing and extracting useful information from unstructured-biomedical data is a tedious job. Hence, mechanisms, tools, processes, and methods are necessary to apply on unstructured biomedical data (text) to get the useful business data. The fast development of these accumulations makes it progressively troublesome for people to get to the required information in an advantageous and viable way. Text mining can help us mine information and knowledge from a mountain of text, and is now widely applied in biomedical research. Text mining is not a new technology, but it has recently received spotlight attention due to the emergence of Big Data. The applications of text mining are diverse and span to multiple disciplines, ranging from biomedicine to legal, business intelligence and security. In this survey paper, the researcher identifies and discusses biomedical data (text) mining issues, and recommends a possible technique to cope with possible future growth.

Keywords—Big data; biomedical data; text mining; information retrieval; feature extraction

I. INTRODUCTION

Currently, the field of biomedical research is booming, a lot of biomedical knowledge is in unstructured form in the form of text file, and now the field has witnessed exponential trend increase; there is a need to solve the contradictions between massive growth of information and knowledge of text slowly and in a credible manner to identify useful patterns in the text which is still a challenge. In recent years, biomedical text mining technology which is one branch of an efficient automatic access to new exploration-related knowledge has witnessed significant progress [1].

Biomedical information is increasing rapidly in size, and helpful outcomes come into sight daily in research publications. However, automatically taking out useful information from such a stupendous quantity of documents is a difficult task because these documents are unstructured and are revealed in natural language. To enable data mining and

knowledge discovery techniques, documents should be in the structured format [2].

The problem faced by the biological researchers is on how to effectively find out the useful and needed documents in such an information-overload environment. Traditional manual retrieval method is impractical. Furthermore, online biological information exists in a combination of different forms, including structured, semi-structured and unstructured forms [3]. It is impossible to keep abreast of all developments. Computational methodologies increasingly become important for research [4]. Text mining techniques which involve the process of information retrieval, information extraction and data mining provide a means of solving this by Ananiadou et al. [5].

The volume of published knowledge in the biomedical region is produced at an unprecedented pace. Biomedical researchers need to explore the big amount of scientific publications to examine findings related to certain biomedical entities such as proteins, diseases, etc. In the biomedical domain, simple keyword based matching may not be adequate because biomedical entities have synonyms and ambivalent names. Biomedical text mining relates to automatically identifying biomedical entities from a given text and to associate them to their correlating entries in knowledge bases. Biomedical text mining enables researchers to recognize useful information more efficiently. Two elementary functions of information extraction are Named entity recognition and Relation extraction. Named entity recognition deals with detecting the name of entities. Relation extraction refers to uncovering the semantic relations between entities [2].

The number of articles that are added to the literature databases is growing at a fast pace [6]. Retrieval of relevant information from literature databases and combining this information with experimental output is time-consuming and requires careful selection of keywords and drafting of queries. This is often a biased and time-consuming process, resulting in incomplete search results, preventing the realization of the full potential that these databases can offer [7]. Automated processing and analysis of text (referred to as Text Mining (TM)) can assist researchers in evaluating scientific literature. Nowadays, TM is applied to answer many different research questions, ranging from the discovery of drug targets and biomarkers from high-throughput experiments [8]–[13] to drug repositioning, the creation of a state-of-the-art overview of a certain disease or therapeutic area and for the creation of

domain-specific databases [14], [15]. Due to the heterogeneous nature of written resources, the automated extraction of relevant biological knowledge is not trivial. As a consequence, TM has evolved into a sophisticated and specialized field in the biomedical sciences where text processing and machine learning techniques are combined with mining of biological pathways and gene expression databases.

The rest of the paper is organized as follows: Section II has discussed the purpose and overview of text mining, the significance of biomedical text mining, task, models and methods used. It presents the definitions of the concepts explored in this study. Section III discussed the previous study which is related to text mining, biomedical text mining, biomedical data with feature extraction approach and biomedical data mapping technique. It critically evaluates methodologies that were available at the time of this research. Section IV discussed research methods used by the reviewed articles. Analysis and discussion are covered in Section V wherein the contextual settings of the reviewed articles are examined. The study findings, conclusion, and recommendation for further research are discussed in Section VI.

II. PURPOSE OF THE STUDY

Due to the rapid growth of data and text, information extraction is a difficult task, especially in biomedical databases [16]. Additionally, the diversity, complexity and volume of the information that need to be mined present challenges in the biomedical domain impacts the biomedical discovery process, stifling researchers working towards novel hypotheses to address critical questions [17]. Subsequently, such information extraction depends on the flexible formulation and common methods for heterogeneous data integration and indefinable discovery of knowledge sources that highly depend on a particular scientific question. It truly influences the effective techniques of storage, extraction and permitting sympathetic of the molecular substructures of biological processes. For this purpose, this paper briefly overviews the major challenges in these areas and discusses the recommendations and implications of this research.

A. Overview of Text Mining

Text mining or text analytics is an umbrella term describing a range of techniques that seek to extract useful information from document collections through the identification and exploration of interesting patterns in the unstructured textual data of various types of documents - such as books, web pages, emails, reports or product descriptions. A more formal definition restricts text mining to the creation of new, nonobvious information (such as patterns, trends or relationships) from a collection of textual documents. Typical text mining tasks include activities of search engines, such as assigning texts to one or more categories (text categorization), grouping similar texts together (text clustering), finding the subject of discussions (concept/entity extraction), finding the tone of a text (sentiment analysis), summarizing documents, and learning relations between entities described in a text (entity relation modeling) [18].

The utilization of the web has expanded the obtainability and access to publications that are the foremost in various cases on data over-burdening [19]. Specifically, biomedical data sets have increased rapidly in large computerized stores [20]. Therefore, searching and organizing these data is always considered as time-consuming and cost ineffective. For example, in the digital library, the MEDLINE is a fast-growing biomedical database, and the information within this data set is stored in text form. Recently, it has stored more than 18 million indexed articles. So the usability and obtainability of this data have become precarious to the researchers and students who are working in the biomedical area [21]. The quick advancement of these accumulations makes it progressively troublesome for analysts to get the required information in a helpful and proficient way.

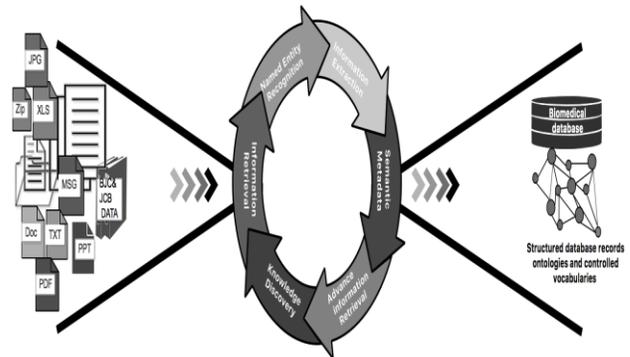


Fig. 1. Text mining eco-system for biomedical data.

Subsequently, the relationship amongst various medical conceptions from medical collected works is a foremost issue for many biological researchers. But, the data gathering level confines the incorporation into choice data frameworks for two reasons. Firstly, it requires more time from therapeutic specialists to create and maintain the learning base. Secondly, sharing and reusing the approved information base is troublesome due to the absence of clearness [22]. The goal is to obtain consistent data, and its extraction is one of the essential objectives of biomedical text mining groups [23]. The term text mining is used when exploring the objects stored in an unstructured data set and offers the capability towards managing and analyzing [24] the large sets of data in an effective manner [25]. While also realizing the significant relationships or correlations amongst variables in the huge dataset [26]. The smart data retrieval system is essential in operating non-standardized entries in order to access the data [27]. Subsequently, there is a robust need to create strategies for programmed extraction of pertinent data from the collected works, which is composed in natural language [28]. Therefore, in this study, the text mining method is towards discovering additional useful information in a more effective way. “Fig. 1” shows the overview of text mining process from the biomedical database.

B. Text Mining

Text mining refers to the automated extraction of knowledge and information from the text by revealing relationships and patterns that are present, but not obvious, in a document collection. Subsequently, it uses a wide range of

utilities including information extraction, text clustering, sentiment analysis, text categorization, document summarization, named entity recognition and question answering and the seven interdisciplinary fields based on computational linguistics: artificial intelligence, data mining, natural language processing and information retrieval [29].

The goal of text mining is to derive implicit knowledge that hides in unstructured text and present in an explicit form. This generally has four phases: information retrieval, information extraction, knowledge discovery, and hypothesis generation. Information retrieval systems aim to get desired text on a certain topic; information extraction systems are used to extract predefined types of information such as relation extraction; knowledge discovery systems help us to extract novel knowledge from the text; hypothesis generation systems infer unknown biomedical facts based on text, as shown in "Fig. 2". Thus, the general tasks of biomedical text mining include information retrieval, named entity recognition and relation extraction, knowledge discovery and hypothesis generation [30].

The text mining-associated text document and database models [31] are identified as:

- Information recovered from web archives with a population of data set patterns.
- Disclosure of data presented in the text as well as the capacity for XML or social groups.
- Incorporation and questioning of content information after it has been stored in databases.
- Deduplication of a data set through utilizing standard information mining strategies like clustering.

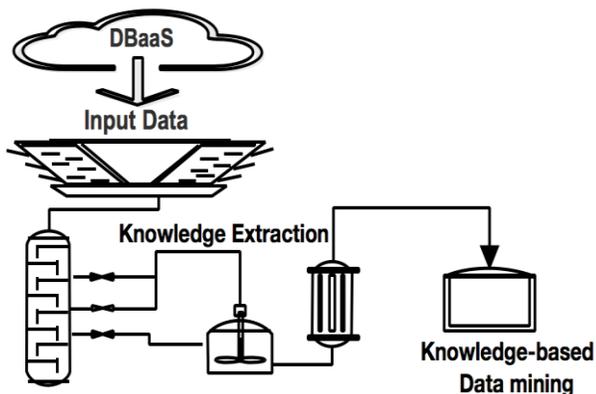


Fig. 2. BDaaS Utilization Model for knowledge extraction.

C. Models and Methods Used in Text Mining

To solve text mining issues, previously many researchers have suggested new methods for relevant information retrieval according to a user's requirement [32]. Based on the information retrieval process, there are four methods: term, phrase, pattern taxonomy and the concept-based method.

D. Biomedical Literature Mining

The era of applying text mining approaches to biology and biomedical fields came into existence in 1999. It was first

applied to the biomedical domain for gene expression profiling [33], as well as the extraction and visualization of protein-protein interaction [34]. It emerged as a hybrid discipline from the edges of three major fields, namely, bioinformatics, information science, and computational linguistics. Biomedical literature mining is concerned with the identification and extraction of biomedical concepts (e.g., genes, proteins, DNA/RNA, cells, and cell types) and their functional relationships [35]. The major tasks include 1) document retrieval and prioritization (gathering and prioritizing the relevant documents); 2) information extraction (extracting information of interest from the retrieved document); 3) knowledge discovery (discovering new biological event or relationship among the biomedical concepts); and 4) knowledge summarization (summarizing the knowledge available across the documents). A brief description of the biomedical literature mining tasks is listed as follows.

E. Biomedical Text Mining Tasks

Document Retrieval: The process of extracting relevant documents from a large collection is called document retrieval or information retrieval [36]. The two basic strategies applied are query-based and document-based retrieval. In query-based retrieval, documents matching with the user specified query are retrieved. In document-based retrieval, a ranked list of documents similar to a document of interest is retrieved.

Document Prioritization: The retrieved documents are usually prioritized to get the most relevant document. Many biomedical document retrieval systems achieve prioritization based on certain parameters including journal-related metrics (e.g., impact factor, citation count) [37] and MeSH index [38], [39] for biomedical articles. The similarity between the documents is estimated with various similarity measurements (e.g., Jaccard similarity, cosine similarity) [40].

Information Extraction: This task aims to extract and present the information in a structured format. Concept extraction and relation/event extraction are the two major components of information extraction [41], [42]. While concept extraction automatically identifies the biomedical concepts present in the articles, relation/event extraction is used to predict the relationship or biological event (e.g., phosphorylation) between the concepts [43], [44].

Knowledge Discovery: It is a nontrivial process to discover novel and potentially useful biological information from the structured text obtained from information extraction. Knowledge discovery uses techniques from a wide range of disciplines such as artificial intelligence, machine learning, pattern recognition, data mining, and statistics [45]. Both information extraction and knowledge discovery find their application in database curation [46], [47] and pathway construction [48], [49].

Knowledge Summarization: The purpose of knowledge summarization is to generate information for a given topic from one or multiple documents. The approach aims to reduce the source text to express the most important key points through content reduction selection and/or generalization [50]. Although knowledge summarization helps to manage the

information overload, state of the art is still open to research to develop more sophisticated approaches that increase the likelihood of identifying the information.

Hypothesis Generation: An important task of text mining is hypothesis generation to predict unknown biomedical facts from biomedical articles. These hypotheses are useful in designing experiments or explaining existing experimental results [51].

Text mining for biomedical literature often involves two major steps. a. First, it must identify biomedical entities and concepts of interests from free text using natural language processing techniques. Many text mining algorithms have been applied to this problem. For example, some morphological clues to recognize the heartache like obesity, blood pressure. b. And then, the converted information is extracted from the text or unstructured documents into the standardized data set, and data mining is applied to the data source. "Fig. 3", shows the typical text Mining Process.

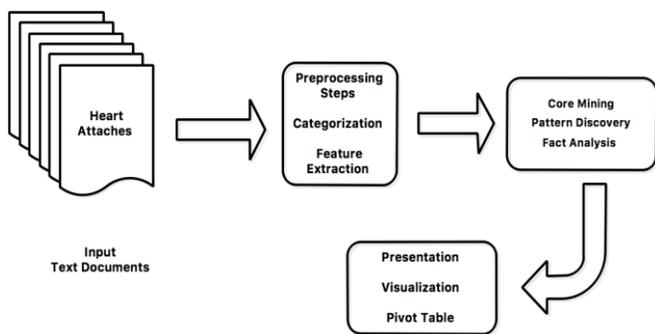


Fig. 3. Typical text mining process.

III. RELATED WORKS

This section provides a brief summary of text mining followed by most recent studies that have been conducted with regard to text mining in the field of biomedical research texts.

A. Biomedical Text Mining Review

Information extraction or (IE) covers the recognition of biomedical identities in biomedicine for extracting information pertaining to a disease, its treatment and its proteins and extracting the association (s) between these identities. The association between two different entities is extracted through different methods. Previous studies related to the extraction of useful information from databases are discussed as follows.

Tan and Lambri [52] suggested a framework for the purpose of selecting a suitable ontology for a specific application for biomedical text mining. Subsequently, an experiment was put forth for biomedical ontology in the context of a gene normalization system by utilizing the framework. Inside the references of the framework, the results of the assessment directed us to a comparatively firm option of ontology for our module. Furthermore, the researchers have planned to evaluate this framework with more applications and ontologies.

Qi et al. [53] conducted a survey about text mining in the realm of bioinformatics with a focus on the application of text

mining. During the course of this study, the primary research focus of text mining in bioinformatics was supported through exhaustive examples. This study, in particular, matched the requirement for a state-of-the-art area of text mining in bioinformatics, primarily due to the swift advancement in both the fields of text mining and bioinformatics. The full ability of this area has remained underutilized.

A framework of a probabilistic combination nature for the purpose of precisely linking citation information with the content-based information retrieval weighting model is suggested by Yin et al. [54]. Through a case study, they were able to observe the model of linking information that was available in the citation graph. Extensive parameter tuning can possibly be done away with through this framework. However, this basically tested the suggested combination framework in the context of a biomedical literature corpus; they researchers of the opinion that the basic premise of their paper could be absorbed for literature retrieval in other areas.

Also, Tari et al. [55] explained the Gene Properties Mining Portal as one that permits retrieving gene-centric data from literature through text mining. This portal acts as a node for scientists to discern vital relationships in an effective and efficient manner from literature. But, the precision of the relations that were extracted were influenced by many issues, for instance, by limiting the methods of extraction in addition to the quality of the sources.

Bchir and Ben Abdesslem Karaa [56] proposed a method for the purpose of extracting relations between disease and drug. To begin with, they deployed Natural Language Processing methods for preprocessing abstracts. Later, features were extracted in the form of a set of preprocessed abstracts. To conclude, a disease drug association was extracted through the utilization of a disease-drug Association through a machine learning classifier. But, they ended up extracting associations among drugs and diseases, with a need to additionally extract other relations among other concepts.

Mala and Lobiyal [57] relied on ontologies for extracting concepts and offered an algorithm to locate and identify concept-based clusters. They then went on to label semantic weightage for all terms for every document. They resorted to using a tagging mechanism commonly known as POS (Part of Speech Tagging) to locate nouns in addition to utilizing Rapid Miner for text mining method such as text processing. The use of medical ontologies can also enhance the outcome of this method.

Roth et al. [58] had an objective to extract from biomedical literature information that was supportive of Protein-Protein Interactions (PPIs) that were of a predictable nature. The demonstrated results of the relation extraction show that an f-score of 0.88 was witnessed on the HPRD50 corpus, and the similarities in semantics that were calculated with an angular distance were also proved to be statistically considerable.

Jimeno Yepes and Berlanga [59] suggested an innovative technique to create word-concept probabilities from knowledge bases (KBs), which could then act as a foundation for numerous text mining jobs. The findings indicated that this

technique secured enhanced accuracy when compared with other state-of-the-art methods, particularly in the context of the MSH WSD data set. However, the present refinement implementation does not attempt to recognize or locate new synonyms for prevailing concepts; rather, it only attempts to tag the data by quantifying the frequency of usage within a specific concept. It does not attempt to locate or unearth new concepts that are not found in the knowledge base. It would be worthwhile to evaluate information extraction methods in order to locate and recognize new synonyms [60] for concepts that are both prevailing and new.

Meaney et al. [61] debated the changes and patterns in the use of techniques in statistics and epidemiology found in medical literature from the last 20 years. Furthermore, the research proposed a method to improve the text-mining approach and incorporated advanced retrieval techniques to gauge the ratio of articles. This method referred to a specific technique that was statistical or epidemiological: this is where further study needs to be undertaken by the team [62], [63]. A statistical machine translation approach [64] and a Bayesian information extraction network for the Medline abstract [65] are used in the proposed text mining system to deal with this problem.

B. Text Mining Methods

Berardi et al. suggested a framework that assists biologists in automatically extracting information from machine-readable documents or texts. These extraction models were later used on unobserved texts in automatic mode. They reported an application that was a real world dataset compiled by publications, which were in turn chosen to aid biologists in annotating an HmtDB database.

An extension of the Okapi retrieval system that was effectual for mining biomedical text has been suggested by [66]. This led to two advantages in the system when compared with other models. First, this method is uncomplicated to implement and is not tagged to any domain. Secondly, it has proven its competence and effectiveness in TREC Genomics experiments. Despite the fact that the suggested extension is effective in discerning the subtle variations in the verbiage of a biological entity, it does not offer any comprehensive solution to encompass all variations in that lexicon. But, this algorithm cannot serve to be its identifying factor. Henceforth, such variations would be discerned through a query expansion algorithm.

A text summarization algorithm that used scientific literature in biomedicine which discerns the focal topic of biomarker cancer discoveries and all information in the literature that is deemed vital was suggested by Islam et al. [67]. The purpose of this study, however, needs to be directed towards extracting more specialized information on protein structure and image data mining. Also, the system needs to be optimized to handle large loads with quick response and must support multiple databases.

Liu et al. [68] introduced a study regarding names in the Bio Thesaurus, which was, in turn, collated from multiple databases present in a free-text by utilizing a data set that was automatically created from cross-referencing in the

UniProtKB. The findings proved that using different resources to put together synonyms for biological identities can result in optimized coverage for nomenclature present in the text while utilizing matching that is able to be adjusted. But, flexible matching creates more ambiguous situations for English words that are common. This results in the need to narrow down the confusion between common English vocabulary and biological identity nomenclature through corpus-based word sense disambiguation.

Leroy et al. [69] created text mining tools that indicate co-occurrence relations among concepts. Engaging subsets of relations are mined through statistical measures. In addition, the researchers proved the manner in which these relations were directed had an effect on the amount of interest. To summarise, the numerous relations and their assistants were quantified. The differences in direction had a remarkable effect on the number of relations, and it also included the firm support of different types of graphs. The consequences of directionality on bigger graphs were not considered, however.

Salahuddin and Rahman [70] attempted to analyze and collate biomedical data from hypertext documents by utilizing text mining methods with the assistance of biomedical ontology. The matching and layout of the biomedical entity from the Metathesaurus were performed through a query on a keyword. However, this study focused on data in documents alone. Documents contain both textual information and visual imagery, and hence, there is a need to take into consideration the relevance of images in medical documents and attempt to give ranking to the documents based on the combined textual and image content.

Ronquillo et al. [23] proposed a program for automatic categorization of biomedical text. The results achieved pertaining to performance and execution timing are more positive when compared to the results obtained earlier and used in Weka, and what is known as the baseline system. This system has certain limitations, however, especially when it needs to show the difference between texts regarding hearing loss classified as syndromic and nonsyndromic. For the purpose of improving categorization, this method will be used to locate and indicate symptoms and genes that are related to both types of hearing loss.

Hou et al. [71] proposed two options to help in directing the relation between genes and diseases (a) utilizing proximity relationship among genes and diseases, and (b) using GO terms that are prevalent among genes and diseases for the purpose of comparing similarity. Experiments demonstrate that relations using GO terms function better than utilizing word proximity. This proves that GO terms serve as a better option for good gene-disease association. But, this only concentrated on the aspect of the relationship. Additionally, there is a need to focus on applying prediction of gene-disease relationships apart from the OMIM database.

A text mining technique that extracts numerous entities from biomedical text had proposed by Javed and Afzal [72] where candidate terms are discerned through the application of an algorithm known as the C-Value. These candidate terms and prevalent terms used in Seed/Ontology are labeled in the corpus. By resorting to the assessment of profiles that were

lexical and contextual in the comparison between candidate terms and the prevailing Seed/Ontological Terms, it was possible for them to discover novel ideas and assess them. This study required an enhancement to the categorization of included measures that resembled each other, such as Word Net to discern the link between two terms.

The summary of text mining methods is presented in Table 1 where each method is briefly identified and then

analyzed. Other methods which include knowledge extraction and data mapping techniques have been classified in the next sections along with their summary in Table 2. The evaluation includes some major limitations in each method which need to be recovered for potential researches and experiments. The summary of previous studies related to biomedical data mapping techniques are discussed in Table 3. Finally, the recommendation and implication of this research are discussed in Table 4.

TABLE I. SUMMARY OF PREVIOUS STUDIES RELATED TO TEXT MINING APPROACH

Ref.	Method	Results	Advantage	Limitation
Berardi et al. [73]	Text extraction rule	Automatic information extraction	Fast and simple	Extract abbreviations and acronyms.
Ming Zhong and Xiangji [66]	Okapi retrieval approach	An effective TREC Genomics experiments	Simple implementation	Cannot serve to be identifying factors
Islam et al. [67]	text summarization algorithm	Discerns the focal topic of biomarker cancer discovery	Simple implementation	Does not support multiple and public databases
Liu et al. [68]	Text classification approach	Nomenclature text optimization	Flexible in text matching	High confusion between common English and biological vocabulary
Leroy et al. [69]	Text mining tool	Multiple text mapping	Secure and support different types of graphs	Does not support bigger graphs
Salahuddin and Rahman [70]	Ontology-based text mining	Documents identification	Effectiveness for fewer parameters.	document data only
Ronquillo et al. [23]	Text classification approach	Small data sets classification	High Performance and execution time	Does not identify some symptoms and genes
Hou et al. [71]	Text mining approach	Utilizing word proximity using GO terms function	Very secure	Does not support multiple and public databases
Javed and Afzal [72]	Text mining methodologies	Automatic biomedical text extraction	High efficiency.	Does not enhance similarity measures

C. Knowledge Extraction Methods

Jahiruddin et al. [74] introduced an innovative Biomedical Knowledge Extraction and Visualization framework (BioKEV) which is used to discern and isolate vital information components from biomedical text documents. The method of information extraction was based on NLP or Natural Language Processing methods and analysis that were also based on semantics. Additionally, it was suggested that a ranking system for documents needed to be in place to refer to retrieved documents in the same relevant order as queried by the user. Furthermore, they improved the format of the query processing module to render it compatible with a high degree of efficiency when searching biomedical queries of a complex nature.

Sharma et al. [75] concentrated on discovering the task and extracting relations that were witnessed between certain bioentities, like green tea and cancer of the breast. Additionally, a verb-centric algorithm was suggested to be put in place. This system locates and extracts the primary verb(s) observed in a sentence; hence, there is no requirement for a separate set of rules or patterns. The algorithm was assessed in numerous datasets and observed an average of F as 0.905, which is considerably more than what had been previously achieved.

However, a framework called Feature Coupling Generalization (FCG) for the purpose of developing novel features from untagged data has been suggested by Li et al. [76]. This framework chooses Example-Distinguishing Features (EDFs) and Class-Distinguishing Features (CDFs) to recognize the gene entity name (NER), extract the protein-protein interaction (PPIE) and classify the gene ontology (GO). Additionally, the performance of baselines that are under supervision was improved by 7.8 %, 5.0 %, and 5.8 %, respectively, in all three tasks. But this study does not justify the reason for the workings of FCG and the reasons that determine EDFs' and CDFs' qualities.

Holzinger et al. [77] proposed a Sequence Memorizer Based Model (SMBM) that had its roots in what was known as the generative model to oversee its functioning. This method resorted to the utilization of the generative strategy in order to avoid the option of selecting work that was time-consuming. While ensuring the advantages of models that were generative in nature, the functionality of this technique can be compared to that of the Maxent model.

Holzinger et al. [77] offered a way to assess knowledge discernment of disease-disease relationships for rheumatic diseases. Also, they resorted to utilizing a Point wise Mutual Information (PMI) calculation to identify a relationship's strength. The output indicates concealed knowledge in articles

pertaining to rheumatic diseases that were indexed by MEDLINE, and which could be used by medical experts and researchers for the purpose of making medical decisions. This study also needs to concentrate on collecting the names of diseases, nomenclature/codes of diagnosis and treatments to observe the extent to which identification of diseases in the searched content can be improved through screening for diagnosis and treatment of such diseases.

Pereira et al. [78] developed an integrated approach for the reconstruction of Transcriptional Regulatory Networks (TRNs), which retrieve the relevant data from important biological databases and insert the result into a unique repository named KREN. Further, they integrated this into the Note software system, which allows some methods from the

Biomedical Text Mining field, including algorithms for Named Entity Recognition (NER), extraction relationships between biological entities and extraction of all relevant terms from publication abstracts. Finally, this tool was extended to allow the reconstruction of TRN using scientific literature.

Landge and Rajeswari [79] conducted an overview of the comparative analysis of numerous techniques employed in determining the relation between chemical entities, and also reviewed the comparative analysis of numerous text mining methods. Further, they suggested to using a parallel approach to text mining towards minimizing the time needed by their method. Conventional algorithms can be parallelized and applied to mine and extract information and knowledge from a large data set.

TABLE II. SUMMARY OF PREVIOUS STUDIES RELATED TO KNOWLEDGE EXTRACTION METHOD

Ref.	Method	Results	Advantage	Limitation
Tangtulyangkul et al.[80]	Keyword mining scheme	External-source based knowledge accumulator	Reduce Information overload	clinical records only
Sharma et al. [75]	Verb-centric algorithm	Biomedical entities identification	Handled complex sentence	private data sets only
Liu et al. [68]	FCG framework	Supervised data learning utility	High supervised baselines performance	Does not confirm FCG, EDFs and CDFs qualities
Holzinger et al. [77]	Sequence memorizer Based Model	Natural language recognition	High performance.	Does not integrate sequence memorized into machine learning model.
Holzinger et al. [77]	Decision-making approach	Medical decision-making processes	High efficiency	Disease names only
Pereira et al. [78]	Transcriptional Regulatory Networks	Gene scientific corpora	Robust extraction	Does not validate the regulatory model.
Landge and Rajeswari [79]	Reviewed text mining approach towards chemical entities	Chemical data machine learning algorithms	High efficient for small samples	Does not use parallel approach

D. Biomedical's Data Mapping Techniques

Cano et al. [81] suggested an approach that was hybrid in nature for the purpose of mining or unearthing the vast knowledge that was accumulated in the scientific literature. This method has its foundations on the utilization of effectively mining text through tools that work in tandem with precise and collaborative human duration. To demonstrate the effectiveness of this method, this study requires quantification of the time that is reduced in performing tasks. This leads to an observable upgrade in the state of information regarding the remaining portion of the knowledge content and ensures that active learning techniques are put into use for assigning priority to the annotation process.

Yang and Dong [82] proposed a mapping-based approach by first mapping bio-entities to terms in an established ontology Medical Subject Headings (MeSH). Specifically, they present two approaches to mapping biomedical entities identified using the Unified Medical Language System Met

thesaurus to MeSH terms. The first approach utilizes a special feature of the MetaMap algorithm, and the second employs an approximate phrase-based match to map entities directly to MeSH terms. These two approaches deliver comparable results with an accuracy of 72% and 75%, respectively, based on two evaluation datasets.

Mohammed and Nazeer [83] suggested an enhanced system of text mining that was focused on the method of matching patterns and heuristics that reduced space and increased the recall and accuracy. The system recalls, f-factor and precision were assessed through three metrics. The output of the experiments resulted in a recall of 98.68% and precision of 98.68%.The system has a drawback, though, in that it placed restrictions on the format of candidate acronym-definition pairs, which means that they needed to appear as either an acronym.

Ji et al. [84] created a Map Reduce algorithm to calculate the strength of association among two biomedical terms

witnessed in biomedical documents. Additionally, they evaluated if the algorithm was scalable by utilizing 3,610 documents retrieved from biomedical journals. Further, they demonstrated that this algorithm was linearly scalable when measured in the context of the number of nodes in a cluster. This method was only tested on a limited number of clusters with a reduced dataset, therefore leaving an additional need to assess the scalability of the algorithm in the context of the dataset size. Moreover, the algorithm needed to be enhanced in efficiency and accuracy.

TABLE III. SUMMARY OF PREVIOUS STUDIES RELATED TO BIOMEDICAL DATA MAPPING TECHNIQUES

Ref.	Method	Results	Advantage	Limitation
Cano et al. [81]	Hybrid mining approach	Scientific literature knowledge mining	High efficient.	Does not quantify the time reduction
Yang and Dong [82]	Mapping-based approach	Biomedical entities mapping and identification	High efficient for dimensional data	Low accuracy.
Mohammed and Nazeer [83]	Pattern matching method	Space reduction heuristics	High accuracy.	Low acronym definition
Ji et al. [84]	Map Reduce	Interestingness calculation	Linear scalability.	Tested on a small number clusters with low- scale datasets.

IV. MATERIALS AND METHODS

A comprehensive literature search of mining for text information in a medical database was conducted using a database such as Google search, Elsevier, IEEE, and Springer digital library and other literature sources. The searches were restricted to the years 2005 to 2016. The retrieved articles had text summarizations, like clinical, biomedical and medical summarization. This kind of search approach was applied to the web supplement. Additionally, searching through the collected database investigated the references of the included articles with an uncommon spotlight on past pertinent surveys. As seen in this review, the search retrieved a total of 112 potentially suitable articles to fulfill the inclusion criteria required for this review. Here, this study included unique examinations concentrated on the created and assessed text summarization techniques in the therapeutic areas, together with a summarization of electronic health records and biomedical collected works. “Fig. 4” shows the study flow diagram based on the PRISMA guidelines for reporting systematic reviews. However, the studies that met any of the following norms were excluded: images and multimedia summarization without a text summarization component, summarization of substance outside the biomedical area and non-English may have missed frameworks that compress content in different languages.

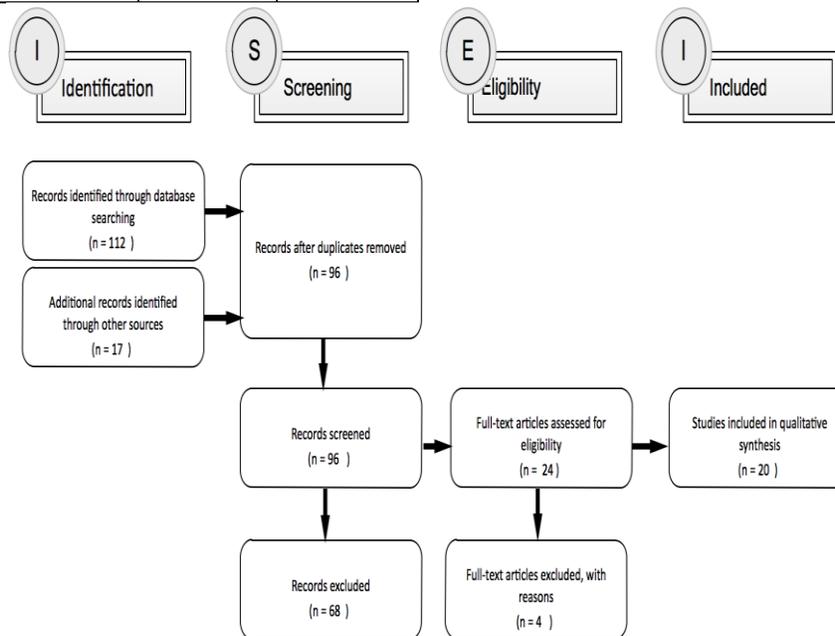


Fig. 4. Analysis of referred article.

V. RESULTS AND DISCUSSION

From the above review, the studies of Tan and Lambri, Salahuddin & Rahman, Yang & Dong, [52], [70], [82] have suggested a structure for choosing a suitable ontology for a specific biomedical text mining application. However, this study needs to focus more on handling the complex biomedical words. Additionally, this study only concentrated

on document data. The archive is improved with both printed data and pictures. Therefore, this study needs to consider the significance of pictures in medicinal reports and attempt to rank archives both on the premise of printed data and picture data [67], [70]. Also, a focus on gaining higher mapping accuracy should be included.

Some studies focused on event extraction applications in biomedical text, such as [85], [86]. On the other hand, some concentration was held on security-based event extraction applications, such as [87]. Hogenboom et al. [88] reviewed event extraction methods from the text for decision support systems. They extracted the biologist's data automatically from the text, though some researchers had proposed [68], [72], [73] a mining based framework. However, this study needs to focus on reducing redundancy in data as well as improving the classification by adding similarity measures in order to extract the biomedical term proposed [71], [84] association rule mining approach. However, this method was tested on a fewer number of the clusters with low-scale datasets. also, there is a need for further refinement of this algorithm to improve the overall efficiency and accuracy. Landge and Rajeswari [79] reviewed the comparative analysis of various text mining methods to find an association amongst various chemical entities. They also discussed that text mining algorithms take a large amount of time [81] for the huge data sets. For this purpose, they suggested using the parallel approach of text mining towards minimizing the time over huge datasets.

Few of the previous studies have proposed a framework for identifying key information components from biomedical text documents, such as [55], [82] and [74], [84]. But, the precision of the extricated relations was influenced by various issues, for example, the impediment of the extraction designs and the nature of the sources. Liu et al. (2007) Aimed to study the Bio Thesaurus. Nonetheless, the examination of sets with names was neglected, which demonstrates that there are a few equivalent words in the content that were neglected to be caught in the Bio Thesaurus [59], [60]. [54], [56], [77], [80] all extracted keywords from biomedical records. However, this study only focused on the biomedical literature corpus and could be adapted to literature retrieval in other domains [89], [90]. Hou et al. Sharma et al. [71], [75] focused on mining associations amongst bioentities, like breast cancer and green tea. However, this study only concentrated on a specific data set. Furthermore, this would take a shot at the undertakings of categorization and relationship integration [23], [72], [76] which proposed an algorithm for categorizing biomedical text in an automatic manner. However, this system needs to improve the classification to achieve a higher performance [58]. A deep validation process in order to compare this method with the existing regulatory model is still necessary. Meaney et al. [61] recommended, enhancing the text mining technique towards a retrieval approach or highly sophisticated preprocessing [35], which could be utilized to evaluate the extent of articles referring to a given epidemiological or statistical technique [62], [63].

It is hence clear that biomedical text mining has great potential. However, that potential is yet unrealized. In the following years, text mining should be able to evaluation validate the results of analytical expression methods in identifying significant groupings of data [91]. Text mining

researcher should co-operate with biology researchers in this interdisciplinary area. The following are some of the potential "New Frontiers" in biomedical text mining: Question-answering, Summarization, Mining data from full text (including figures and tables), User-driven systems, Evaluation [92] Now, this is an exciting time in biomedical text mining, full of promise.

VI. CONCLUSION

In this research, the researcher discussed and analyzed text mining techniques for biomedical data retrieving from the pool of documents on the web. From the literature, the biomedical record recovery strategy demonstrates about ideal results. In any case, the significance of a web report significantly relies on upon client's need that implies how much applicable the web record is as indicated by the client question. More effective text processing approach will provide an ideal result for the retrieval of the document from the web. Proficiency in processing mainly depends on time, but the calculation of time for ranking is a critical issue in implementation. As the web contains a large number of reports, offline estimation approach is not estimated effectively by any of existing approaches. Due to the complexity of Natural language processing, there is a broad examination in this field. So in future, it is necessary to concentrate more towards an effective method for capturing the meaning as well as relationships of words present in the document.

Based on the above review, future studies need to focus on:

- Cognitive aspects of text summarization which include visualization techniques, and evaluations of the impact of text summarization systems in work settings.
- Need to enable summarization corpora and reference standards to support the development of summarization tools in various applications.
- The increasing interest of users in efficiently retrieving and extracting relevant information, the need to keep up with new discoveries described in the literature or in biological databases, and the demands posed by the analysis of high-throughput experiments, are the underlying forces motivating the development of text-mining applications in molecular biology. Those technologies should provide the foundation for future knowledge-discovery tools able to identify previously undiscovered associations, something that will assist in the formulation of models of biological systems.
- Need to enable publicly available summarization corpora and reference standards to support the development of summarization tools.
- Need to improve the classification and mine the data towards getting higher performance

TABLE IV. RECOMMENDATIONS AND IMPLICATIONS OF THIS RESEARCH

Recommendation	Definition
Text Summarization	Further research is required in the subjective parts of text summarization, together with visualization method and the assessments towards the effect of text summarization systems in work settings.
Summarization Tool	Need to permit the reference standards and summarization corpora towards supporting the advancement of summarization tools in different applications.
Databases	The expanding enthusiasm of clients in productively recovering and separating important data, the need to stay aware of new disclosures depicted in the collected work or inorganic databases. Also, the requests postured by the investigation of high-throughput investigates, investigation are the basic powers spurring the improvement of text mining applications in sub-atomic science. Those innovations ought to provide an establishment of future information disclosure devices ready to distinguish already unfamiliar affiliations that will help with planning models of organic frameworks.
Higher Performance	Need to concentrate towards increasing the classification and mining the data for the attainment of higher performance.

ACKNOWLEDGMENT

This research project is supported by the ministry of higher education in Saudi Arabia and the National Natural Science Foundation of China (Grant No: 61303029,61602353), National Social Science Foundation of China (Grant No: 15BGL048), 863 Program (2015AA015403), Hubei Province Science and Technology Support Project(2015BAA072).

REFERENCE

[1] K. Meijing, Z. Le, and W. Jun, "Review of Research on Biological literature text mining," worldcomp, 2017. [Online]. Available: <http://worldcomp-proceedings.com/proc/p2014/BIC3034.pdf>. [Accessed: 21-Jun-2017].

[2] R. P. Saste and S. S. Patil, "Extraction of incremental information using query evaluator," in 2014 First International Conference on Networks & Soft Computing (ICNSC2014), 2014, pp. 324–328.

[3] M. Ghanem, A. Chortaras, Yike Guo, A. Rowe, and J. Ratcliffe, "A grid infrastructure for mixed bioinformatics data and text mining," in The 3rd ACS/IEEE International Conference on Computer Systems and Applications, 2005., 2005, pp. 185–188.

[4] G. C. Black and P. E. Stephan, "Bioinformatics: Recent Trends in Programs, Placements and Job Opportunities," Biotech, 2004. [Online]. Available: http://biotech35.tripod.com/private/ReportBioinfSloan_June04.pdf. [Accessed: 21-Jun-2017].

[5] S. Ananiadou, D. B. Kell, and J. Tsujii, "Text mining and its potential applications in systems biology.," Trends Biotechnol., vol. 24, no. 12, pp. 571–9, Dec. 2006.

[6] W. W. M. Fleuren and W. Alkema, "Application of text mining in the biomedical domain," Methods, vol. 74, pp. 97–106, Mar. 2015.

[7] L. J. Jensen, J. Saric, and P. Bork, "Literature mining for the biologist: from information retrieval to biological discovery," Nat. Rev. Genet., vol. 7, no. 2, pp. 119–129, Feb. 2006.

[8] C. Plake, L. Royer, R. Winnenburger, J. Hakenberg, and M. Schroeder, "GoGene: gene annotation in the fast lane," Nucleic Acids Res., vol. 37, no. Web Server, pp. W300–W304, Jul. 2009.

[9] Z.-X. Huang, H.-Y. Tian, Z.-F. Hu, Y.-B. Zhou, J. Zhao, and K.-T. Yao, "GenCLIP: a software program for clustering gene lists by literature profiling and constructing gene co-occurrence networks related to custom keywords," BMC Bioinformatics, vol. 9, no. 1, p. 308, 2008.

[10] A. Kentsis, F. Monigatti, K. Dorff, F. Campagne, R. Bachur, and H. Steen, "Urine proteomics for profiling of human disease using high

accuracy mass spectrometry," PROTEOMICS - Clin. Appl., vol. 3, no. 9, pp. 1052–1061, Sep. 2009.

[11] F. Al-Shahrour et al., "FatiGO +: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments," Nucleic Acids Res., vol. 35, no. suppl_2, pp. W91–W96, Jul. 2007.

[12] A. S. Haqqani, J. Kelly, E. Baumann, R. F. Haseloff, I. E. Blasig, and D. B. Stanimirovic, "Protein Markers of Ischemic Insult in Brain Endothelial Cells Identified Using 2D Gel Electrophoresis and ICAT-Based Quantitative Proteomics," J. Proteome Res., vol. 6, no. 1, pp. 226–239, Jan. 2007.

[13] W. W. M. Fleuren et al., "CoPub update: CoPub 5.0 a text mining system to answer biological questions," Nucleic Acids Res., vol. 39, no. suppl, pp. W450–W454, Jul. 2011.

[14] K. Jensen, G. Panagiotou, and I. Kouskoumvekaki, "Integrated Text Mining and Chemoinformatics Analysis Associates Diet to Health Benefit at Molecular Level," PLoS Comput. Biol., vol. 10, no. 1, p. e1003432, Jan. 2014.

[15] D. G. Jamieson, M. Gerner, F. Sarafraz, G. Nenadic, and D. L. Robertson, "Towards semi-automated curation: using text mining to recreate the HIV-1, human protein interaction database," Database, vol. 2012, p. bas023-bas023, Apr. 2012.

[16] W. Hersh, "Evaluation of biomedical text-mining systems: Lessons learned from information retrieval," Brief. Bioinform., vol. 6, no. 4, pp. 344–356, 2005.

[17] G. Gonzalez et al., "Text and Data Mining For Biomedical Discovery," 2014.

[18] M. Truyens and P. Van Eecke, "Legal aspects of text mining," Comput. Law Secur. Rev., vol. 30, no. 2, pp. 153–170, Apr. 2014.

[19] G. Leroy et al., "Genescene: Biomedical Text And Data Mining," in Proceedings of the 2003 Joint Conference on Digital Libraries (JCDL'03), 2003.

[20] S. Bleik, M. Song, A. Smalter, J. Huan, and G. Lushington, "CGM: A biomedical text categorization approach using concept graph mining," in 2009 IEEE International Conference on Bioinformatics and Biomedicine Workshop, 2009, pp. 38–43.

[21] PubMed, "<http://www.ncbi.nlm.nih.gov/pubmed>," 2016. .

[22] S. L. Achour, M. Dojat, C. Rieux, P. Bierling, and E. Lepage, "A UMLS-based Knowledge Acquisition Tool for Rule-based Clinical Decision Support System Development," J. Am. Med. Assoc., vol. 8, no. 4, pp. 351–360, 2001.

[23] F.-I. Ronquillo, C. P. de Celis, G. Sierra, I. da Cunha, and J.-M. Torres-Moreno, "Automatic classification of biomedical texts: experiments with a hearing loss corpus," in 4th International Conference on Biomedical Engineering and Informatics (BMEI), 2011, pp. 1674–1679.

[24] A. Browne, B. D. Hudson, D. C. Whitley, M. G. Ford, and P. Picton, "Biological data mining with neural networks: implementation and application of a flexible decision tree extraction algorithm to genomic problem domains," Neurocomputing, vol. 57, pp. 275–293, Mar. 2004.

[25] D. Luo et al., "Searching association rules of traditional Chinese medicine on Ligusticum wallichii by text mining," in 2013 IEEE International Conference on Bioinformatics and Biomedicine, 2013, pp. 162–167.

[26] Y. He et al., "Using association rules mining to explore pattern of Chinese medicinal formulae (prescription) in treating and preventing breast cancer recurrence and metastasis," J. Transl. Med., vol. 10, no. Suppl 1, p. S12, 2012.

[27] A. Li, Q. Zang, D. Sun, and M. Wang, "A text feature-based approach for literature mining of lncRNA?protein interactions," Neurocomputing, vol. 206, pp. 73–80, Sep. 2016.

[28] Y. Liu, M. Brandon, S. Navathe, R. Dingledine, and B. J. Ciliax, "Text mining functional keywords associated with genes," Stud. Health Technol. Inform., vol. 107, pp. 292–296, 2004.

[29] T. Magerman, B. Van Looy, B. Baesens, and K. Debackere, "Assessment of Latent Semantic Analysis (LSA) text mining algorithms for large scale mapping of patent and scientific publication documents," katholieke Universiteit Leuven, 2011.

- [30] F. Zhu et al., "Biomedical text mining and its applications in cancer research," *J. Biomed. Inform.*, vol. 46, no. 2, pp. 200–211, Apr. 2013.
- [31] A. Stavrianou, P. Andritsos, and N. Nicoloyannis, "Overview and Semantic Issues of Text Mining," *SIGMOD Rec.*, vol. 36, no. 3, pp. 23–34, 2007.
- [32] S. V. Gaikwad, A. Chaugule, and P. Patil, "Text Mining Methods and Techniques," *Int. J. Comput. Appl.*, vol. 85, no. 17, pp. 42–45, 2014.
- [33] L. Tanabe, U. Scherf, L. H. Smith, J. K. Lee, L. Hunter, and J. N. Weinstein, "MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling," *Biotechniques*, vol. 27, no. 6, pp. 1210–4, 1216–7, Dec. 1999.
- [34] C. Blaschke, M. A. Andrade, C. Ouzounis, and A. Valencia, "Automatic extraction of biological information from scientific text: protein-protein interactions," *Proceedings. Int. Conf. Intell. Syst. Mol. Biol.*, pp. 60–7, 1999.
- [35] M. Krallinger and A. Valencia, "Text-mining and information-retrieval methods for molecular biology," *Genome Biol.*, vol. 6, no. 7, p. 224, 2005.
- [36] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval: The Concepts and Technology Behind Search*, 2nd ed. Boston: Addison Wesley, 2011.
- [37] Y. Lin, W. Li, K. Chen, and Y. Liu, "A Document Clustering and Ranking System for Exploring MEDLINE Citations," *J. Am. Med. Informatics Assoc.*, vol. 14, no. 5, pp. 651–661, Sep. 2007.
- [38] S. J. Darmoni et al., "Improving information retrieval using Medical Subject Headings Concepts: a test case on rare and chronic diseases," *J. Med. Libr. Assoc.*, vol. 100, no. 3, pp. 176–183, Jul. 2012.
- [39] M. Petrova, P. Sutcliffe, K. W. M. B. Fulford, and J. Dale, "Search terms and a validated brief search filter to retrieve publications on health-related values in Medline: a word frequency analysis study," *J. Am. Med. Inform. Assoc.*, vol. 19, no. 3, pp. 479–88, 2012.
- [40] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval International Student Edition*. Chennai: Cambridge University Press, 2008.
- [41] R. Leaman and G. Gonzalez, "BANNER: an executable survey of advances in biomedical named entity recognition," *Pac. Symp. Biocomput.*, pp. 652–63, 2008.
- [42] K. Raja, S. Subramani, and J. Natarajan, "A hybrid named entity tagger for tagging human proteins/genes," *Int. J. Data Min. Bioinform.*, vol. 10, no. 3, pp. 315–28, 2014.
- [43] M. Torii, C. N. Arighi, G. Li, Q. Wang, C. H. Wu, and K. Vijay-Shanker, "RLIMS-P 2.0: A Generalizable Rule-Based Information Extraction System for Literature Mining of Protein Phosphorylation Information," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 12, no. 1, pp. 17–29.
- [44] K. Raja, S. Subramani, and J. Natarajan, "PPInterFinder—a mining tool for extracting causal relations on human proteins from literature," *Database*, vol. 2013, p. bas052-bas052, Jan. 2013.
- [45] J. Natarajan, D. Berrar, C. J. Hack, and W. Dubitzky, "Knowledge discovery in biology and biotechnology texts: a review of techniques, evaluation strategies, and applications," *Crit. Rev. Biotechnol.*, vol. 25, no. 1–2, pp. 31–52.
- [46] K. E. Ravikumar, K. B. Waghlikar, D. Li, J.-P. Kocher, and H. Liu, "Text mining facilitates database curation - extraction of mutation-disease associations from Bio-medical literature," *BMC Bioinformatics*, vol. 16, p. 185, Jun. 2015.
- [47] S. Matos et al., "Mining clinical attributes of genomic variants through assisted literature curation in Egas," *Database*, vol. 2016, p. baw096, Jun. 2016.
- [48] S. Subramani, R. Kalpana, P. M. Monickaraj, and J. Natarajan, "HPMiner: A text mining system for building and visualizing human protein interaction networks and pathways," *J. Biomed. Inform.*, vol. 54, pp. 121–31, Apr. 2015.
- [49] J. Czarniecki, I. Nobeli, A. M. Smith, and A. J. Shepherd, "A text-mining system for extracting metabolic reactions from full-text articles," *BMC Bioinformatics*, vol. 13, p. 172, Jul. 2012.
- [50] R. Mishra et al., "Text summarization in the biomedical domain: A systematic review of recent research," *J. Biomed. Inform.*, vol. 52, pp. 457–467, Dec. 2014.
- [51] F. Zhu et al., "Biomedical text mining and its applications in cancer research," *J. Biomed. Inform.*, vol. 46, no. 2, pp. 200–11, Apr. 2013.
- [52] H. Tan and P. Lambri, "Selecting an ontology for biomedical text mining," in *Proceeding BioNLP '09 Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, 2009, pp. 55–62.
- [53] Y. Qi, Y. Zhang, and Min Song, "Text Mining for Bioinformatics: State of the Art Review," in *2009 2nd IEEE International Conference on Computer Science and Information Technology*, 2009, pp. 398–401.
- [54] X. Yin, J. X. Huang, and Z. Li, "Mining and modeling linkage information from citation context for improving biomedical literature retrieval," *Inf. Process. Manag.*, vol. 47, no. 1, pp. 53–67, Jan. 2011.
- [55] L. Tari et al., "Mining Gene-centric Relationships from Literature to Support Drug Discovery," in *2011 IEEE International Conference on Bioinformatics and Biomedicine*, 2011, pp. 639–644.
- [56] A. Bchir and W. Ben Abdesslem Karaa, "Extraction of drug-disease relations from MEDLINE abstracts," in *2013 World Congress on Computer and Information Technology (WCCIT)*, 2013, pp. 1–3.
- [57] V. Mala and D. K. Lobiyal, "Concepts extraction for medical documents using ontology," in *2015 International Conference on Advances in Computer Engineering and Applications*, 2015, pp. 773–777.
- [58] A. Roth, S. Subramanian, and M. K. Ganapathiraju, "Towards extracting supporting information about predicted protein-protein interactions," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, pp. 1–1, Dec. 2015.
- [59] A. Jimeno Yepes and R. Berlanga, "Knowledge based word-concept model estimation and refinement for biomedical text mining," *J. Biomed. Inform.*, vol. 53, pp. 300–307, Feb. 2015.
- [60] D. R. Blair, K. Wang, S. Nestorov, J. A. Evans, and A. Rzhetsky, "Quantifying the Impact and Extent of Undocumented Biomedical Synonymy," *PLoS Comput. Biol.*, vol. 10, no. 9, p. e1003799, Sep. 2014.
- [61] C. Meaney, R. Moineddin, T. Voruganti, M. A. O'Brien, P. Krueger, and F. Sullivan, "Text mining describes the use of statistical and epidemiological methods in published medical research," *J. Clin. Epidemiol.*, vol. 74, pp. 124–132, Jun. 2016.
- [62] D. Jurafsky and J. Martin, *Natural Language Processing*. Englewood Cliffs, NJ: Pearson, 2008.
- [63] C. Manning, *Foundations of Statistical Natural Language Processing*. Cambridge, MA: The MIT Press, 1999.
- [64] A. Bodile and M. Kshirsagar, "Text mining in radiology reports by statistical machine translation approach," in *2015 Global Conference on Communication Technologies (GCCT)*, 2015, pp. 238–241.
- [65] M. Mannai and W. Ben Abdesslem Karaa, "Bayesian information extraction network for Medline abstract," in *2013 World Congress on Computer and Information Technology (WCCIT)*, 2013, pp. 1–3.
- [66] Ming Zhong and Xiangji Huang, "An effective extension to okapi for biomedical text mining," in *2006 IEEE International Conference on Granular Computing*, 2006, pp. 615–618.
- [67] M. T. Islam, D. Bollina, A. Nayak, and S. Ranganathan, "Intelligent Agent System for Bio-medical Literature Mining," in *2007 International Conference on Information and Communication Technology*, 2007, pp. 57–63.
- [68] H. Liu, M. Torii, Z. Hu, and C. Wu, "Mapping Gene/Protein Names in Free Text to Biomedical Databases," in *Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007)*, 2007, pp. 101–106.
- [69] G. Leroy, M. Fiszman, and T. C. Rindfleisch, "The Impact of Directionality in Predications on Text Mining," in *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008)*, 2008, pp. 228–228.
- [70] S. Salahuddin and R. M. Rahman, "Mining biomedical data from hypertext documents," in *14th International Conference on Computer and Information Technology (ICCCIT 2011)*, 2011, pp. 417–422.
- [71] W.-J. Hou, L.-C. Chen, and C.-S. Lu, "Identifying gene-disease associations using word proximity and similarity of Gene Ontology

- terms,” in 2011 4th International Conference on Biomedical Engineering and Informatics (BMEDI), 2011, pp. 1748–1752.
- [72] Z. Javed and H. Afzal, “Biomedical text mining for concept identification from traditional medicine literature,” in International Conference on Open Source Systems and Technologies, 2014, pp. 206–211.
- [73] M. Berardi, D. Malerba, and M. Attimonelli, “Mining Information Extraction Models for HmtDB annotation,” in Sixth IEEE International Conference on Data Mining - Workshops (ICDMW’06), 2006, pp. 207–212.
- [74] Jahiruddin, M. Abulaish, and L. Dey, “A concept-driven biomedical knowledge extraction and visualization framework for conceptualization of text corpora,” *J. Biomed. Inform.*, vol. 43, no. 6, pp. 1020–1035, Dec. 2010.
- [75] A. Sharma, R. Swaminathan, and H. Yang, “A Verb-Centric Approach for Relationship Extraction in Biomedical Text,” in 2010 IEEE Fourth International Conference on Semantic Computing, 2010, pp. 377–385.
- [76] Y. Li, X. Hu, H. Lin, and Z. Yang, “A Framework for Semisupervised Feature Generation and Its Applications in Biomedical Literature Mining,” *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 8, no. 2, pp. 294–307, Mar. 2011.
- [77] A. Holzinger, K.-M. Simoncic, and P. Yildirim, “Disease-Disease Relationships for Rheumatic Diseases: Web-Based Biomedical Textmining and Knowledge Discovery to Assist Medical Decision Making,” in 2012 IEEE 36th Annual Computer Software and Applications Conference, 2012, pp. 573–580.
- [78] R. T. Pereira, H. Costa, S. Carneiro, M. Rocha, and R. Mendes, “Reconstructing transcriptional Regulatory Networks using data integration and Text Mining,” in 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2015, pp. 1552–1558.
- [79] M. A. Landge and K. Rajeswari, “A Survey on Chemical Text Mining Techniques for Identifying Relationship Network between Drug Disease Genes and Molecules,” *Int. J. Comput. Appl.*, vol. 146, no. 1, pp. 5–9, 2016.
- [80] P. Tangtulyangkul, T. S. Hocking, and C. C. Fung, “Intelligent information mining from veterinary clinical records and open source repository,” in TENCON 2009 - 2009 IEEE Region 10 Conference, 2009, pp. 1–6.
- [81] C. Cano, A. Labarga, A. Blanco, and L. Peshkin, “Collaborative semi-automatic annotation of the biomedical literature,” in 2011 11th International Conference on Intelligent Systems Design and Applications, 2011, pp. 1213–1217.
- [82] H. Yang and Y. Dong, “Recognizing hierarchically related biomedical entities using MeSH-based mapping,” *Tsinghua Sci. Technol.*, vol. 17, no. 6, pp. 609–618, Dec. 2012.
- [83] S. Mohammed and A. Nazeer, “An improved method for extracting acronym-definition pairs from biomedical literature,” in 2013 International Conference on Control Communication and Computing (ICCC), 2013, pp. 194–197.
- [84] Y. Ji, Y. Tian, F. Shen, and J. Tran, “High-Performance Biomedical Association Mining with MapReduce,” in 2015 12th International Conference on Information Technology - New Generations, 2015, pp. 465–470.
- [85] J. Björne, F. Ginter, S. Pyysalo, J. Tsujii, and T. Salakoski, “Complex event extraction at PubMed scale,” *Bioinformatics*, vol. 26, no. 12, pp. i382–390, Jun. 2010.
- [86] M. Miwa, R. Saetre, J.-D. Kim, and J. Tsujii, “Event extraction with complex event classification using rich features,” *J. Bioinform. Comput. Biol.*, vol. 8, no. 1, pp. 131–46, Feb. 2010.
- [87] M. Naughton, N. Kushmerick, and J. Carthy, “Event Extraction from Heterogeneous News Sources,” 2006.
- [88] F. Hogenboom, F. Frasincar, U. Kaymak, F. de Jong, and E. Caron, “A Survey of event extraction methods from text for decision support systems,” *Decis. Support Syst.*, vol. 85, pp. 12–22, May 2016.
- [89] C. Jonquet et al., “NCBO Resource Index: Ontology-based search and mining of biomedical resources,” *Web Semant. Sci. Serv. Agents World Wide Web*, vol. 9, no. 3, pp. 316–324, Sep. 2011.
- [90] A. B. Can and N. Baykal, “MedicoPort: a medical search engine for all,” *Comput. Methods Programs Biomed.*, vol. 86, no. 1, pp. 73–86, Apr. 2007.
- [91] M. Ghanem, Y. Guo, and A. S. Rowe, “Integrated Data Mining and Text Mining In Support of Bioinformatics,” in Poster at Proceedings of the UK e-Science All Hands Meeting, 2004, pp. 1–3.
- [92] P. Zweigenbaum, D. Demner-Fushman, H. Yu, and K. B. Cohen, “New Frontiers in Biomedical Text Mining,” *Pacific Symp. Biocomput.*, vol. 12, pp. 205–208, 2007.

A Non-Linear Regression Modeling is used for Asymmetry Co-Integration and Managerial Economics in Iraqi Firms

Karrar Abdulellah Azeez

Harbin Institute of Technology, Kufa university
School of Management: Accounting
Harbin, China

Han DongPing

Harbin Institute of Technology
School of Management: accounting
Harbin, China

Marwah Abdulkareem Mahmood

Harbin Institute of Technology, Kufa university
School of Management: business administration
Harbin, China

Abstract—This paper analyzes the cost asymmetry through managerial expectations in a nonlinear regression function. Two development determinants, asymmetry co-integration and managerial expectations are also considered. The results revealed that managerial expectation had an impact on the wholesale cost asymmetry response. The managerial optimism is pronounced that show cost asymmetry response for sales, and inventory assets increased higher than decreased with the changing of the expectation basic coefficient and the values of contract parameters. Finally, the impacts of the managerial expectations, cost basic coefficient, and values of the contract parameters are analyzed for illustrating the results of the proposed nonlinear models with the help of numerical experiments. The research examined the short-run and the long-run effects of asymmetry co-integration and managerial expectation changes on the cost behavior in Iraq using the nonlinear regression function.

Keywords—Cost asymmetry; managerial expectations; co-integration; nonlinear regression function

I. INTRODUCTION

In the critical business environment, interdisciplinary concepts like the behavioral theory of the firms, which draws on economics, political science and organization theory, are imported in accounting works since the beginning of the research stream [1], [2]. Management expectation might be managerial optimism strengths or be managerial pessimism strengths about resource adjustments called sticky or anti-sticky behavior on cost asymmetry [3]. Managers likely rely on additional signals when their expectations are positive in the current period, and the activity level realization is high. They like to adjust capacity resources [4]. Moreover, many studies argued that relationship between cost and activity is not linear, but they depended on one driver to measure cost behavior [5], [6] found that traditional cost behavior model unsuitable to measure cost behavior, they provide an asymmetric response to cost and sales changes. Second explanation examined the managerial expectations about the future activity level, which is in turn driven by future demand that relates managerial

optimism and pessimism [7]. Some studies focused on the agency problem when managers make self-maximizing decisions that might not be in the best interest of the stockholders [8], [9].

Recently, literature has discussed the scientific question is there asymmetry co-integration between managerial expectations and cost response? To explain empirically how costs behave when management adjusts its costs and makes deliberate decisions as responding to certain factors [10], this evidence ignores the model of fixed and variable cost that assumed a mechanical relation between costs and activity change, and argues that the traditional model of cost behavior is not a fit framework to determine a benefit of the current period for future. Kama and Weiss [11] found the deliberate decisions to lessen the degree of costs sticky rather than induce cost sticky. While Bradbury and Scott [12] documented the deliberate managers decisions have not effect on costs respond to activity changes. In this study, we build a model of costs asymmetry by Cannon [13] and Chen, et al. [14]. Furthermore, costs are likely to vary with the levels of price, inventory, and demand differently than the level of sales [15]. The adjustment of costs in response to changes in activity volume is a primary issue in the company [16]. Chen, et al. [17] expected the managerial confidence has affected the degree of sticky costs. This adjustment may be cut or keep excess cost resources when sales increase and decrease because of future demand. This thinking considers that conscious adjustment of costs in the short term will be delayed. Management has an adjustment plan related to operational activities in the company [18].

In this context, this study attempts to provide some basis for responding to the evaluation of the impact of the managerial expectations and asymmetric cost information. One approach to addressing this problem is to employ the matching methods originally developed by Banker, et al. [19], credited by offering an interesting alternative to the analysis through these of nonlinear estimators. The works mentioned above studied the cost asymmetry in two determinants from sales and

assets, and considered the actors as managerial expectations. The study extends their works to asymmetry co-integration and managerial expectations using nonlinear regression function, and analyze the impact of the managerial expectations, the cost basis coefficient, and the values of contract parameters on the market policies. The paper is organized as follows. In Section II, we developed the centralized managerial expectations. In Sections III, two numerical examples are given to illustrate the solutions for proposed models. Section IV, summarizes the work.

II. MODELS AND SOULATION APPROACHES

The paper has applied an established methodology to develop the costs and activity relationships [13], [20], managers understand and performance in different situations [11], [21].

A. Sampling

Research examined monthly data for the Iraq over the period of 1 January 2006 to 31 December 2015 using industrial firms. The final samples consisted monthly of 600 usable observations of each variable but inventory assets were 400. We calculated all changes using the financial and performance statements across periods as indexes of total costs (Iraqi dinner), sales volumes (Ton) and output selling price and inventory value for using a non-linear function of multiple regression analysis. This data is described in Table 1.

TABLE I. DATA OF SAMPLING FROM 2006 - 2015

Number	Factory	Total cost. C/q	Sales	Inventory value. q*C
1	Najaf	120	120	64
2	Kufa	120	120	64
3	Smeawa	120	120	64
3	Busra	120	120	64
5	Karbala	120	120	64
Total sample		600	600	320

These items are determined from monthly statements of factories. Total costs are collection from operations costs plus selling and administrative costs by five activities (manufacturing, engineering & services, quality control, marketing, and administration). Sales revenue is (P*V). Inventory value is store quantity from produce last period based on factories statements.

B. Empirical Models

It is now a well-established fact to include the measures of economic activities in five industrial firms as well as a measure of managerial expectations. Sales and inventory assets level change as two main determinants of the cost stickiness. Therefore, we have designed our model with the following long-run specification [3], [14]:

$$\ln \frac{TC_{i,t} - TC_{i,t-1}}{TC_{i,t-1}} = \varphi_0 + \varphi_1 \ln \left(\frac{REV_{i,t} - REV_{i,t-1}}{REV_{i,t-1}} \right) + \varphi_2 DEC_{i,t} \ln \left(\frac{REV_{i,t} - REV_{i,t-1}}{REV_{i,t-1}} \right) + \alpha_{i,t} \quad (1)$$

Where: $TC_{i,t}$ is a total cost for firm i time t . $R_{i,t}$ is sales revenue for firm i time t . $DEC_{i,t}$ is an indicator variable set value of 1 when $R_{i,t} < R_{i,t-1}$ for firm i time t , and set value of 0 otherwise. φ_0 is a parameter that estimates the asymmetric cost changes unassociated with revenues change. φ_1 is the parameter that estimates the association between cost change and revenue increase. φ_2 is the parameter of “asymmetry measure” that estimates the association between cost response and revenue change during increasing and decreasing. $\alpha_{i,t}$ is an error term for variability cost change estimation for firm i time t . As argued by Anderson et al. [5], this measure of the cost stickiness is unit free and it allows us to specify the model in the logarithmic form that fits the macro data better. Furthermore, the measure is defined as the ratio of revenues over cash flows so that if this measure is to improve due to a depreciation of firm’s performance, an estimate of φ_2 to be negative. However, as argued by Kama and Weiss [11], these income elasticities could also be negative and positive respectively, if prior sales decrease and increase as it grows. The parameters of activity function and manufacturing cost are all characterized as fuzzy variables [22].

Proposition 1: Optimistic expectations generate stickiness behavior of cost by sales change. The cost response is a non-linear function for managerial expectations.

The coefficient estimates we discussed above are the long-run estimates. In order to also infer the short-run effects of all the exogenous variables we need to turn (1) into an error-correction specification [23].

Prior to the introduction of asymmetry co-integration by Chen et al. [14] it was a common practice to just estimate (2) and judge the managerial expectations as a short-run positive or insignificant β_1 coefficient combined with a significant negative $-\beta_2$ coefficient. However, as mentioned before, Balakrishnan et al. [24], [25] demonstrated that the insignificance of the short-run and long-run estimates could be due to assuming the effects of asset intensity changes to be symmetric. They then followed Shin et al. (2014) and modified (2) so that one can assess the asymmetry effects of Asset Intensity changes. Under this new method, we first form $\Delta \ln \text{INVAS}$ which includes negative values reflecting decreasing prior flows and positive values, reflecting increasing current flows. Using these changes, we then construct two new parameters and define them as $\Delta \ln \text{INVAS}$, partial sum of positive changes and $DEC * \Delta \ln \text{INVAS}$, partial sum of negative changes. These two new variables now reflect only prior and current flows, respectively. Thus, we estimate variables to arrive at:

$$\ln \frac{TC_{i,t} - TC_{i,t-1}}{TC_{i,t-1}} = \beta_0 + \beta_1 \ln \left(\frac{\text{INVAS}_{i,t} - \text{INVAS}_{i,t-1}}{\text{INVAS}_{i,t-1}} \right) + \beta_2 DEC_{i,t} \ln \left(\frac{\text{INVAS}_{i,t} - \text{INVAS}_{i,t-1}}{\text{INVAS}_{i,t-1}} \right) + \vartheta_{i,t} \quad (2)$$

Since construction of $\Delta \ln \text{INVAS}$ and $DEC * \Delta \ln \text{INVAS}$ variables using partial sum methods introduce non-linearity to the adjustment process, Chen et al. [26] and Chen et al. [14] call specification (2) as the non-linear regression model.

Where: $INV_{i,t}$ is a total inventory assets value for firm i time t . $DEC_{i,t}$ is an indicator variable set value of 1 when $INVAS_{i,t} < INVAS_{i,t-1}$ for firm i time t , and set value of 0 otherwise. β_0 is a parameter that estimates the asymmetric cost changes unassociated with inventory assets changes. β_1 is the parameter that estimates the association between cost response and inventory assets change during periods when inventory asset is increasing. β_2 is the parameter of “asymmetry measure” that estimates the difference in the association between cost response and inventory assets change during increasing and decreasing. $\vartheta_{i,t}$ is an error term for variability cost change estimation for firm i time t .

Proposition 2: Optimistic expectations generate stickiness behavior of cost by assets change. The cost response is a non-linear function for managerial expectations.

This is expected to be the case in most cost stickiness models due to the fact that assets increase and decrease in five different firms that are subject to different business rules and regulations. Second, the long-run asymmetric effects of managerial expectations on the cost stickiness will be established if Max Eigenvalue Statistics coefficient is higher than scheduled coefficient, at each variable. Third, impact asymmetry will be established if $\Sigma \Delta \ln INVAS, \Delta \ln REV_{i,t} \neq \Sigma DEC * \Delta \ln INVAS, DEC * \Delta \ln REV_{i,t}$ 1,2. Finally, long-run asymmetric effects of Managerial expectations on the cost stickiness will be established if φ_2, β_2 have negative values. The cost behavior is stickiness when the average percentage

increase is higher or less than average percentage decrease in costs. The empirical hypothesis for sticky behavior means that $\varphi_2, \beta_2 < 0$ and opposite that means anti-sticky behavior. This finding provides an empirical test of H1 and H2. The variables definitions are presented in Table 2.

TABLE II. DATA DEFINITION AND RELATIONS AMONG VARIABLES IN MODELS

Variable	Calculation	Description
$\ln\left(\frac{TC_{i,t} - TC_{i,t-1}}{TC_{i,t-1}}\right)$	Percent total cost change	Log-change in total costs by dinar. Payments of all industrial, marketing and administration activities.
$\ln\left(\frac{R_{i,t} - R_{i,t-1}}{R_{i,t-1}}\right)$	Percent total sales revenue.	Log-change in total net revenue by dinar.
$\ln\left(\frac{INVAS_{i,t} - INVAS_{i,t-1}}{INVAS_{i,t-1}}\right)$	Percent total inventory asset	Log-change in total inventory asset by dinar.

C. Preliminary Analysis

Descriptive statistics from a sample for costs, capacities, and their changes are presented in Table 3. The mean sales revenue is IQD 1932 million (median IQD 1332 million). The mean total cost is IQD 2131 million (median IQD 1433 million). The mean inventory assets is IQD 7 million (median IQD 0.37 million). On average, the magnitude of changes in total cost, sales and assets, mean (median) sales revenue is 3579 (0.00) percent. Total cost is 42 (13) percent. Inventory asset is 1328 (0.00). Consistent with prior studies [6], [13].

TABLE III. DESCRIPTION STATISTICS

Variable	Mean	Standard Dev.	Median	Maximum	Minimum
Total costs	2131174860	1813379509	1433865019	9973095303	36103999
Total costs %	0.423	2.223	0.130	27.750	0.000
Sales revenue	1932273460	2020397695	1332414963	8982796000	0.000
Sales revenue %	3.579	41.212	0.000	899.64	-1.00
inventory assets	7685291	11012039	372882	41347328	230
inventory assets %	1.328	20.07	0.000	356.65	-1.00

All numbers of costs reported in Iraqi dinar (IQD).

TABLE IV. RESULTS OF AUGMENTED DICKEY-FULLER TESTS: STATIONARY ANALYSIS

Variable	Coefficient	Standard Error	Critical value	t-statistics (Prob.*)
$\ln\left(\frac{TC_{i,t} - TC_{i,t-1}}{TC_{i,t-1}}\right)$	-1.179 (-)	0.04	(-2.866)	-29.28*** (0.000)
$\ln\left(\frac{REV_{i,t} - REV_{i,t-1}}{REV_{i,t-1}}\right)$	-1.009 (-)	0.04	(-2.866)	-24.65*** (0.000)
$DEC_{i,t} \ln\left(\frac{REV_{i,t} - REV_{i,t-1}}{REV_{i,t-1}}\right)$	-1.15 (-)	0.04	(-2.866)	-28.61*** (0.000)
$\ln\left(\frac{INVAS_{i,t} - INVAS_{i,t-1}}{INVAS_{i,t-1}}\right)$	-1.004 (-)	0.06	(-2.87)	-17.88 (0.000)
$DEC_{i,t} \ln\left(\frac{INVAS_{i,t} - INVAS_{i,t-1}}{INVAS_{i,t-1}}\right)$	-0.86 (-)	0.05	(-2.87)	-15.46*** (0.000)

Reject the null of non-stationarity at the 5% level, significant indicates *, **, *** at the 1%, 5%, 10% level respectively.

D. Test of Stationarity

The stationary test is a good econometric practice to restricted co-integrating vectors to establish whether relevant restrictions are rejected or not [25], [27]. Table 4 presents the results of Augmented Dickey-Fuller tests. All variables are rejected the null hypothesis of a unit root that the empirical variables are stationary. Next, we test for co-integration applying the Johansen technique in four separate models.

As expected, all empirical variables were negative (δ_1 (0.04 = -1.179, $p < 0$), and the results from the test for existence or not of a unit root in the log levels of our variables. The statistical values are greater than the critical values rejecting the null hypothesis of the unit root. Therefore, all our variables are integrated [28].

E. Co-integration Tests

Multivariate results are from the Johansen trace and maximum eigenvalue statistics on co-integration for the empirical models are presented in Table 5. The theory of co-integration provides a natural setting for testing cross-variables relationships in permanent output movements [29]. The two statistics for the test give full co-integrating vectors for study variables. The cointegrating test explains that the relationship between managerial expectation and costs asymmetry is long-run. The Johansen trace and the maximum eigenvalue statistics are rejected the null hypothesis implies that there are co-integrating vectors at the 5% level for the entire two-model variables ($r \geq 0$, $r \geq 1$ and $r \geq 2$).

TABLE V. RESULTS FROM JOHANSEN CO-INTEGRATION TESTS

Model	Null	Eigenvalue	Trace Statistics	Max. Eigen. Stat.
$\ln \frac{TC_{i,t} - TC_{i,t-1}}{TC_{i,t-1}} = \varphi_0 + \varphi_1 \ln \left(\frac{REV_{i,t} - REV_{i,t-1}}{REV_{i,t-1}} \right) + \varphi_2 DEC_{i,t} \ln \left(\frac{REV_{i,t} - REV_{i,t-1}}{REV_{i,t-1}} \right) + \alpha_{i,t}$	None *	0.49	15.54** (0.050)	10.95 (0.156)
	At most 1 *	0.24	4.50** (0.033)	4.50** (0.033)
	At most 2 *	0.14	91.56*** (0.000)	91.56*** (0.000)
$\ln \frac{TC_{i,t} - TC_{i,t-1}}{TC_{i,t-1}} = \beta_0 + \beta_1 \ln \left(\frac{INVAS_{i,t} - INVAS_{i,t-1}}{INVAS_{i,t-1}} \right) + \beta_2 DEC_{i,t} \ln \left(\frac{INVAS_{i,t} - INVAS_{i,t-1}}{INVAS_{i,t-1}} \right) + \vartheta_{i,t}$	None *	0.23	175.47*** (0.001)	83.89*** (0.000)
	At most 1 *	0.17	91.57*** (0.000)	61.47*** (0.000)
	At most 2 *	0.091	30.09*** (0.000)	30.09*** (0.000)

Reject the null of no co-integration among empirical variables at the 5% level.

The results indicate that co-integration is accepted all of the empirical models in the full estimates of co-integrating vectors at the 5% level. This suggests an evidence of nonlinear

modeling linkages between managerial expectation and costs asymmetry relationship and allows examining the hypotheses by nonlinear regression analysis in the next part.

TABLE VI. ESTIMATED REGRESSION MODEL AND LONG RUN COEFFICIENT: NONLINEAR ANALYSIS

Variable	Parameter	Coefficient	Standard Error	T-ratio [prob]
Panel A: Regression analysis: effects of sales revenue – model 1				
Intercept		0.55 (?)		
$\ln \left(\frac{REV_{i,t} - REV_{i,t-1}}{REV_{i,t-1}} \right)$	φ_0	0.792 (+)	0.03	1.77[0.038]**
$DEC_{i,t} \ln \left(\frac{REV_{i,t} - REV_{i,t-1}}{REV_{i,t-1}} \right)$	φ_1	-0.753 (-)	0.25	3.096 [0.022]**
Adjusted R ²		0.36	0.13	-5.71[0.00] ***
F-value		45.32		
Significant level		0.000		
Panel B: Regression analysis: effects of inventory asset – model 2				
Intercept		0.084 (?)		
$\ln \left(\frac{INVAS_{i,t} - INVAS_{i,t-1}}{INVAS_{i,t-1}} \right)$	β_0	0.854 (+)	0.06	1.45[0.146]**
$DEC_{i,t} \ln \left(\frac{INVAS_{i,t} - INVAS_{i,t-1}}{INVAS_{i,t-1}} \right)$	β_1	-0.208 (-)	0.41	2.08[0.029] **
Adjusted R ²	β_2	0.40	0.04	-4.55[0.00] ***
F-value		16.51		
Significant level		0.000		

All t-statistics were calculated by using significant indicate *, **, *** at the 1%, 5%, 10% level respectively.

III. NUMERICAL EXAMPLES

In this section, we provide two numerical examples to show determinants of Asymmetric effects of managerial expectations on cost response. Results of non-linear regression analysis show the effect of managerial expectations on cost asymmetry (H1-H2). Results show the models are significant as a whole (F-value 45.32, 16.51 for model 1 and 2, respectively, p -value <0.001), and reasonably explains the dependent variables (Adj. R^2 31 and 36 percent for two models respectively). All explanatory variables show the significant main effects. Their details are shown above in Table 6.

As Table 6 shows, sales change is asymmetrically and significantly related to asymmetric behavior of costs, costs behavior is sticky ($\varphi_1 > 0$, $\varphi_2 < 0$, $p < 0.01$) and different from zero at the 1% (t-statistics -5.71), the adjusted R^2 is 36%. On average, costs increase 0.80% per 1% increase in sales revenues (φ_1) and they decrease by 0.05% per 1% decrease in sales revenues ($\varphi_1 + \varphi_2$); see model 1. The result shows a direct effect of sales change on cost behavior during increasing and decreasing periods. Thus, proposition 1 is supported.

Model 2 shows, inventory assets change is asymmetrically and significantly related to asymmetric behavior of costs, costs behavior is sticky ($\beta_1 > 0$, $\beta_2 < 0$, $p < 0.001$) and different from zero at the 1% (t-statistics -4.55), the adjusted R^2 is 40%. On average, costs increase 0.85% per 1% increase in inventory change (β_1) and they decrease by 0.64% per 1% decrease in inventory change ($\beta_1 + \beta_2$); see model 2. The results show a direct effect of sales change on cost behavior during increasing and decreasing periods. The difference between these coefficients captures the degree of cost asymmetry. Thus, proposition 2 is supported.

IV. DISCUSSION

These results proved that costs are the description of a broader pattern of asymmetric cost behavior, which extends to all the major components of costs for physical input quantity (sales and assets) for cost behavior. Results suggest that asymmetric behavior of costs may be difficult to reduce inventory assets costs related to managerial expectations in the short term, the evidence provides direct support for the managerial expectations on the cost structure. On the contrary, Bradbury and Scott [12] found no differences between actual and forecast sample when sales revenues increase and decrease, The estimated value of φ_2 in actual and forecast regression is equal to -0.35%, and -0.21%, respectively. Whilst [11] agree with our results they found there is an effect on cost asymmetry with and without sensitive. The estimated value of φ_2 regression is equal to -0.025%, and -0.092%. Furthermore, Ibrahim [21] agrees with results found that the costs behavior is sticky in prosperity periods, and cost behavior is anti-sticky in recession periods. The estimated value of φ_2 regression is equal to -0.48%, and 0.20% during prosperity and recession respectively. This finding means estimation of costs asymmetry has associated with inventory changes by setting the cost based on competition and considers the inventory changes are a new driver to measure asymmetric cost behavior. Inventory increase relates to sales increase may the demand for capacity utilization is falling or there are positive expectations

about future [30]. Anderson, et al. [20] Argue when we add the asset's elements to the basic asymmetric cost behavior model, we can find economic meaning. The effect of demand uncertainty on the order quantity and wholesale price has investigated by fuzzy random methods, and compared to the conditions of buyback policies [31]. The significant anti-sticky costs made when activity changes decrease in previous periods, and significant sticky costs when activity changes increase in previous periods [3]. These differences in estimates of cost behavior due to managers do not consider the effect of managerial optimism about future [20]. This finding applies the managerial optimism future and moves asymmetric behavior phenomenon for providing a new evidence that associated the managerial estimation with anti-sticky and sticky cost behavior in different positions.

V. CONCLUSIONS

This article examines the asymmetry co-integration between managerial expectations and cost response, as well as sales and inventory change, in Iraqi industry using nonlinear function modeling developed by Anderson et al. [5] and Chen et al. [14]. Once non-linear modeling and co-integration were introduced, the definition of the cost asymmetry was modified to mean short-run expectations combined with long-run improvement. Now that asymmetry co-integration has been advanced, the definition has also been modified further to mean short-run expectations or insignificant effects combined with long-run improvement only due to only expectations or short-run insignificant effects and long-run expectations only due to adjustment costs. The last approach is which requires separating currency expectations from appreciations and using a non-linear cost asymmetry model. This approach also allows us to determine if cost level changes have symmetric or asymmetric effects via managerial expectations. The results revealed that the change in the expectation basic coefficient impact on the wholesale cost response. Second, evidence of short-run asymmetric effects of sales and assets changes in cost response, significant short-run and long-run asymmetric effects were established in Iraqi industry. Third, asymmetric cost behavior was found for managerial expectations by non-linear function in Iraqi industry. Finally, asymmetry analysis revealed that while managerial expectations against the competitive environment have favorable effects on the asymmetric cost behavior of the industry.

One limitation of this article is that we only consider one determinant of the cost asymmetry phenomenon. Therefore, one possible extension work is to study the cost asymmetry with multiple determinants in non-linear function modeling. In fact, the cost function of the asymmetric model can be non-linear. One can consider the case the sales and assets are asymmetric random variables. This study contributes to our knowledge of how and when managerial expectations can be influenced into costs. Our study also empirically validates asymmetry co-integration as a mechanism that accounts for the relationship between managerial expectations and costs response. In addition, this research emphasizes the importance of managerial economics, which determines whether managerial expectations have a positive or negative effect on the cost structure. We hope that this study will stimulate future

endeavors to advance our understanding of the relationship between managerial expectations and cost asymmetry.

ACKNOWLEDGEMENT

The authors would like to thank Iraqi Cultural Attaché in China and national company for supporting a research about data collecting and their corresponding for our academic goal.

REFERENCES

- [1] E. H. Caplan, "Behavioral assumptions of management accounting," *The Accounting Review*, vol. 41, pp. 496-509, 1966.
- [2] R. D. Banker, R. Huang, and R. Natarajan, "Equity incentives and long-term value created by SG&A expenditure," *Contemporary Accounting Research*, vol. 28, pp. 794-830, 2011.
- [3] R. D. Banker, D. Byzalov, M. Ciftci, and R. Mashruwala, "The moderating effect of prior sales changes on asymmetric cost behavior," *Journal of Management Accounting Research*, vol. 26, pp. 221-242, 2014.
- [4] B. De Langhe, S. M. van Osselaer, and B. Wierenga, "The effects of process and outcome accountability on judgment process and performance," *Organizational Behavior and Human Decision Processes*, vol. 115, pp. 238-252, 2011.
- [5] M. C. Anderson, R. D. Banker, and S. N. Janakiraman, "Are selling, general, and administrative costs "sticky"?," *Journal of Accounting Research*, vol. 41, pp. 47-63, 2003.
- [6] M. L. Weidenmier and C. Subramaniam, "Additional evidence on the sticky behavior of costs," 2003.
- [7] F. Rouxelin, W. Wongsunwai, and N. Yehuda, "Aggregate Cost Stickiness in GAAP Financial Statements and Future Unemployment Rate," Available at SSRN 2547789, 2015.
- [8] I. Kama and D. Weiss, "Do earnings targets and managerial incentives affect sticky costs?," *Journal of Accounting Research*, vol. 51, pp. 201-224, 2013.
- [9] B. Qin, A. W. Mohan, and Y. F. Kuang, "CEO Overconfidence and Cost Stickiness," *Management Control & Accounting* (2), pp. 26-32, 2015.
- [10] M. Porporato and E. M. Werbin, "Active cost management in banks: evidence of sticky costs in Argentina, Brazil and Canada," 2011.
- [11] I. Kama and D. Weiss, "Do managers' deliberate decisions induce sticky costs? Israel: Henry Crown Institute of Business Research 2010.
- [12] M. E. Bradbury and T. Scott, "Do Managers Understand Asymmetric Cost Behavior," presented at the Auckland Region Accounting Conference 2014.
- [13] J. N. Cannon, "Determinants of "sticky costs": An analysis of cost behavior using United States air transportation industry data," *The Accounting Review*, vol. 89, pp. 1645-1672, 2014.
- [14] J. V. Chen, I. Kama, and R. Lehavy, "Management Expectations and Asymmetric Cost Behavior," pp. 1-34, 2015.
- [15] M. Bugeja, M. Lu, and Y. Shan, "Cost Stickiness in Australia: Characteristics and Determinants," *Australian Accounting Review*, vol. 25, pp. 248-261, 2015.
- [16] N. Dalla Via, "Three essays in behavioral management accounting," 2012.
- [17] C. X. Chen, T. Gores, and J. Nasev, "Managerial overconfidence and cost stickiness," *global management accounting research symposium*, pp. 1-40, 2013.
- [18] G. Blue, E. Moazed, D. Khanhossini, and M. Nikoonesbati, "The Relationship between Perspective Managers and 'Sticky Costs' in the Tehran Stock Exchange," Available at SSRN 2216631, 2013.
- [19] R. D. Banker, M. Ciftci, and R. Mashruwala, "Managerial optimism, prior period sales changes, and sticky cost behavior," 2008.
- [20] M. C. Anderson, J. H. Lee, and R. Mashruwala, "Cost Stickiness and Cost Inertia: A Two-Driver Model of Asymmetric Cost Behavior," *Accounting, Organizations, and Society* pp. 1-36, 2016.
- [21] A. E. A. Ibrahim, "Economic growth and cost stickiness: evidence from Egypt," *Journal of Financial Reporting and Accounting*, vol. 13, pp. 119-140, 2015.
- [22] S. WANG, "A Manufacturer Stackelberg Game in Price Competition Supply Chain under a Fuzzy Decision Environment," *International Journal of Applied Mathematics*, vol. 47, 2017.
- [23] M. Bahmani-Oskooee, M. Aftab, and H. Harvey, "Asymmetry cointegration and the J-curve: New evidence from Malaysia-Singapore commodity trade," *The Journal of Economic Asymmetries*, 2016.
- [24] R. Balakrishnan, E. Labro, and N. S. Soderstrom, "Cost structure and sticky costs," *Journal of management accounting research*, vol. 26, pp. 91-116, 2014.
- [25] F. Zanella, P. Oyelere, and S. Hossain, "Are costs really sticky? Evidence from publicly listed companies in the UAE," *Applied Economics*, vol. 47, pp. 6519-6528, 2015.
- [26] C. X. Chen, T. Gores, and J. Nasev, "Managerial overconfidence and cost stickiness," Available at SSRN 2208622, 2013.
- [27] P. Österholm and E. Hjalmarsson, "Testing for cointegration using the Johansen methodology when variables are near-integrated: International Monetary Fund, 2007.
- [28] J. G. MacKinnon, "Numerical distribution functions for unit root and cointegration tests," *Journal of applied econometrics*, pp. 601-618, 1996.
- [29] D. Asteriou, S. Karagianni, and C. Siriopoulos, "Testing The Convergence Hypothesis Using Time Series Techniques: The Case Of Greece 1971-1996," *Journal of Applied Business Research (JABR)*, vol. 18, 2011.
- [30] S. W. Anderson and W. N. Lanen, "Understanding Cost Management: What Can We Learn from the Evidence on 'Sticky Costs'?", Available at SSRN 975135, 2007.
- [31] S. Sang, "Buyback Contract with Fuzzy Demand and Risk Preference in a Three Level Supply Chain," *International Journal of Applied Mathematics*, vol. 46, 2016.

AUTHOR PROFILE

Han DongPing is a professor of Accounting at Harbin institute of technology in China, Harbin. She is Dean of management school at Harbin Institute of Technology on Weihai city. Karrar Abdulelah Azeez is a lecturer of Accounting at Kufa University in Iraq, Najaf. He is PhD candidate at Harbin institute of technology in China. Marwah Abdulkareem Mahmood is a lecturer of Business Administration at Kufa University in Iraq, Najaf. She is PhD candidate at Harbin institute of technology in China.

DDoS Attacks Classification using Numeric Attribute-based Gaussian Naive Bayes

Abdul Fadlil

Department of Electrical Engineering
Ahmad Dahlan University
Yogyakarta, Indonesia

Imam Riadi

Department of Information System
Ahmad Dahlan University Ahmad
Dahlan University
Yogyakarta, Indonesia

Sukma Aji

Department of Information
Technology
Ahmad Dahlan University
Yogyakarta Indonesia

Abstract—Cyber attacks by sending large data packets that deplete computer network service resources by using multiple computers when attacking are called Distributed Denial of Service (DDoS) attacks. Total Data Packet and important information in the form of log files sent by the attacker can be observed and captured through the port mirroring of the computer network service. The classification system is required to distinguish network traffic into two conditions, first normal condition, and second attack condition. The Gaussian Naive Bayes classification is one of the methods that can be used to process numeric attribute as input and determine two decisions of access that occur on the computer network service that is “normal” access or access under “attack” by DDoS as output. This research was conducted in Ahmad Dahlan University Networking Laboratory (ADUNL) for 60 minutes with the result of classification of 8 IP Address with normal access and 6 IP Address with DDoS attack access.

Keywords—Distributed Denial of Service (DdoS); Gaussian Naive Bayes; Numeric

I. INTRODUCTION

Distributed Denial of Service (DDoS) attacks still top the list of Cyber Attacks. In Open Source Intelligence by month January reported an unusually low number of Attack Techniques shows 34% of the cases, the reason was not specified. Where as DDoS leads the chart of the known techniques with 22.3%, ahead Hijacks (13.8%), and Defacements (10%). Targeted attacks are immediately behind with a remarkable 7.4%. Fig. 1 shows attacks technique until January 2016. Data shows that DDoS attacks are still always very interesting to be the object of the research.¹

DDoS attacks through computer networks, especially Local Area Network (LAN) are detected using a multi-classification technique, that is, by combining data mining method to get better accuracy. In pre-processing data, before loading data sets into data mining software, relevant attributes are selected to get accurate and unused classification omitted because it will add noise that affect accuracy [1].

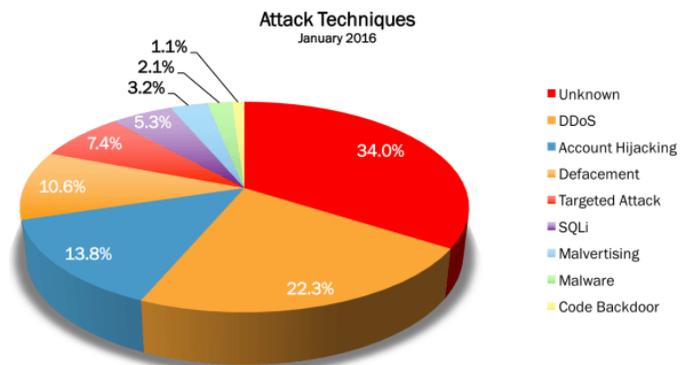


Fig. 1. Top 9 of Attack Techniques January 2016.

In research [2] the Comparative Analysis of Different DDoS Detection Techniques used Statistical Method, Intrusion Detection System (IDS), IDS based Dempster-Shafer Theory, Host Based IDS, Network IDS, and Real Time IDS of Throughput, Fault Tolerance, Performance, Overheads, Response Time, and Detection Rate.

Gülay Öke [3] used Multiple Bayesian Classifier and Random Neural Network to detect Denial of Service attacks. Naive Bayes Classifier makes a decision by collecting offline input features. The input feature is bit rate, an increase in bit rate, entropy value of the incoming bit rate, Hurst parameter, delay, and Delay rate. Bharti Nagpal [4] comparing 5 DDoS attack tools Trinity, Low Orbit Ion Cannon (LOIC), Tribal Flood Network, Mstream, and Trinoo as Architecture used, Type of Flooding used for attacking, Type of DDoS method used, Possible damage caused, Channel encryption. Gnanapriya [5] research Software-Defined Networking (SDN) shows that SDN provides a new opportunity to defeat DDoS attacks in cloud computing environments, and summarizes the excellent SDN features to defeat DDoS attacks. Then review the study of the launch of DDoS attacks on SDN and methods against DDoS attacks on the SDN.

¹ <http://www.hackmageddon.com>

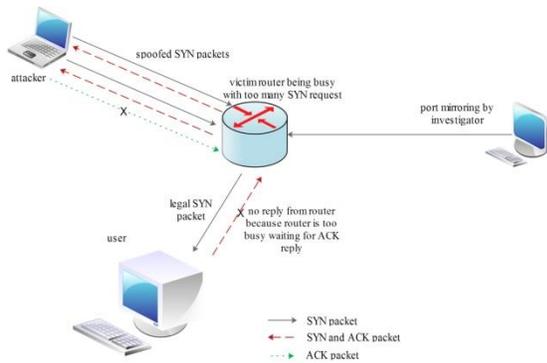


Fig. 2. TCP SYN flood attack.

Normal TCP connections usually start transmitting from the user by sending SYN to the router, and the router will allocate the buffer to the user and respond with SYN and ACK packets. This stage, the connection is in a half open state, waiting for an ACK response from the user to complete the connection settings. When the connection is completed, this is called 3-way linkage and TCP SYN Flood attacks manipulate this 3-way linkage by making the router busy with SYN request [6]. TCP SYN Flood is a common form of Denial of Service attack. Fig. 2 shows the TCP SYN Flood happened. TCP SYN Flood can be observed with a Packet Capture application by using a port mirroring to observe a copy of router activity. TCP SYN flood features are often the emergence of one of the IP Address to the router. The source IP Address that always appears to the router is calculated within a specified time range and used as feature extraction as a DDoS attack [7].

Based on earlier research regarding packet classification with Naive Bayes, in this paper, we provide a detailed understanding of how to process numerical attributes on a network traffic activity based on the Gaussian Naive Bayes method.

II. BASIC THEORY

A. Gaussian Method

The Gaussian method is one of the common and important methods in probability and statistics, introduced by Gauss in his study of error theory. Gauss uses it to describe errors. Experience shows that many random variables, the height of adult males, and reaction time in psychological experiments, all of which can be solved by the Gaussian Method [8], [9]. The Gaussian method is:

$$P(x) = \frac{1}{\delta\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\delta^2}} \quad (1)$$

Where, μ is average and δ is standard deviation, to calculate μ and δ values for numerical attributes using formula

$$\mu = \frac{\sum_{i=1}^n x_i}{n} \quad (2)$$

$$\delta^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n-1} \quad (3)$$

B. Naive Bayes Method

Bayes method is used to calculate the probability of occurrence of an event based on the observed effects of observation. Naive Bayes method is simple probabilistic-based prediction technique based on Bayes's method application with strong independence assumptions [10]. Naive Bayes method is:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (4)$$

Where,

$P(A|B)$ is the posterior of class (target) given predictor (attribute).

$P(B|A)$ is the likelihood which is the probability of predictor given class.

$P(A)$ is the prior probability of class.

$P(B)$ is the prior probability of predictor.

C. Accuracy

The accuracy of a classification system is described as the data output level compared to the desired value. Accuracy in classification is calculated from:

- Normal access data in a normal class (True Positives (TP)).
- Normal access data outside the normal class (False Positives (FP)).
- Attacks access data outside the attack class (False Negatives (FN)).
- Attack access data in the attack class (True Negatives (TN)) [9].

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (5)$$

III. RESEARCH METHODOLOGY

A. Topology

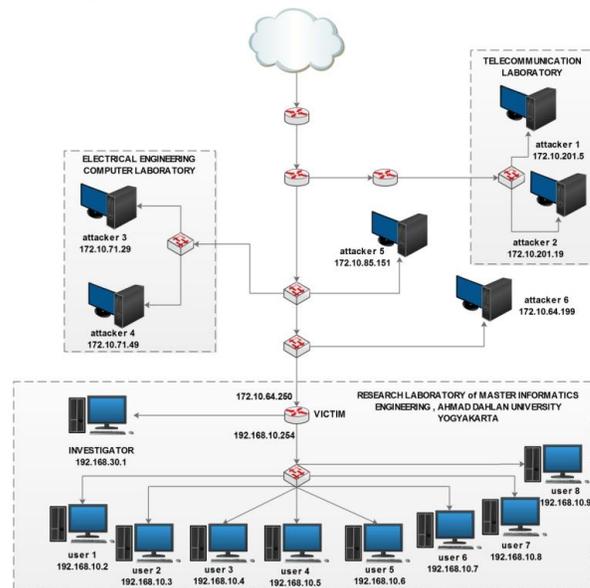


Fig. 3. Research laboratory of master informatics engineering topology.

Computer Network of ADUNL topology shown in Fig. 3 is distributed, the development of star topology. Router with IP Address 172.10.64.250 and 192.168.10.254 become the network service center and access divider of each user within the scope of ADUNL.

B. Attacks Scenario

IP address 192.168.10.64.2; 192.168.10.64.3; 192.168.10.64.4; 192.168.10.64.5; 192.168.10.64.6; 192.168.10.64.7; 192.168.10.64.8; and 192.168.10.64.9 (user) perform normal activities by accessing the site www.detik.com and www.youtube.com and run the function in the site by pressing play movie button.

The attack is done from outside ADUNL to victim router with IP address 172.10.64.250 by an attacker with IP address 172.10.64.199; 172.10.85.151; 172.10.71.29; 172.10.71.49; 172.10.201.5; and 172.10.201.19 using DDoS attack tool Low Orbit Ion Canon (LOIC).

Investigator use port mirroring access with IP address 192.168.30.1. To retrieve log data of network traffic from within and to ADUNL.

C. Methodology

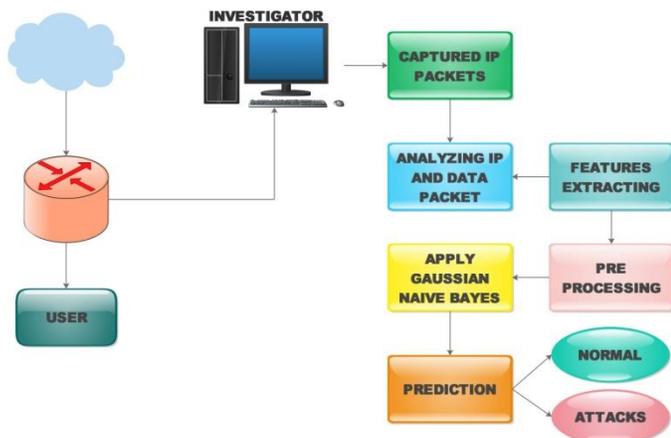


Fig. 4. Methodology of DDoS attacks classification.

DDoS attacks classification step of the methodology is shown in Fig. 4.

- Captured IP packet is used to retrieve data in the form of log file network traffic with port mirroring access in .pcap format.
- Analyzing IP and data packet, in this step is to analyze the IP address who is doing the attack and how long the packet is sent.
- Extraction, in this stage log files with the .pcap format, is converted into spreadsheet files so they can be processed using Gaussian Naive Bayes method.
- Pre-processing, at this step the making of input parameters can be used in the classification method.

- Apply Gaussian Naive Bayes, at this stage Gaussian Naive Bayes classification method, is used to process data that already has input parameters.
- Prediction, at this step Gaussian Naive Bayes method, determines the data that has been processed into two decisions that are normal access or under attack.

IV. RESULT AND ANALYSIS

Object research result capture network traffic at ADUNL. The methodological step is carried out coherently to produce maximum research.

A. Captured IP Packet Result

Log file of captured network traffic for 60 minutes divide within 3 minutes each time access through port mirroring ADUNL by the investigator using Wireshark packet capture in .pcap format. Fig. 5 shows capture result in .pcap format.

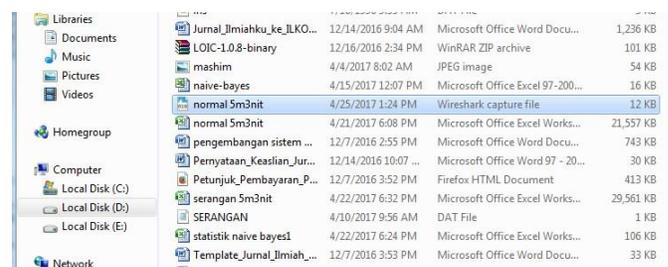


Fig. 5. Capture result in .pcap format.

B. Analyzing IP and data packet

IP address that accesses ADUNL and estimates how many packets of data transmitted by and from the IP address that is doing the activity calculated based on log files that have been obtained. Fig. 6 shows the IP address accessing ADUNL.

No.	Time	Source	Destination	Protocol	Length	Info
2353	2017-02-13 14:37:08.956900	192.168.10.8	101.203.171.78	QUIC	128	Payload
2353	2017-02-13 14:37:08.957171	101.203.171.78	192.168.10.8	QUIC	1439	Payload
2353	2017-02-13 14:37:08.957175	192.168.10.8	101.203.171.78	QUIC	128	Payload
2353	2017-02-13 14:37:08.957177	101.203.171.78	192.168.10.8	QUIC	1439	Payload
3330	2017-02-13 15:25:08.981862	172.10.64.250	172.10.64.199	TCP	101	80->6214
3331	2017-02-13 15:25:08.982126	172.10.64.199	172.10.64.250	TCP	149	[TCP se
3332	2017-02-13 15:25:08.982127	172.10.64.250	172.10.64.199	TCP	101	80->6214
3333	2017-02-13 15:25:08.982127	172.10.201.5	172.10.64.250	TCP	133	[TCP se

Fig. 6. IP address accessing in ADUNL.

C. Extraction

Capture results of network traffic log files in .pcap format can not be processed into columns and rows required in the classification process. To be processed into columns and rows of .pcap format are extracted into the .csv format and then extracted into xlsx format. Fig. 7 shows extracting .pcap format into .csv format.

D. Pre-processing

At this stage, it is processing the results of network traffic extraction into the main parameters that can identify normal access or attack. The main parameters used as input parameters shown in Table 1. In this research, two input parameters taken are:

- Incoming of IP address (IIP) within specified time range (2nd column is x attribute).
- Packet length (PL) within a specified time range (3rd column is y attribute).

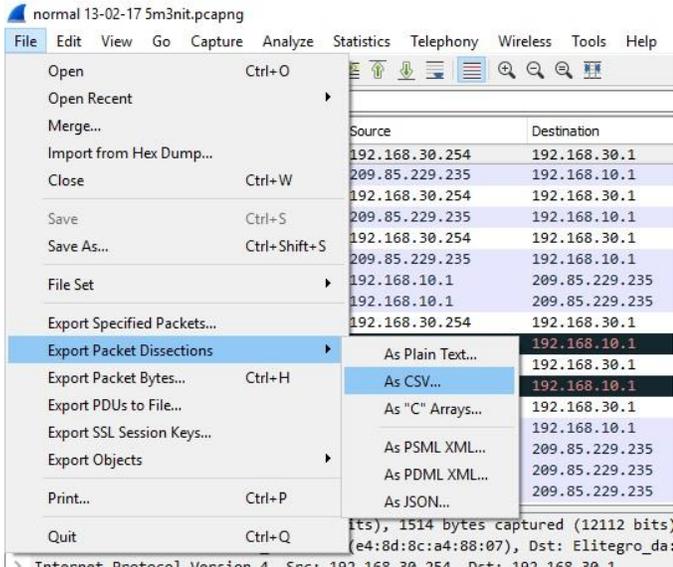


Fig. 7. Extracting .pcap format into .csv format.

TABLE I. INPUT PARAMETERS IN TIME RANGE 0-3 MINUTES

IP address	Incoming IP (IIP) in time range (x attribute)	Packet length (PL) in time range (y attribute)	Access	Time range (minutes)
192.168.10.2	81	16134	Normal	0-3
192.168.10.3	2939	405244	Normal	0-3
192.168.10.4	803	118889	Normal	0-3
192.168.10.5	1173	165510	Normal	0-3
192.168.10.6	1074	154472	Normal	0-3
192.168.10.7	1566	207772	Normal	0-3
192.168.10.8	1105	155560	Normal	0-3
192.168.10.9	1963	268497	Normal	0-3
172.10.64.199	3386	1088676	Attack	0-3
172.10.85.151	14323	2432059	Attack	0-3
172.10.201.5	10787	2282970	Attack	0-3
172.10.201.19	7658	1831513	Attack	0-3
172.10.71.29	8899	2525711	Attack	0-3
172.10.71.49	9437	1433478	Attack	0-3

E. Apply Gaussian Naive Bayes method

Average (μ) and Standard deviation (δ) are calculated for every normal access and attack on x and y attributes used (2) and (3).

- Average of incoming IP (μ) normal = 1338
- Standard deviation of incoming IP (δ) normal = 847

- Average of packet length (μ) normal = 186510
- Standard deviation of packet length (δ) normal = 114045
- Average of incoming IP (μ) attack = 9082
- Standard deviation of incoming (δ) attack = 3606
- Average of packet length (μ) attack = 1932401
- Standard deviation of packet length (δ) attack = 582331

Formula (1) is used to calculate the likelihood of Incoming IP address (IIP) normal and attack.

$$P(IIP|normal) = \frac{1}{\delta(normal)\sqrt{2\pi}} e^{-\frac{(x-\mu(normal))^2}{2\delta(normal)^2}}$$

- $P(192.168.10.2 = 81|normal) = \frac{1}{847\sqrt{2\pi}} e^{-\frac{(81-1338)^2}{2.847^2}} = 0,0001566$
- $P(192.168.10.3 = 2939|normal) = 7,892E - 05$
- $P(192.168.10.4 = 803|normal) = 0,0003858$
- $P(192.168.10.5 = 1173|normal) = 0,0004622$
- $P(192.168.10.6 = 1074|normal) = 0,0004487$
- $P(192.168.10.7 = 1566|normal) = 0,0004542$
- $P(192.168.10.8 = 1105|normal) = 0,0004535$
- $P(192.168.10.9 = 1963|normal) = 0,0003587$
- $P(172.10.64.199 = 3386|normal) = 2,532E - 05$
- $P(172.10.85.151 = 14323|normal) = 4,342E - 55$
- $P(172.10.201.5 = 10787|normal) = 4,451E - 31$
- $P(172.10.201.19 = 7658|normal) = 3,83E - 16$
- $P(172.10.71.29 = 8899|normal) = 2,339E - 21$
- $P(172.10.71.49 = 9437|normal) = 6,591E - 24$

$$P(IIP|attack) = \frac{1}{\delta(attack)\sqrt{2\pi}} e^{-\frac{(x-\mu(attack))^2}{2\delta(attack)^2}}$$

- $P(192.168.10.2 = 81|attack) = \frac{1}{3306\sqrt{2\pi}} e^{-\frac{(81-9082)^2}{2.3306^2}} = 4,908E - 06$
- $P(192.168.10.3 = 2939|attack) = 2,592E - 05$
- $P(192.168.10.4 = 803|attack) = 7,93E - 06$
- $P(192.168.10.5 = 1173|attack) = 9,984E - 06$
- $P(192.168.10.6 = 1074|attack) = 9,397E - 06$
- $P(192.168.10.7 = 1566|attack) = 1,261E - 05$
- $P(192.168.10.8 = 1105|attack) = 9,578E - 06$
- $P(192.168.10.9 = 1963|attack) = 1,576E - 05$
- $P(172.10.64.199 = 3386|attack) = 3,178E - 05$
- $P(172.10.85.151 = 14323|attack) = 3,848E - 05$

- $P(172.10.201.5 = 10787|attack) = 9,893E - 05$
- $P(172.10.201.19 = 7658|attack) = 0,0001023$
- $P(172.10.71.29 = 8899|attack) = 0,0001105$
- $P(172.10.71.49 = 9437|attack) = 0,0001101$

Formula (1) also used to calculate the likelihood of Packet Length (PL) normal and attack.

$$P(PL|normal) = \frac{1}{\delta(normal)\sqrt{2\pi}} e^{-\frac{(y-\mu(normal))^2}{2\delta(normal)^2}}$$

- $P(192.168.10.2 = 16134|normal) = \frac{1}{114045\sqrt{2\pi}} e^{-\frac{(16134-186510)^2}{2.114045^2}} = 1,146E - 06$
- $P(192.168.10.3 = 405244|normal) = 5,56E - 07$
- $P(192.168.10.4 = 118889|normal) = 2,934E - 06$
- $P(192.168.10.5 = 165510|normal) = 3,439E - 06$
- $P(192.168.10.6 = 154472|normal) = 3,363E - 06$
- $P(192.168.10.7 = 207772|normal) = 3,438E - 06$
- $P(192.168.10.8 = 155560|normal) = 3,372E - 06$
- $P(192.168.10.9 = 268497|normal) = 2,702E - 06$
- $P(172.10.64.199 = 1088676|normal) = 9,021E - 20$
- $P(172.10.85.151 = 2432059|normal) = 2,272E - 90$
- $P(172.10.201.5 = 2282970|normal) = 1,46E - 79$
- $P(172.10.201.19 = 1831513|normal) = 2,317E - 51$
- $P(172.10.71.29 = 2525711|normal) = 1,541E - 97$
- $P(172.10.71.49 = 1433478|normal) = 3,832E - 32$

$$P(PL|attack) = \frac{1}{\delta(attack)\sqrt{2\pi}} e^{-\frac{(y-\mu(attack))^2}{2\delta(attack)^2}}$$

- $P(192.168.10.2 = 16134|attack) = \frac{1}{582331\sqrt{2\pi}} e^{-\frac{(16134-1932401)^2}{2.582331^2}} = 3,050E - 09$
- $P(192.168.10.3 = 405244|attack) = 2,2E - 08$
- $P(192.168.10.4 = 118889|attack) = 5,367E - 09$
- $P(192.168.10.5 = 165510|attack) = 6,865E - 09$
- $P(192.168.10.6 = 154472|attack) = 6,48E - 09$
- $P(192.168.10.7 = 207772|attack) = 8,534E - 09$
- $P(192.168.10.8 = 155560|attack) = 6,517E - 09$
- $P(192.168.10.9 = 268497|attack) = 1,156E - 08$

- $P(172.10.64.199 = 1088676|attack) = 2,398E - 07$
- $P(172.10.85.151 = 2432059|attack) = 4,741E - 07$
- $P(172.10.201.5 = 2282970|attack) = 5,715E - 07$
- $P(172.10.201.19 = 1831513|attack) = 6,749E - 07$
- $P(172.10.71.29 = 2525711|attack) = 4,077E - 07$
- $P(172.10.71.49 = 1433478|attack) = 4,746E - 07$

Probabilities for the nominal attributes are then calculated based on Table 1.

$$P(normal) = \frac{8}{14} = 0,5714$$

$$P(attack) = \frac{6}{14} = 0,4286$$

$$P(IP\ address\ 192.168.10.2) = \frac{1}{14} = 0,0714$$

Formula (4) is used to calculate $P(normal|IP)$ and $P(attack|IP)$

$$P(normal|IP) = \frac{P(IIP|normal)P(PL|normal)P(normal)}{P(IP)}$$

$$P(attack|IP) = \frac{P(IIP|attack)P(PL|attack)P(attack)}{P(IP)}$$

- $P(normal|192.168.10.2) = \frac{0,0001566 \times 1,146E-06 \times 0,5714}{0,0714} = 1,436E - 09$

- $P(attack|192.168.10.2) = \frac{4,908E-06 \times 3,050E-09 \times 0,4286}{0,0714} = 8,983E - 14$

- $P(normal|192.168.10.3) = 3,51E-10$

- $P(attack|192.168.10.3) = 3,421E-12$

- $P(normal|192.168.10.4) = 9,057E-09$

- $P(attack|192.168.10.4) = 2,554E-13$

- $P(normal|192.168.10.5) = 1,272E-08$

- $P(attack|192.168.10.5) = 4,112E-13$

- $P(normal|192.168.10.6) = 1,207E-08$

- $P(attack|192.168.10.6) = 3,654E-13$

- $P(normal|192.168.10.7) = 1,249E-08$

- $P(attack|192.168.10.7) = 6,454E-13$

- $P(normal|192.168.10.8) = 1,223E-08$

- $P(attack|192.168.10.8) = 3,745E-13$

- $P(normal|192.168.10.9) = 7,753E-09$

- $P(attack|192.168.10.9) = 1,093E-12$

- $P(normal|172.10.64.199) = 1,827E-23$

$P(\text{attack}|172.10.64.199) = 4,572E-11$

- $P(\text{normal}|172.10.85.151) = 7,892E-144$
 $P(\text{attack}|172.10.85.151) = 1,094E-10$
- $P(\text{normal}|172.10.201.5) = 5,198E-109$
 $P(\text{attack}|172.10.201.5) = 3,393E-10$
- $P(\text{normal}|172.10.201.19) = 7,099E-66$
 $P(\text{attack}|172.10.201.19) = 4,144E-10$
- $P(\text{normal}|172.10.71.29) = 2,884E-117$
 $P(\text{attack}|172.10.71.29) = 2,703E-10$
- $P(\text{normal}|172.10.71.49) = 2,02E-54$
 $P(\text{attack}|172.10.71.49) = 3,135E-10$

F. Prediction

Decisions are predicted by comparison $P(\text{normal}|IP)$ and $P(\text{attack}|IP)$. If $P(\text{normal}|IP) > P(\text{attack}|IP)$ then the decision is normal, and if $P(\text{normal}|IP) < P(\text{attack}|IP)$ then the decision is under attack.

- $P(\text{normal}|192.168.10.2) > P(\text{attack}|192.168.10.2)$, then IP address 192.168.10.2 categorized in a normal class.
- $P(\text{normal}|192.168.10.3) > P(\text{attack}|192.168.10.3)$, then IP address 192.168.10.3 categorized in a normal class.
- $P(\text{normal}|192.168.10.4) > P(\text{attack}|192.168.10.4)$, then IP address 192.168.10.4 categorized in a normal class.
- $P(\text{normal}|192.168.10.5) > P(\text{attack}|192.168.10.5)$, then IP address 192.168.10.5 categorized in a normal class.
- $P(\text{normal}|192.168.10.6) > P(\text{attack}|192.168.10.6)$, then IP address 192.168.10.6 categorized in a normal class.
- $P(\text{normal}|192.168.10.7) > P(\text{attack}|192.168.10.7)$, then IP address 192.168.10.7 categorized in a normal class.
- $P(\text{normal}|192.168.10.8) > P(\text{attack}|192.168.10.8)$, then IP address 192.168.10.8 categorized in a normal class.
- $P(\text{normal}|192.168.10.9) > P(\text{attack}|192.168.10.9)$, then IP address 192.168.10.9 categorized in a normal class.
- $P(\text{normal}|172.10.64.199) < P(\text{attack}|172.10.64.199)$, then IP address 172.10.64.199 categorized in attack class.
- $P(\text{normal}|172.10.85.151) < P(\text{attack}|172.10.85.151)$, then IP address 172.10.85.151 categorized in attack class.
- $P(\text{normal}|172.10.201.5) < P(\text{attack}|172.10.201.5)$, then IP address 172.10.201.5 categorized in attack class.
- $P(\text{normal}|172.10.201.19) < P(\text{attack}|172.10.201.19)$, then IP address 172.10.201.19 categorized in attack class.
- $P(\text{normal}|172.10.71.29) < P(\text{attack}|172.10.71.29)$, then IP address 172.10.71.29 categorized in attack class.

- $P(\text{normal}|172.10.71.49) < P(\text{attack}|172.10.71.49)$, then IP address 172.10.71.49 categorized in attack class.

G. Visualization of Classification

Two-dimensional images can be used to display the classification results, so it can detect the level of accuracy. Matlab is the right tool to display the result of the classification.

```

1 - t = 0:pi/10000:2*pi;
2 - x1 = 1338 + 2541*cos(t); % 1xSD=847, 1,5xSD=127
3 - y1 = 186510 + 342135*sin(t); % 1xSD=114045, 1,5xSD=171067
4 - x2 = 9082 + 9015*cos(t); % 1xSD=3606, 1,5xSD=5409
5 - y2 = 1932401 + 1455827.5*sin(t); % 1xSD=582331, 1,5xSD=873496
6 - h2 = plot(x1, y1, 'g', x2, y2, 'r');
7 - set(h2, 'LineWidth', 2)
    
```

Fig. 8. Create set with average (μ) + Standard Deviation (δ) in Matlab.

Fig. 8 shows how to create a set based on average (μ) + standard deviation (δ) in Matlab; x_1, y_1 is the set of normal access (green), whereas x_2, y_2 is the set of attack (red).

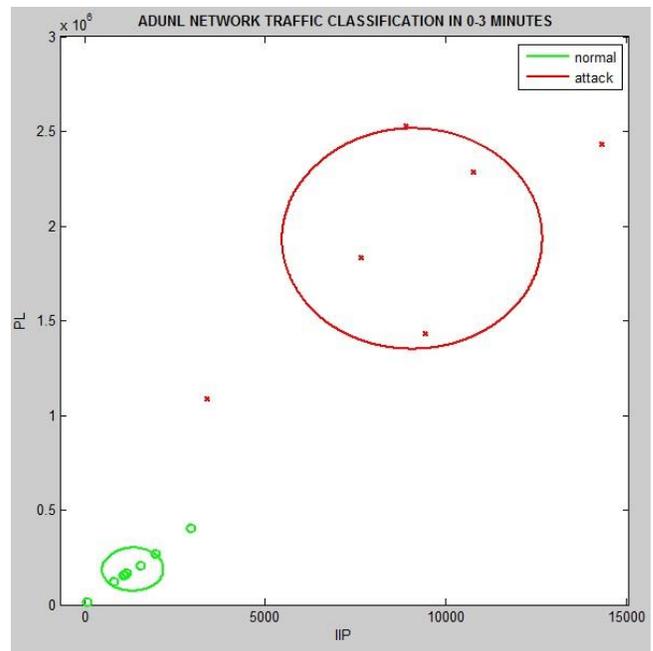


Fig. 9. Network traffic classification with class area $\mu+\delta$.

Fig. 9 shows a visualization of ADUNL network traffic classification in 3 minutes time range with an area of class $\mu+\delta$ using Matlab. The normal class area and the attack with $\mu+\delta$ based on Fig. 9 have not precisely shaded the members of the set. The accuracy obtained using the formula (5) is 57,14%, then searched again the value of δ to get the broad class that can shelter its members.

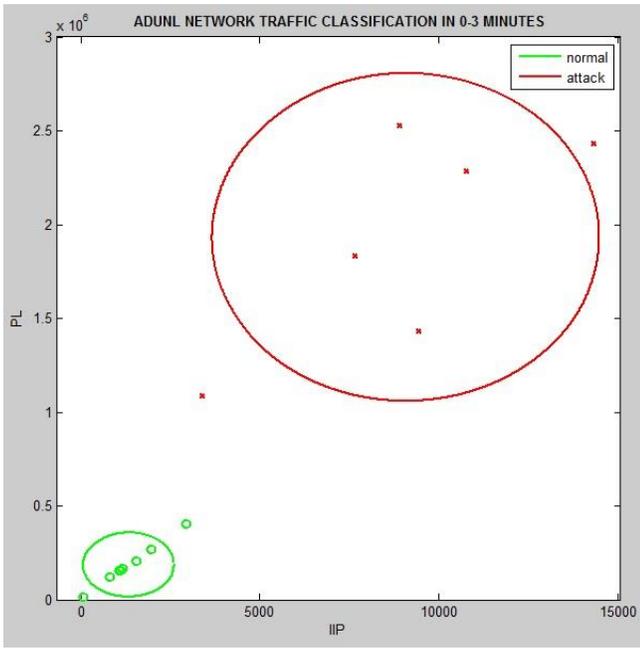


Fig. 10. Network traffic classification with class area $\mu+(1,5\delta)$

The normal class area and the attack with $\mu+(1,5\delta)$ based on Fig. 10 still have not precisely shaded the members of the set. The accuracy obtained using the formula (5) is 71,43%, then searched again the value of δ to get the broad class that can shelter its members.

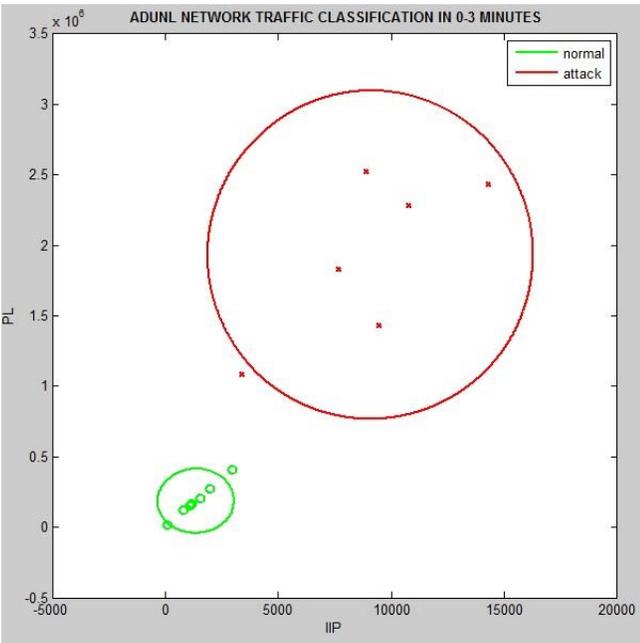


Fig. 11. Network traffic classification with class area $\mu+(2\delta)$.

The normal class area and the attack with $\mu+(2\delta)$ based on Fig. 11 still have not precisely shaded the members of the set. The accuracy obtained using the formula (5) is 78,57%, then

searched again the value of δ to get the broad class that can shelter its members.

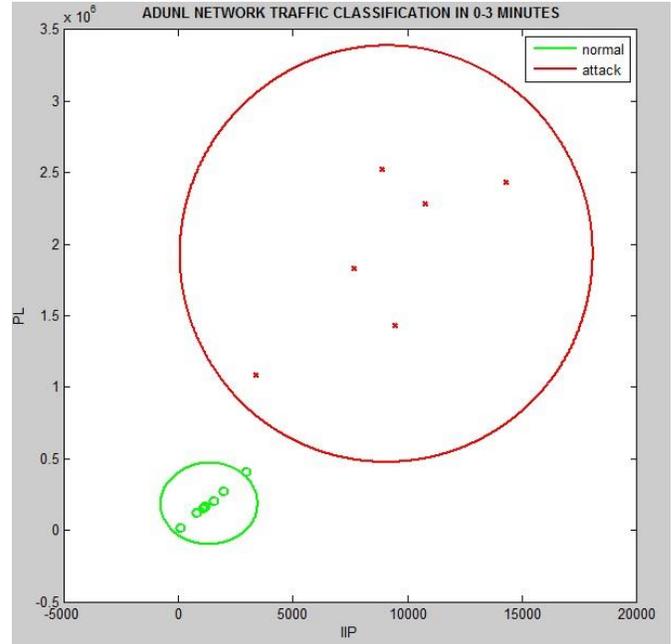


Fig. 12. Network traffic classification with class area $\mu+(2,5\delta)$.

The normal class area with $\mu+(2,5\delta)$ based on Fig. 12 has not precisely overshadowed the set members, while the attack class is right to cover the set members. The accuracy obtained using the formula (5) is 92,86%, then searched again the value of δ from the normal class to obtain the extent of class that can shelter its members.

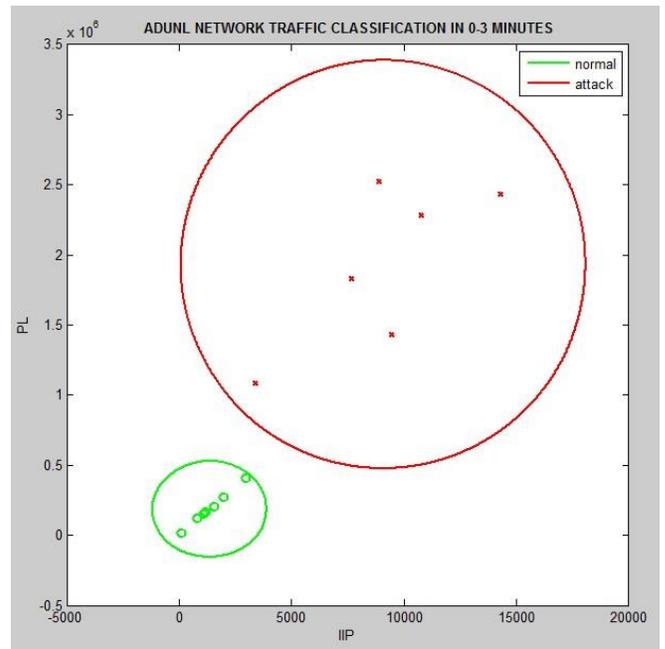


Fig. 13. Network traffic classification with class normal area $\mu+(3\delta)$ and class attack area $\mu+(2,5\delta)$.

TABLE II. CLASSIFICATION WITH NEW STANDARD DEVIATION IN TIME RANGE 0-3 MINUTES

No	IP Address	Incoming IP (IIP) in time range (x attribute)	Packet length (PL) in time range (y attribute)	Access	P(normal IP)	><	P(attack IP)	CLASS
1	192.168.10.2	81	16134	NORMAL	1.145E-09	>	1.859E-11	NORMAL
2	192.168.10.3	2939	405244	NORMAL	9.789E-10	>	3.328E-11	NORMAL
3	192.168.10.4	803	118889	NORMAL	1.405E-09	>	2.197E-11	NORMAL
4	192.168.10.5	1173	165510	NORMAL	1.459E-09	>	2.371E-11	NORMAL
5	192.168.10.6	1074	154472	NORMAL	1.45E-09	>	2.326E-11	NORMAL
6	192.168.10.7	1566	207772	NORMAL	1.456E-09	>	2.548E-11	NORMAL
7	192.168.10.8	1105	155560	NORMAL	1.452E-09	>	2.336E-11	NORMAL
8	192.168.10.9	1963	268497	NORMAL	1.381E-09	>	2.772E-11	NORMAL
9	172.10.64.199	3386	1088676	ATTACK	3.272E-11	<	5.038E-11	ATTACK
10	172.10.85.151	14323	2432059	ATTACK	1.383E-24	<	5.793E-11	ATTACK
11	172.10.201.5	10787	2282970	ATTACK	1.023E-20	<	6.943E-11	ATTACK
12	172.10.201.19	7658	1831513	ATTACK	6.346E-16	<	7.169E-11	ATTACK
13	172.10.71.29	8899	2525711	ATTACK	1.237E-21	<	6.695E-11	ATTACK
14	172.10.71.49	9437	1433478	ATTACK	1.189E-14	<	6.856E-11	ATTACK

The normal class area with $\mu+(3\delta)$ and the attack class area with $\mu+(2,5\delta)$ based Fig. 13 is appropriate to cover the set members. The accuracy obtained using the formula (5) is 100%, then counted once again using the Gaussian Naive Bayes classifier to ensure the correctness of each set member. Average and new standard deviation is:

- Average of incoming IP (μ) normal = 1338
- Standard deviation of incoming IP (3δ) normal = $3 \times 847 = 2541$
- Average of packet length (μ) normal = 186510
- Standard deviation of packet length (3δ) normal = $3 \times 114045 = 342135$
- Average of incoming IP (μ) attack = 9082
- Standard deviation of incoming IP ($2,5\delta$) attack = $2,5 \times 3606 = 9015$
- Average of packet length (μ) attack = 1932401
- Standard deviation of packet length ($2,5\delta$) attack = $2,5 \times 582331 = 1455827,5$.

Table 2 shows the recalculating of Gaussian Naive Bayes classifier using a match standard deviation. The class of the normal and attack set corresponds to the access of each IP address.

The average and match standard deviation are finally used to calculate all new data of network traffic at ADUNL in time-range 3 – 60 minutes using Gaussian Naive Bayes classifier shown in Fig. 14.

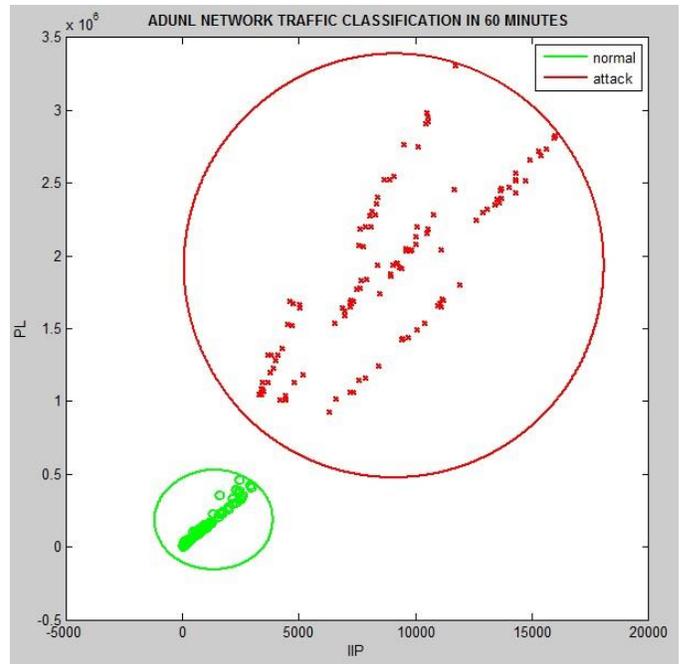


Fig. 14. ADUNL Network Traffic Classification in 60 minutes.

V. CONCLUSION AND FUTURE WORK

Gaussian Naive Bayes classification can be used to process numeric attributes on a computer network service. Numeric attributes such as Incoming IP and Packet Length are the main features to know the access that occurs in a computer network. The average and standard deviation are important for processing data based on Gaussian method, which is also used to visualize in the Matlab. Traffic on a computer network service such as normal access and DDoS attacks can be

grouped according to their class. Classes using the Gaussian Naive Bayes method more specifically cover all of its members based on the average and standard deviation. This method makes it very easy to detect the flow of data packets that are characteristic of DDoS attacks. Furthermore, this paper is expected to process more attributes as well as various parameters to be able to produce DDoS attack detection with better accuracy.

REFERENCE

- [1] M. Tabash and T. Barhoom, "An Approach for Detecting and Preventing DoS Attacks in LAN," vol. 18, no. 6, pp. 265–271, 2014.
- [2] N. Singh, A. Hans, K. Kumar, M. Pal, and S. Birdi, "Comprehensive Study of Various Techniques for Detecting DDoS Attacks in Cloud Environment," *Int. J. Grid Distrib. Comput.*, vol. 8, no. 3, pp. 119–126, 2015.
- [3] G. Oke, G. Loukas, and E. Gelenbe, "Detecting Denial of Service Attacks with Bayesian Classifiers and the Random Neural Network," *IEEE, no. Fuzzy Systems Conference*, 2007.
- [4] B. Nagpal, P. Sharma, N. Chauhan, and A. Panesar, "DDoS Tools : Classification , Analysis and," pp. 2–6.
- [5] Gnanapriya and K. R, "Denial Of Service Attack By Feature Reduction Using Naive Bayes Classification," vol. 4, no. 1, 2016.
- [6] A. S. Tanennbaum, *Computer Networks*, 5th ed. Pearson, 2011.
- [7] S. H. C. Haris, "Anomaly Detection of IP Header Threats," *Int. J. Comput. Sci. Secur. y*, vol. 4, no. 5, pp. 497–504, 2011.
- [8] J. Yang, X. Yu, Z. Xie, and J. Zhang, "A novel virtual sample generation method based on Gaussian distribution," *Knowledge-Based Syst.*, vol. 24, no. 6, pp. 740–748, 2011.
- [9] E. Balkanli, "Supervised Learning to Detect DDoS Attacks," *IEEE Int. Conf. Comput. Commun. Informatics*, 2014.
- [10] J. K. Bains, "Intrusion Detection System with Multi Layer using Bayesian Networks," *Int. J. Comput. Appl.*, vol. 67, no. 5, pp. 1–4, 2013.

A Features-based Comparative Study of the State-of-the-Art Cloud Computing Simulators and Future Directions

Ahmad Waqas, M. Abdul Rehman, Abdul Rehman Gilal, Mohammad Asif Khan, Javed Ahmed, *Zulkefli Muhammed Yusof

Department of Computer Science, Sukkur IBA University, Sukkur, Pakistan

*Department of Computer Science, International Islamic University Malaysia, Kuala Lumpur, Malaysia

Abstract—Cloud computing has emerged during the last decade and turned out to be an essential component for today’s business. Therefore, many solutions are being proposed to optimize and secure the cloud computing environment. To test and validate the proposed solutions before deploying in real cloud infrastructure, a cloud computing simulator is the key requirement. There are several cloud computing simulators that have been used by research community for this purpose. In this paper, we have discussed modern cloud simulators and presented comprehensive comparison based on their features.

Keywords—Cloud computing; simulation; cloud simulator; cloud performance analysis; simulator features

I. INTRODUCTION

Cloud introduced a new way of distributed computing by providing users with on-demand access to resources with minimal efforts and management overheads [1]-[3]. Further, it offers the cost-effective solution for utilizing computing resources and, at the same time, a good attractive business for the enterprises who own computing resources for rent. Cloud computing provides many features including adaptability, accessibility, cost reduction, flexibility, reliability, and scalability [4]. According to National Institute of Standards and Technology (NIST), a cloud ensures the five important features that are “on-demand self-service, broad network access, resource pooling, rapid elasticity and measured service” [5], [6]. Besides this, the goal is to provide on-demand computing services to cloud consumers with the guarantee of reliability, availability and scalability. Cloud provides three basic service models termed as “SaaS (Software as a Service), PaaS (Platform as a Service) and IaaS (Infrastructure as a Service)” [7], [8] and is deployed in four ways known as “Public, Private, Community and Hybrid Clouds”. Fig. 1 summarizes the cloud services and deployment models along with some application domains that may consume the cloud resources [9], [10].

The simulation of newly developed applications, algorithms, and architectural components is vital before its deployment in a real cloud environment. This is essential for analyzing the behavior of new algorithms and for further improvements. Therefore, cloud simulators play an important role and facilitate researchers for rapid evaluation of the efficiency, reliability, and functioning of proposed algorithms on the big heterogeneous cloud infrastructure [11], [12]. Many

cloud simulators are available out of which some are commercial and some are open source. These simulators emphasize on the simulation of particular cloud computing component. For example, some simulators target the simulation of large-scale data centers, some of them simulate the cloud applications and analyze their behavior and some focus on the workload distribution and fault tolerance analysis. Author in [13]-[15] presented the study and comparison of the cloud simulators but they do not impart the simulation emphasis of the presented simulators and the in-depth analysis of their features. This paper aims to review the state-of-the-art cloud simulators in order to explore their features, limitations and research opportunities.

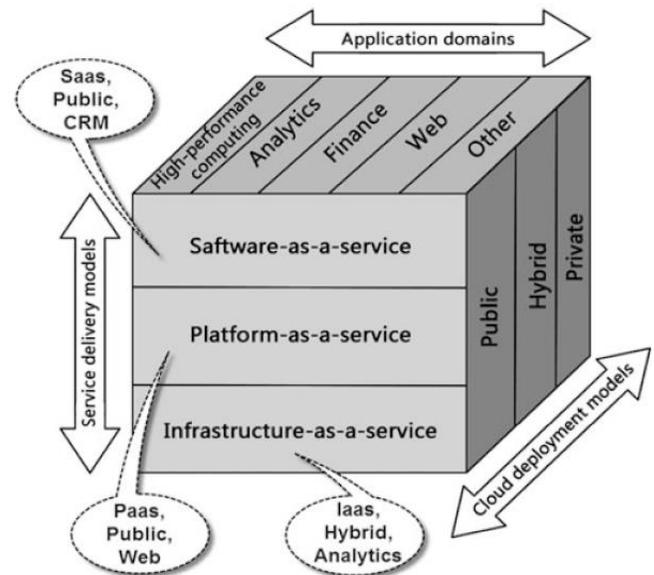


Fig. 1. Cloud computing services and deployment models.

The rest of the paper is structured as follows. Section 2 followed by an introduction, discusses the issues related to performance analysis in cloud computing. Section 3 gives the overview of modern cloud simulators. Section 4 provides the discussion and the detailed analysis and feature based comparison of these simulators. Section 5 concludes the paper and future research directions.

II. PERFORMANCE ANALYSIS AND ISSUES IN CLOUD COMPUTING

The performance analysis of a cloud computing system refers to evaluate it from different perspectives that include the assessment of the SLAs, resources distribution and its complex path generation for service delivery, and secure cloud storage. For evaluating the performance of a cloud computing system, one of the key performance requirements is to assure that it is a SLA- driven system performance [16]. To determine the cloud performance, the major challenges in large-scale cloud computing, are massive scalability, dynamic configuration, and complexity of component interactions. The cloud systems are managed dynamically by SLAs that are settled an agreement between cloud consumer and service provider for service usage and policies. The performance metric for SLA is response time that deals with a response time of requested services to be delivered.

The cloud users connect with cloud and its service components through the internet. The cloud service components may be located at multiple hosts and the hosts may be in federated clouds. The delivery of services to users generates different types of executions and delivery paths that become complex and challenging to determine cloud behavior for performance analysts. An example for such challenge is to identify the service component that causes the main problem when system performance is not satisfying the expectations.

Cloud storage and evaluating its performance is very critical in terms of security, reliability, and availability because it contains the valuable business data. A 10-point execution assessment structure for existing distributed storage framework proposed in [17] is useful for the evaluation and simulation of distributed storage.

III. STATE-OF-THE-ART CLOUD SIMULATORS

Cloud computing evolved in last few years and facilitates with a new way of delivering on-demand computing services. The evolution roots of cloud computing are in Grid Computing and Cluster Computing. Academic and industrial researchers are participating in this field as it is yet evolving. For deploying cloud infrastructure, cloud services, policies related to service delivery and users for example SLA (Service Level Agreement), energy efficient computing, risk management and related policies, load balancing, communication between components and federated clouds, QoS etc. requires cloud simulators for researchers to test, evaluate and improve their ideas and findings before actual implementation in real clouds. There are few cloud simulators available that offer different features to researcher community for simulation of some aspects of clouds. The list below contains the names of modern cloud simulators that are discussed in proceeding sections.

- CloudSim
- CloudAnalyst
- GreenCloud
- iCanCloud

- DCSim
- MDCSim
- NetworkCloudSim
- EMUSIM
- SPECI
- D-Cloud
- eXo Cloud-IDE
- CloudSim

CloudSim, with the latest release of version 3.03 at May 2, 2013, is a simulation and modeling framework for cloud services and infrastructure, developed by the “University of Melbourne’s Cloud Computing and Distributed Systems (CLOUDS) Laboratory”. Its development started with the motivation of providing a simplified, comprehensive and extensible structure to the researchers, developers and cloud analyst for modeling, simulation, experimentations and performance analysis [18]. The main advantages of CloudSim are time effectiveness, flexibility, and applicability. It facilitates the researchers to seamlessly perform experiments and investigate results focusing towards the certain system design issues regardless of low-level infrastructural implementation details of cloud computing. CloudSim eases to model and simulate large-scale data centers, host server’s virtualization with customizable policies for resource provisioning, topologies for data centers, applications that use MPI (Message Passing Interface) and federated clouds. It further offers run-time inclusion of simulation components with stop and resumes feature. Defining user-defined policies enables provisioning of resources and to analyze it [19].

CloudSim has the layered architecture and existing simulation libraries of GridSim and SimJava are exploited for development. As shown in Fig. 2, the first layer is User Code that allows cloud application developers to configure applications, design scenarios for cloud availability and testing, and implement policies for resource provisioning. It contains basic user-level entities such as the number of machines including specifications, required applications, virtual machines, the number of users and policies etc. The most important layer, CloudSim layer is the simulation layer that contains “user interface structures, virtual machine services, cloud services, cloud resources and network implementation details”. CloudSim has the following main components:

- Data center cops the core cloud services. It comprises a set of host entities that are allocated to virtual machines using allocation policy. VMs perform the “low level” processing. Minimum one data center need to be created for simulation.
- Host referred to a computing server that has the processing power, memory, storage and authority for assigning processor cores to virtual machines from a pool of VMs managed by the host.

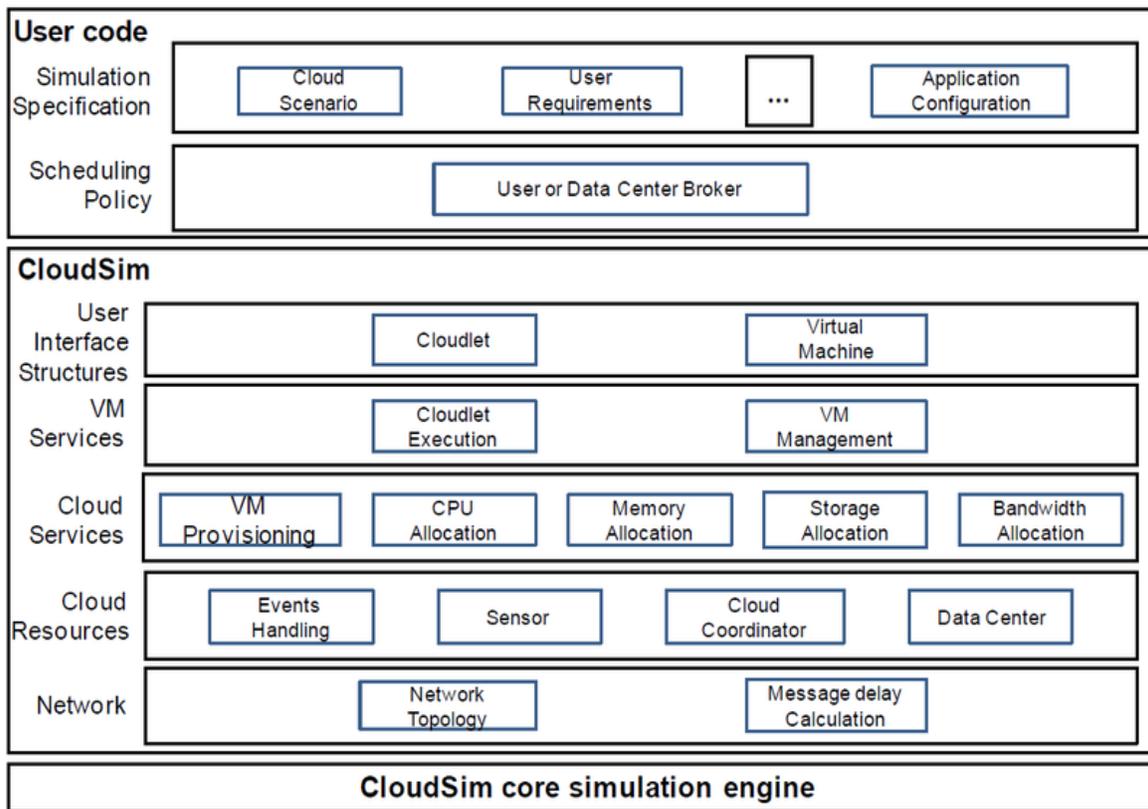


Fig. 2. Architecture of CloudSim [18].

- The provisioning of virtual machine is the assignment of different VMs to different hosts. These are assigned so that host can schedule processing cores to virtual machines. This assignment of VM depends on the application, and the default policy is “First-Come-First-Serve”.
- Datacenter broker plays a role between users and service providers to identify the best provider for user subject to conditions of QoS imposed by the user.
- Cloudlet component of CloudSim signifies the application service whose complexity is modeled in terms of the computational requirements.
- CloudCoordinator is responsible for communicating with other CloudCoordinator, services, and brokers. Furthermore, it observes the internal state of a data center periodically.

A. CloudAnalyst

CloudAnalyst is a visual tool for analyzing the cloud computing environment and applications. It is developed based on CloudSim [19]. The motivation behind CloudAnalyst was the unavailability of tools that can help to estimate the requirements related to the workload on computing servers and user for geographically distributed cloud applications [20]. This requirement and performance analysis is important because cloud contains a distributed infrastructure and applications may run in different geographical locations. This distribution of application effects the performance. CloudAnalyst enables to analyze the performance of extensive

cloud applications based on various deployment setups by simulating them.

The CloudAnalyst is built with the extensions of CloudSim Toolkit as shown in Fig. 3. It extends the GUI package to ease with separation of programming and simulation exercises. The existing CloudSim libraries are used to model the simulation and analysis of applications behavior. CloudAnalyst has the following main components:

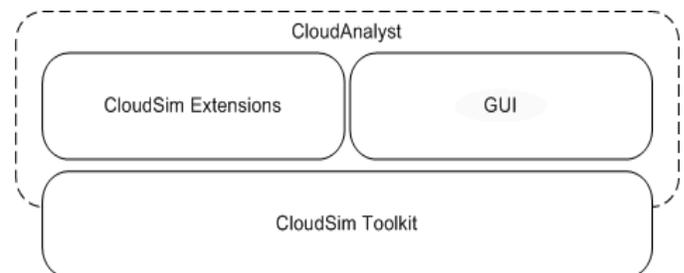


Fig. 3. CloudAnalyst architecture [20].

- GUI Package: The front-end is the graphical user interface to control screen transitions and related functionalities.
- Simulation: This important component enables the development and execution of simulation by retaining the simulation parameters.
- UserBase: It is used to model the users and users’ traffic.

- DataCenterController: This module tackles with the activities related to the data center.
- Internet: This is used to exhibit the Internet and traffic routing.
- InternetCharacteristics: This is used to define the Internet characteristics that are used for simulation including latency, bandwidth, and region etc.
- VmLoadBalancer: Used to implement the load balancing policies for data centers.
- CloudAppServiceBroker: This defines the cloud service broker who is responsible for managing traffic and service delivery between user and service provider.

B. GreenCloud

The absence of a meticulous simulator accessible in the market was the inspiration to develop GreenCloud that enables researchers to watch, communicate and measure cloud executions. GreenCloud is a modern open source cloud computing simulator which has been expounded with regards to the GreenIT and ECO-Cloud projects [21].

GreenCloud is NS2 (Network Simulator) simulator that focuses the packet-level simulation of cloud data centers for energy-aware computing [22]. It enables the simulation of workload distribution along with acquiring the information related to energy utilized by data center components including servers, switches, and links. Furthermore, it simulates the packet level communication patterns. Fig. 4 elaborates the architecture of GreenCloud.

The architecture of GreenCloud is deployed on the basis of three-tier network architecture. Data Center layer is added on top of core network layer which provides an interface for cloud users. The components of Data Center layer are Data Center Characteristics and Task Scheduler. Data Center Characteristics is used to define the data center and Task Scheduler is responsible for schedule and workload distribution. At the “core network, aggregation network and access network layers”, TaskCom Agent are responsible for routing and communication between computing servers, switches and data center components. These layers also contain the L3 and L2 energy model to capture the energy consumed by layer 3, layer 3 and rack switches. For each computing server, there is a scheduler for scheduling computing tasks, an energy model to capture energy consumed by server and server characteristics component to define the server capabilities.

C. iCanCloud

iCanCloud is an adaptable and scalable cloud computing simulation tool intended for modeling and mimicking the large cloud environments both existent and fictional cloud architectures. It was developed by the “Computer Architecture, Communications and Systems (ARCOS) Research Group at Universidad Carlos III de Madrid, Spain” [23]. It provides the valuable information to users about expenses by forecasting the trade-offs between price and performance of specific applications running on certain hardware. iCanCloud can be utilized by a variety of clients, from basic active users to developers of large distributed applications.

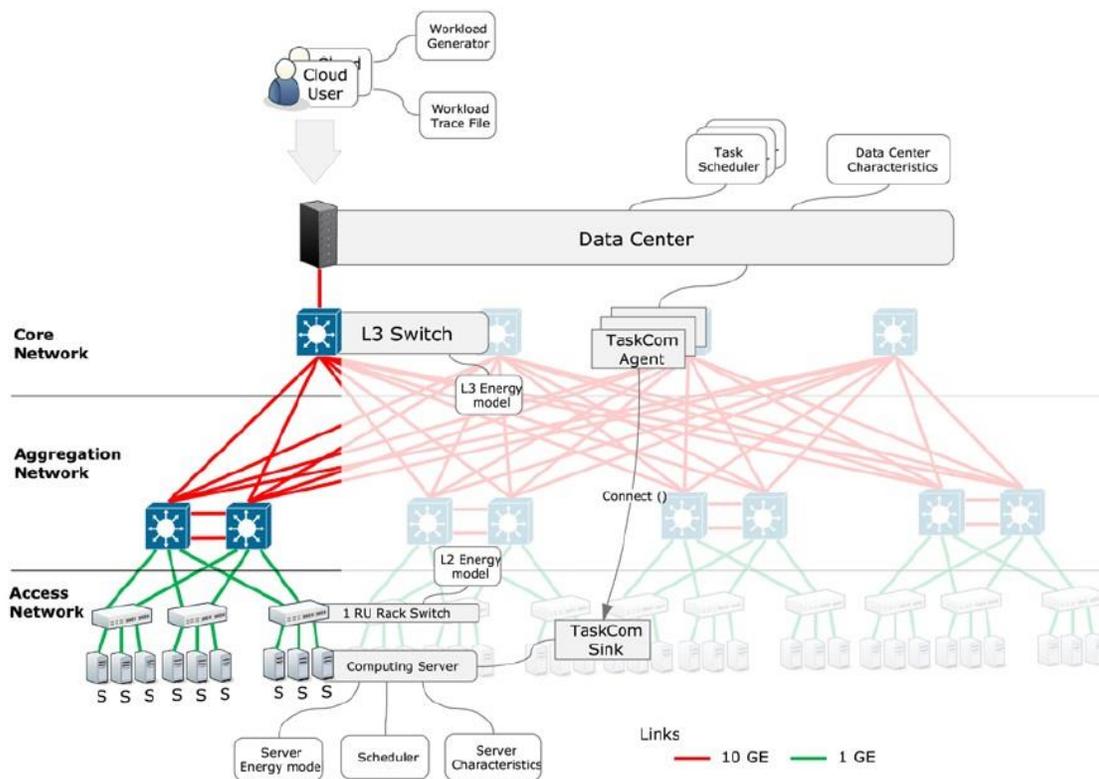


Fig. 4. GreenCloud architecture [22].

iCanCloud has a layered architecture [24] as elaborated in Fig. 5. The lowest layer contains the hardware models such as “CPU, memory, storage and network system”. It also contains the basic systems API that is responsible for providing and interfaces between hardware models and applications. The second layer is VMs repository layer on top of hardware models. It is an essential element for creating cloud system and contains the list of previously created virtual machines and Amazon EC2 VM instances. The application repository layer, on top of VMs repository layer, contains the pre-defined cloud applications that can be customized by users. The cloud hypervisor layer contains three components: 1) job management manages the incoming jobs and VMs instances where the job is performed, 2) brokering policies to define policies related to brokers and 3) cost policies to define the cost of utilized resources. The topmost layer is the cloud system layer that defines the architecture of cloud and deployment of virtual machines.

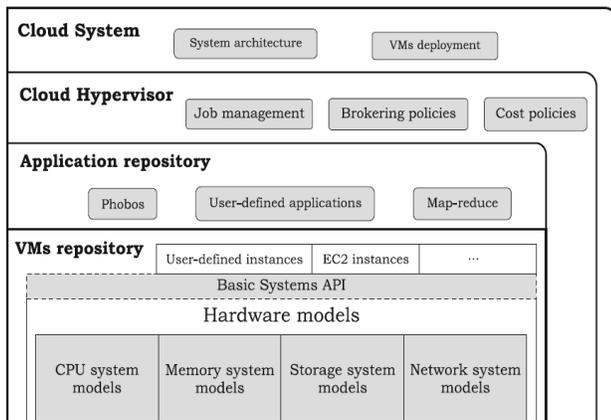


Fig. 5. Architecture of iCanCloud [24].

D. MDCSim

MDCSim is an adaptable and scalable simulation toolkit for comprehensive evaluation of multi-tier data centers [25]. It mainly focuses the simulation of three-tier clustered datacenters. MDCSim is comprehensive because all characteristics of a multi-tier data center are implemented in detail. It is flexible as different characteristics of the data center can be manipulated. For instance, users have provision to change the number of tiers, customize the algorithms for scheduling, and alter the communication mechanisms and interconnect.

CSIM [26] is the underlying platform for MDCSim and the simulation is constituted into three layers called “communication layer, kernel layer, and user-level layer”. Fig. 6 gives an overview of MDCSim platform.

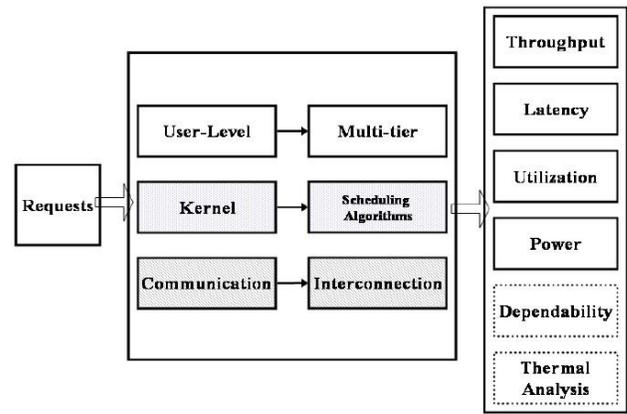


Fig. 6. Overview of multi-tier simulation platform.

E. DCSim

DCSim is a datacenter simulator for assessing management of dynamic virtualized resource [27]. It is an event-driven simulation tool that aims to simulate IaaS offering of a data center to various clients. It focuses on modeling transactional, continuous workloads (such as a web server), but can be extended to model other workloads as well. The foundation of the DCSim development has many motivational reasons including the unavailability of customizable and extensible simulation tool for, modeling multi-tenant data centers, simulating the interactions and dependencies between many VMs, resource management, VM migration and modeling of host power states. Furthermore, it enables the computation and recording of SLA violation, active host, host hours, active host utilization, the number of migrations performed, power consumption, and simulation and algorithm running time.

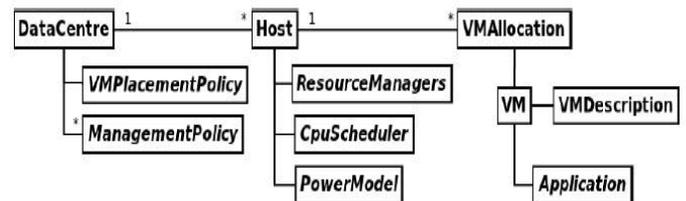


Fig. 7. Architecture of DCSim [27].

Fig. 7 gives an overview of the architecture of DCSim. The key class is termed as DataCenter that contains hosts, virtual machines and various components for management and defining policies. A host is a collection of VMs who is responsible for hosting the task and a data center consists of many interconnected host machines. The responsibilities of the host include VM allocation, resource management, CPU scheduling and power consumption modeling.

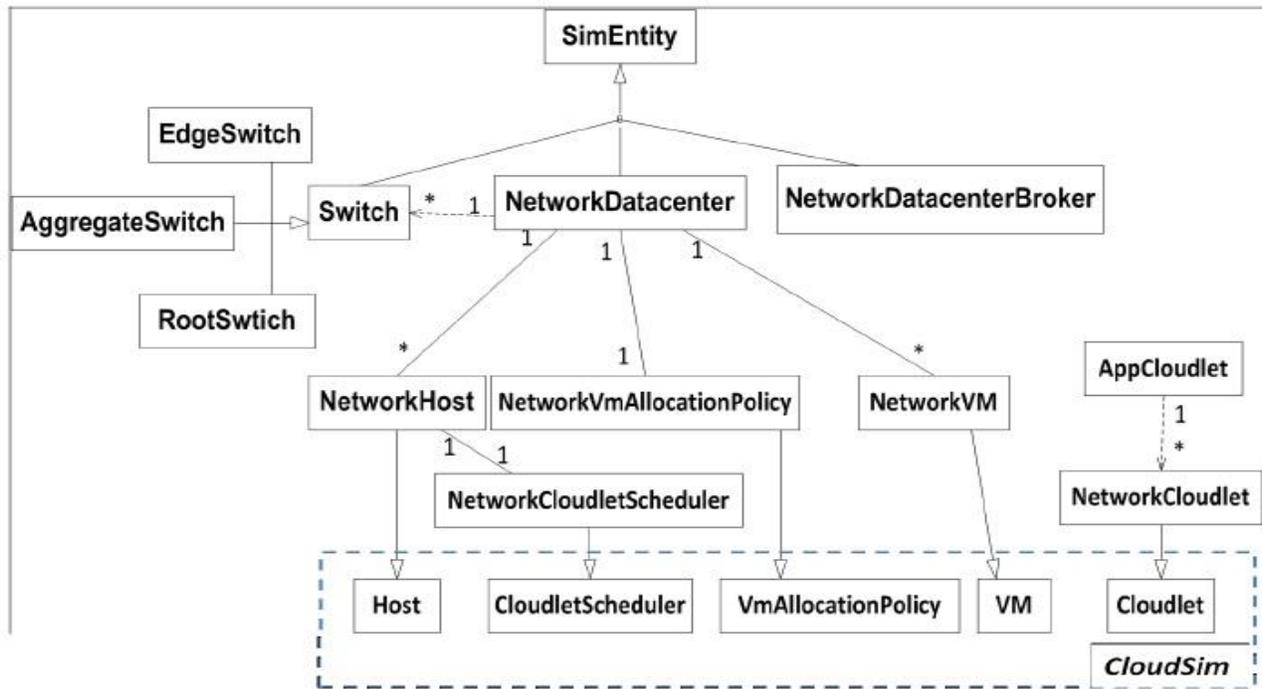


Fig. 8. NetworkCloudSim class diagram [11].

F. NetworkCloudSim

NetworkCloudSim is an extension of CloudSim to model the parallel applications in real networked data centers [11]. Particularly, it addresses the problems and solutions with respect to modeling the internal network and applications of a data center. Furthermore, it supports applications with communicating elements or tasks such as MPI and workflows. In NetworkCloudSim, there are three main actors called Switch, NetworkDatacenter and NetworkDatacenterBroker as shown in Fig. 8. There are two main components of NetworkCloudSim, one of which contains the Switch, NetworkPacket and HostPaket classes for modeling a network topology within the data center. The second component contains the NetworkCloudlet and AppCloudlet classes for application modeling and simulation between different tasks.

G. EMUSIM

EMUSIM is targeted to support the public cloud providers for simulation and analyzing the behavior of cloud applications. It is an incorporated emulation and simulation tool for modeling, assessment, and confirmation of the execution of cloud computing applications [28], [29]. Fig. 9 gives the overview of EMUSIM. It uses Automated Emulation Framework (AEF) [30] for automatically extracting information from the application performance. This extracted information is then used to model the simulation and CloudSim are used for the simulation. It also uses the QAppDeployer [31] – a QoS-aware application deployer – for load generation and automated application deployment to virtual machines. Emulation empowers the execution of the authentic application in a limited environment that models the concrete creation framework, while, simulation permits evaluation of how a system/application acts light of various conditions.

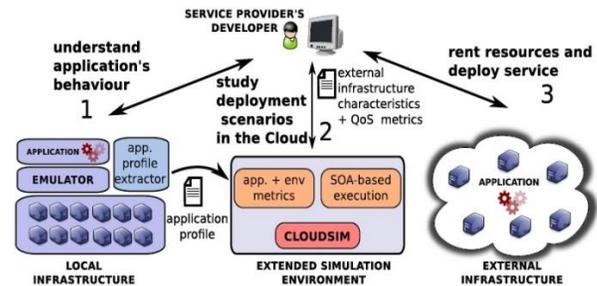


Fig. 9. EMUSIM overview [28].

H. SPECI

SPECI is a simulator for resilient cloud infrastructures that aims to explore the facets of scaling and performance properties of cloud-scale data centers [32]. SPECI is built on two packages, where first specifies the layout and topology of a data center, and second comprises of the elements execution of various experiments and evaluating the results. The data center layout package is a collection of classes for each type of component that is part of the data center, for instance, nodes and network links. This is used to design the layout and topology of the data center for observation and experiments of interest. The experiment component of the SPECI uses SimKit [33] for testing and recording results. The latest public release is named as SEPCI-2 [34] in which the data center layout is structured in a hierarchical fashion.

I. D-Cloud

D-Cloud facilitates with a parallel software testing environment for reliable distributed systems that uses cloud computing technology and VMs with the facility of fault instillation [35]. D-Cloud supports the fault tolerance analysis related to the failures of hardware that happen in the computing

machine. For this, the virtual machine layer of D-Cloud offers the facility of fault injection. Furthermore, it enables to manage computing resources flexibly and automatically, for instance, simulation test can be performed quickly by simultaneous use of resources if available. Moreover, it automates the process of system setup including fault instillation based on test scenario provided by the tester. Additionally, it automates testing phenomena by utilizing the descriptions for system configuration, and test-cases to perform tests on cloud computing systems.

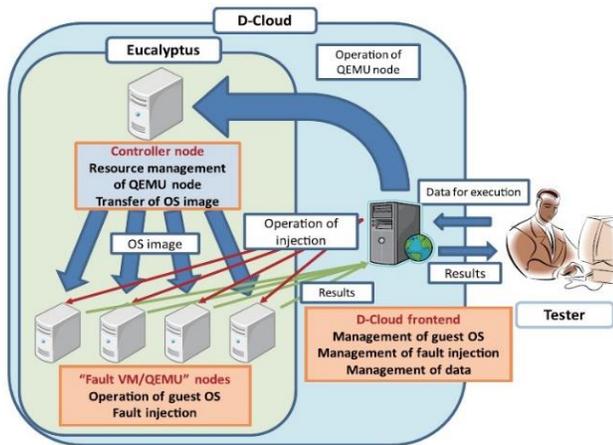


Fig. 10. Architecture of D-Cloud [35].

The design of D-Cloud exploits QEMU [36] for virtualization and Eucalyptus [37], [38] for mimicking cloud environment to manage the computing resources. D-Cloud has three components namely, QEMU nodes, Controller node, and Frontend as depicted in Fig. 10.

J. eXo Cloud-IDE

eXo IDE [39] is a web application that facilitates with an extensive test environment to develop various scripts, content, and services. The advantage of using eXo is that, it does not need installations, it runs in a browser and facilitate for retrieving and operating files online. It is capable of working with remote file system with the help of virtual file system (VFS). Additionally, it provides code editor enriched with popular programming languages such as HTML, XML, Java, PHP, Ruby and JSP for designing and testing applications. Furthermore, it includes many tools for client-side and server-side application development. Also, it facilitates with the deployment of services using “sandbox” technique.

IV. DISCUSSION

There are several simulators available for simulating cloud environment that have been discussed above. These simulators have many common and distinct features along with providing the environment for the different type of simulations. For instance, CloudSim focuses the simulation of large-scale data centers, host server’s virtualization with customizable policies for resource provisioning, topologies for data centers, applications that use MPI, and federated clouds. The CloudAnalyst emphasizes on the cloud applications behavior. Further, the GreenCloud cover the simulation of the data center with respect to energy consumption. The iCanCloud aims to

forecast the trade-off between price and performance. The MDCSim focuses the simulation of multitier data center applications while DCSim covers the workload and VM management. The NetworkCloudSim emphasis on the internal network of data centers. The EMUSIM targets the evaluation of cloud applications behavior. SPECI explores the requirements for scalable data centers. The D-Cloud focuses the analysis of fault-tolerance in hardware and eXo Cloud-IDE provides the environment for designing and testing the cloud applications. Table 1 gives the summary of these cloud simulators with respect to simulation emphasis.

TABLE I. SUMMARY OF CLOUD SIMULATORS

Simulation Tool	Simulation Focus
CloudSim	Large-scale data centers, host server’s virtualization with customizable policies for resource provisioning, topologies for data centers, applications that use MPI and federated clouds.
CloudAnalyst	Analysis of the behavior of extensive cloud applications with a geographical distribution that affects the application performance.
GreenCloud	Workload distribution with gathering information of energy utilized by datacenter components. Packet level communication pattern in cloud data centers for energy-aware computing.
iCanCloud	Forecasting the trade-offs between price and performance of applications running on certain hardware.
MDCSim	Multi-tier clustered datacenters for applications deployment and communication.
DCSim	Modeling of transactional and continuous workloads with VM management in a data center providing an IaaS.
Network CloudSim	Modeling of data centers’ internal network for parallel applications and communication.
EMUSIM	Modeling, evaluation, analysis and validation of performance and behavior of cloud computing applications for the public cloud providers.
SPECI	Exploring the requirements of scaling for elastic cloud infrastructures and performance properties of cloud-scale data centers.
D-Cloud	Fault tolerance analysis related to hardware failures and cloud applications.
eXo Cloud-IDE	Designing and testing applications. Accessing and online manipulation of files.

The simulation results generated with these simulators for the focused domain, applications and algorithms are helpful for

testing and evaluating the cloud environment for the improvement of newly developed applications. Further, facilitation with the specified domain and components for focused simulation enables the accurate evaluation and upgrading of the specified cloud applications and algorithms for the service providers to deploy in real cloud environments.

The above discussed cloud simulators have a different underlying platform and are written in different programming languages. Similarly, some of them are open source while some are available for commercial purposes. Table 2 gives the comparison of different features provided by current cloud simulators.

TABLE II. FEATURE-BASED COMPARISON OF CLOUD SIMULATORS

	CloudSim	Cloud Analyst	GreenCloud	iCanCloud	MDCSim	DCSim	Network CloudSim	EMUSIM	SPECI	D-Cloud	eXo Cloud-IDE
Platform	GridSim	CloudSim	NS2	OMNET	CSIM	-	CloudSim	AEF, CloudSim	SimKit	Eucalyptus/XML	Web-based
Programming Language	Java	Java	C++/Otel	C++	C++/Java	Java	Java	Java	Java	Java	-
GUI Support	No	Yes	Limited	Yes	No	No	No	Limited	No	No	Yes
Availability	Open Source	Open Source	Open Source	Open Source	Commercial	Open Source	Open Source	Open Source	Open Source	Open Source	-
License	LGPL	LGPL	GPL	GNU	-	GPL	LGPL	GPL	-	GNU	-
Federation Support	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	-	Limited	Limited
Full Multi-Cloud Simulation	No	No	No	No	No	No	No	No	No	No	No

V. CONCLUSION AND FUTURE DIRECTIONS

Simulation is extremely important in order to validate cloud applications, algorithms, protocols, and infrastructure. During last decade, many cloud simulators have been developed by the academic and industrial researchers to simulate cloud environments for testing and evaluation. These simulators are built on different architectures and emphasis on simulating different components of cloud including applications and their behavior, data centers, virtualization, workload balancing, and internetworking of data center components. Many of these simulators facilitate with multiple features to test the applications in one or more specified areas based on some evaluation matrix.

Some of the available cloud simulators support cloud federation and semi multi-cloud simulations but up to the best of our knowledge, there is no simulator available that supports the full simulation of multiple clouds with different ownership, administration, and policies at the same time. The full multi-cloud simulation refers to simulating the multiple clouds in a single environment with complete privileges. Such privileges include the inter-connection among multiple clouds with administrative control and with running different administrative, access and security policies at multiple cloud. Further, it includes the simulation of multiple clients of multiple clouds that accesses the resources in real time.

The full multi-cloud simulation is important if we need to design and validate the cross-cloud communication models and intercommunication protocols among cloud networks. Further,

it is of utmost important to test the scenarios of resource sharing and load balancing among connected clouds. In situations, when a cloud may be overloaded with requests of resources and the other cloud with under-utilized resources is available in cloud network, they may facilitate to share the cloud resources for business purposes. To design and validate such a cloud network, full multi-cloud simulation is required that may facilitate with multiple clouds running with different access and security policies.

REFERENCES

- [1] M. Armbrust et al., "A view of cloud computing," *Communications of the ACM*, vol. 53, no. 4, p. 50, 2010.
- [2] A. Waqas, A. W. Mahessar, and N. Mahmood, "TRANSACTIOM MANAGEMENT TECHNIQUES AND PRACTICES IN CURRENT CLOUD COMPUTING ENVIRONMENTS: A SURVEY," vol. 7, no. 1, pp. 41–59, 2015.
- [3] A. Waqas, Z. M. Yusof, A. Shah, and N. Mahmood, "Sharing of Attacks Information across Clouds for Improving Security: A Conceptual Framework," in *IEEE 2014 International Conference on Computer, Communication, and Control Technology (I4CT 2014)*, 2014, pp. 255–260.
- [4] A. Waqas, Z. M. Yusof, and A. Shah, "A security-based survey and classification of cloud architectures, state of art and future directions," in *Proceedings - 2013 International Conference on Advanced Computer Science Applications and Technologies, ACSAT 2013*, 2014, pp. 284–289.
- [5] P. Mell and T. Grance, "The NIST Definition of Cloud Computing (Draft) Recommendations of the National Institute of Standards and Technology," in *NIST Special Publication 800-145 (Draft)*, Computer Security Division, Information Technology Laboratory (ITL), National Institute of Standards and Technology (NIST), U.S. Department of Commerce, Gaithersburg, MD, USA., 2011.

- [6] L. Badger, T. Grance, R. Patt Corner, and J. Voas, "Cloud Computing Synopsis and Recommendations," 2011.
- [7] A. Waqas, Z. M. Yusof, and A. Shah, "Fault tolerant cloud auditing," 5th Int. Conf. Inf. Commun. Technol. Muslim World, ICT4M 2013, pp. 1–5, 2013.
- [8] A. Waqas, Z. M. Yusof, A. Shah, and M. A. Khan, "ReSA: Architecture for Resources Sharing Between Clouds," in Conference on Information Assurance and Cyber Security (CIACS2014), 2014, pp. 23–28.
- [9] B. Furht, "Cloud Computing Fundamentals," in Handbook of Cloud Computing, USA: Springer US, 2010, pp. 3–19.
- [10] A. Waqas, M. A. Rehman, A. R. Gilal, and M. A. Khan, "Cloud Web: A Web-Based Prototype for Simulation of Cross-Cloud Communication Framework (C3F)," J. Bahria Univ. Inf. Commun. Technol., vol. 9, no. 2, 2016.
- [11] S. K. Garg and R. Buyya, "NetworkCloudSim: Modelling Parallel Applications in Cloud Simulations," in 2011 Fourth IEEE International Conference on Utility and Cloud Computing, 2011, pp. 105–113.
- [12] A. Waqas, Z. M. Yusof, A. Shah, Z. Bhatti, and N. Mahmood, "Simulation of Resource Sharing Architecture between Clouds (ReSA) using Java Programming," in 2014 International Conference on Information and Communication Technology for Muslim World (ICT4M), 2014, pp. 5–10.
- [13] X. Bai, M. Li, B. Chen, W.-T. Tsai, and J. Gao, "Cloud testing tools," in Proceedings of 2011 IEEE 6th International Symposium on Service Oriented System (SOSE), 2011, no. Sose, pp. 1–12.
- [14] A. Oujani and R. Jain, "A Survey on Cloud Computing Simulations and Cloud Testing," 2012.
- [15] R. Pandey and S. Gonnade, "Comparative Study of Simulation Tools in Cloud Computing Environment," Int. J. Sci. Eng. Res., vol. 5, no. 5, pp. 110–116, 2014.
- [16] H. Mi, H. Wang, H. Cai, Y. Zhou, M. R. Lyu, and Z. Chen, "P-Tracer: Path-Based Performance Profiling in Cloud Computing Systems," in 2012 IEEE 36th Annual Computer Software and Applications Conference, 2012, pp. 509–514.
- [17] M. F. Ali, A. M. Barnawi, A. Bashar, and I. Technology, "Modeling and Simulation Strategies for Performance Evaluation of Cloud Computing," Int. J. Inf. Stud., vol. 4, no. 3, pp. 148–160, 2012.
- [18] R. N. Calheiros, R. Ranjan, A. Beloglazov, and A. F. De Rose, "CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms," Softw. – Pract. Exp., vol. 41, no. 1, pp. 23–50, 2011.
- [19] A. Beloglazov, R. Ranjan, S. Garg, M. D. de Assuncao, and B. Wickremasinghe, "CloudSim: A Framework For Modeling And Simulation Of Cloud Computing Infrastructures And Services," CloudSim. [Online]. Available: <http://www.cloudbus.org/cloudsim/>. [Accessed: 31-Oct-2013].
- [20] B. Wickremasinghe, R. N. Calheiros, and R. Buyya, "CloudAnalyst: A CloudSim-Based Visual Modeller for Analysing Cloud Computing Environments and Applications," in 24th IEEE International Conference on Advanced Information Networking and Applications, 2010, pp. 446–452.
- [21] GreenCloud, "GreenCloud: Simulating Energy-Efficient Clouds." [Online]. Available: <http://greencloud.gforge.uni.lu/>. [Accessed: 27-Apr-2015].
- [22] D. Kliazovich, P. Bouvry, and S. U. Khan, "GreenCloud: a packet-level simulator of energy-aware cloud computing data centers," J. Supercomput., vol. 62, no. 3, pp. 1263–1283, Nov. 2010.
- [23] ARCOS, "iCanCloud," 2015. [Online]. Available: <http://icancloud.org/Home.html>. [Accessed: 27-Apr-2015].
- [24] A. Núñez, J. L. Vázquez-Poletti, A. C. Caminero, G. G. Castañé, J. Carretero, and I. M. Llorente, "iCanCloud: A Flexible and Scalable Cloud Infrastructure Simulator," J. Grid Comput., vol. 10, no. 1, pp. 185–209, Apr. 2012.
- [25] S.-H. Lim, B. Sharma, G. Nam, E. K. Kim, and C. R. Das, "MDCSim: A multi-tier data center simulation, platform," in 2009 IEEE International Conference on Cluster Computing and Workshops, 2009, pp. 1–9.
- [26] "Mesquite Software." [Online]. Available: <http://www.mesquite.com/>. [Accessed: 28-Apr-2015].
- [27] M. Tighe, G. Keller, M. Bauer, and H. Lutfiyya, "DCSim: A Data Centre Simulation Tool for Evaluating Dynamic Virtualized Resource Management," in 2012 8th international conference on Network and service management (cnsm), and 2012 workshop on systems virtualization management (svm), 2012.
- [28] R. N. Calheiros, M. A. S. Netto, and R. Buyya, "EMUSIM: An Integrated Emulation and Simulation Environment for Modeling , Evaluation , and Validation of Performance of Cloud Computing Applications," Softw. – Pract. Exp., vol. 0, no. 1, pp. 1–18, 2012.
- [29] "EMUSIM: Integrated Emulation And Simulation For Evaluation Of Cloud Computing Applications." [Online]. Available: <http://www.cloudbus.org/cloudsim/emusim/>. [Accessed: 25-Apr-2015].
- [30] R. N. Calheiros, R. Buyya, and C. esar A. F. De Rose, "Building an automated and self-configurable emulation testbed for grid applications," Softw. - Pract. Exp., vol. 40, pp. 405–429, 2010.
- [31] V. Emeakaroha and R. Calheiros, "DeSVi: An architecture for detecting SLA violations in cloud computing infrastructures," in Proceedings of the 2nd International ICST Conference on Cloud Computing (CloudComp'10), 2010.
- [32] I. Sriram, "SPECI, a simulation tool exploring cloud-scale data centres," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 5931 LNCS, pp. 381–392, 2009.
- [33] A. Buss, "Component Based Simulation Modeling with SIMKIT," in Proceedings of the 2002 Winter Simulation Conference, 2002, pp. 243–249.
- [34] I. L. Sriram and D. Cliff, "SPECI-2: An open-source framework for predictive simulation of cloud-scale data-centres," 2011.
- [35] T. Banzai, H. Koizumi, R. Kanbayashi, T. Imada, T. Hanawa, and M. Sato, "D-Cloud: Design of a Software Testing Environment for Reliable Distributed Systems Using Cloud Computing Technology," in 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, 2010, pp. 631–636.
- [36] QEMU, "QEMU, open source processor emulator." [Online]. Available: http://wiki.qemu.org/Main_Page. [Accessed: 29-Apr-2015].
- [37] D. Nurmi et al., "The Eucalyptus Open-source Cloud-computing System," in 9th IEEE/ACM International Symposium on Cluster Computing and the Grid, 2009, pp. 124–131.
- [38] Eucalyptus, "Eucalyptus|Open Source Private Cloud Software." [Online]. Available: <https://www.eucalyptus.com/>. [Accessed: 29-Apr-2015].
- [39] EXo, "eXo Cloud-IDE." [Online]. Available: <https://www.cloud-ide.com/>. [Accessed: 06-Jan-2014].

An Innovative Cognitive Architecture for Humanoid Robot

Muhammad Faheem Mushtaq^{1,2}, Urooj Akram², Adeel Tariq², Irfan Khan^{1,2}, Muhammad Zulqarnain², Umer Iqbal²

¹Department of Computer Science and Information Technology,
The Islamia University of Bahawalpur,
Bahawalpur, Pakistan

²Faculty of Computer Science and Information Technology,
Universiti Tun Hussein Onn Malaysia (UTHM),
Johor, Malaysia

Abstract—Humanoid robot is appearing as most popular research tool and emerging research field. The greatest challenge in the development of robot is cognition, advancement and the understanding in the human like cognition. Humanoid robot requires a self-learning behavior like the humans that is able to get the experience from environment. Based on experience, it can modify their actions, or having conscious intellectual capability to reduce empirical factual knowledge. In this regard, we propose a novel framework called an Innovative Cognitive Architecture for Humanoid Robot (ICAHR) that is capable to develop cognitive through social interaction and autonomous exploration. It combines the modules of active memory, decision processor, and sensor listener that has capability to perform self-learning behavior like human, to make decisions in dynamic environment, and perform more valid and intelligent actions with better precision. The proposed architecture may result in safe, robust, flexible, and reliable machines that can be substitute of human beings in different tasks. The feasibility of new proposed ICAHR design has been examined through real-world case studies.

Keywords—Humanoid robots; cognition; cognitive architecture; self-learning behavior; dynamic environment

I. INTRODUCTION

Cognitive is the mental process of knowing that includes the characteristics such as perception, judgment, reasoning, self-examining, decision making, awareness, imagination, memory and emotion. Cognition process or mental process terms are used for all those actions that human being can perform with their mind. The Greek philosopher highlights the importance of cognitive field which is based on experimental proof and accurate information that are collected through complete observation and careful experimentation [1], [2]. The development of machines like human beings required to combine many useful and desirable features including real human like movement, cognition, and human-friendly behavior and design. Humanoid robotics is a challenging and emergent research area. It brings dominant attention in various models about the description of mind and its related phenomena, and play a vital role in 21st century in the field of robotics. The engagement of humanoid robots in some familiar problems likes the processing of information same as human brain and to deal with the real world to perform more tasks with better precision. The physical appearance and

locomotion has resemblance like the human in their reasoning, actions, and communication about the real environment [3], [4]. The development of such kind of robot need vast range of discipline in their integration and coordination research efforts in several fields such as human machine interaction, control theory, artificial intelligence, computational and psychological computational and perception neuroscience [5]–[10].

Humanoid robots are especially desirable when the human is unable to do unsafe and unhealthy work. It gives assistance to the human because they never get tired and bored on repeated actions, have ability to do tasks in dangerous and uncomfortable situations and perform every task that unlike humans. Humanoid robots are utilized in assisting human activities because of their flexible and friendly appearance [11], [12]. However, it finds that control mechanisms are significantly less advanced. Most of the robots having fixed set of routine codes, their controlling mechanism perform inflexible reactions according to the situation. The learning systems of that robot was not based on psychological findings, so human feel difficulties to interpreting the obtained resulting knowledge. Cognitive architecture gives solution for that problem because their software systems provide support to develop human like cognition and set goals for general intelligence [13].

A cognitive architecture describes the fundamental structure of an intelligent system. It includes those aspects of cognitive systems that share different theoretical assumptions over time. Mostly of these theories belong to the skill acquisition, human memory, problem solving, and reasoning. These architectures based on effective and efficient construction of knowledge-based system that are developed using the programming languages and software environment [14]. Cognitive systems forecast future happening through selecting actions and learn from these events they perform and after that they modify successive expectations. Previously, the cognition was used as the symbol processing unit of the brain that concerns with rational planning and reasoning. Now these early approaches are changed and observe a strong relationship between the cognition, perception and action [15], [16]. The important role of cognitive architecture appears as the central goal of the artificial intelligence and cognitive science that support the similar abilities like human [17].

Humanoid robots can be beneficial from the strong cognitive abilities of these architectures provided. It is called a biological system that takes different decisions and behaves in the environment, learns from them and adapts how to react in new situation and provides solutions based on previous experience. The outstanding feature of the humanoid is the probability to communicate with it, to teach, to interact, and even to demonstrate. They can be used to predict the plans and effects regarding actions in order to accomplish the desire goals [19]. It is a challenging task to develop the connection of human with a robot and examines the effects of how systematically providing knowledge, intentions, and personality to the robot continuously over time. Many of the software are developed for humanoid robots that are control under Open Source paradigm, which shows numerous developers will be capable to modify these robots with powerful cognitive abilities. There are number of humanoid robot in worldwide, some recently developed robots such as NimbRo-OP [18], CASLHR [19], OpenCog [20], [21], and iCub [22], [23] having open source framework that have capabilities to interact with the environment and trying to achieve the required goals. They utilized their already defined pattern, stored memory, receive some information from the surrounding that are not enough for robot to work accurately and efficiently. These architectures have lack of learning and adapting ability, and exhibit limited behaviors. They take actions based on the pre-defined instructions that are given to them but they are unable to learn from previous attempt and update these actions for future use. They show only targeted or specific behaviors toward critical situation. Much of the progress has been done in the development of humanoid robots that are able to perform human-like behavior but still there are some deficiencies in the robots that need to overcome to perform like a real world. Humanoid robots must show behavior like the human being and having perfect intelligence, cognitive process, and perception used multiple behaviors according to the particular situation.

This paper demonstrates the consequences for the inactive method to cognition, phylogenetic pattern, create perception and perform the valid action, the significance of humanoid embodiment, reduce the empirical factual knowledge through the experience gets from the surrounding and their cognitive problems are determined based on user experience. In this regard, this research proposed a new architecture that has powerful cognitive ability for self-learning in humanoid robot with an interactive manner. A new proposed Innovative Cognitive Architecture for Humanoid Robot (ICAHR) having three modules: 1) Active Memory (AM); 2) Decision Processor (DP); 3) Sensor Listener (SL). The proposed architecture has ability to assist continuously for validating and adapting the real-world tasks, as it will demonstrate the attributes are very closer to the human behavior. In addition to this proposed structure, the sensor listener module will enable experimentation and evaluation of sensing capabilities, and emphasize their importance in the development of cognitive abilities. ICAHR architecture can perceive and feel like humans and has ability to learn based on its experience from

previous attempt and modify its actions according to successive rate of its attempt to accomplish goals. The results show that the proposed architecture is suitable for the development of ICAHR architecture and performs much better than the existing humanoid robot architectures. ICAHR architecture shows real human like behavior having perfect cognitive process, perception and intelligence that engage in multiple behaviors according to the specific situation.

The remaining paper is organized as follows: Section II presents the materials and methods that explain the proposed Innovative Cognitive Architecture for Humanoid Robot (ICAHR). Section III includes the results and discussion in which explain the case studies between proposed ICAHR architecture and the conventional robots and Section IV includes the conclusion and future work of this research.

II. MATERIALS AND METHODS

This section explains the proposed Innovative Cognitive Architecture for Humanoid Robot (ICAHR) that is capable to work same as human cognition. Due to this similar cognition, it can be able to do more decision-making power in various situations, thinking, modify and respond actions based on previous experience. This architecture will try to produce intelligence same as the human by evaluating the internal operations and the architecture of the human brain. ICAHR architecture consist of three modules that describe the overall structure of humanoid robots of self-learning process. It contains autonomous philosophical memory like human which means that it can be able to think independently and show different attitudes towards same situation. Autonomous philosophical memory includes the AM and DP modules. Firstly, the AM module contains entity, entity relationship, pattern, action, and goals that have autonomous learning capabilities by actively explore and interact with the environment. AM module is capable to memorize new and old happening as well as memorize the social connections with the other people. Secondly, the DP module includes the perception, feeling processor, execution manager, and validator. This module is used for input-output, it processes the Mata data or information that is to perform more valid action to accomplish goals. Finally, the SL module that contains the active performer and action analyzer. It gets sense from external environment using action analyzer and for processing the observe information send it to the other modules. Based on processed information, the action performer conveys this information to the external environment according to the situations. ICAHR architecture are trying to train the learning behavior of humanoid robot to simulate the human brain activities, understand the planning approach, reasoning, judgment, awareness, representing in traditional artificial intelligence and show dynamic behavior of robot like a human in the surroundings. The self-modification in proposed architecture depends on learning process and structure of the model due to that it is able to alter its model dynamics on the basis of experience, expand and increase its repositories of actions, and as a result modify new situation with more valid set of actions.

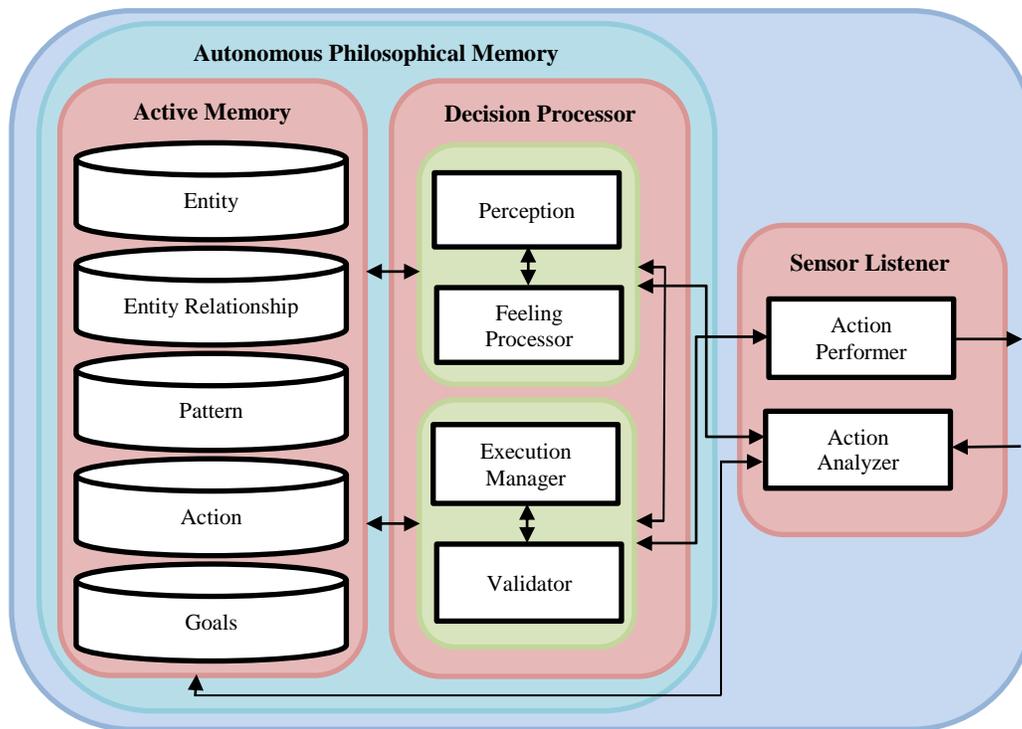


Fig. 1. Innovative Cognitive Architecture for Humanoid Robot (ICAHR).

Each module has individual relationship with the other module as explained in Fig. 1. Action analyzer get inputs from the environment and investigate the set of actions, make perception according to the situation and analyze the previously existence of similar actions in AM module. Based on the feelings and perception, it generates particular actions that fulfil the input commands for the purpose of achieving goals. The action performer is responsible to perform valid set of actions that are taken from DP module. More details of these modules are demonstrated as follows:

A. Active Memory

Active memory is used to memorize the set of actions and used to predict the effects of actions and takes the right decision for actions to accomplish the desire goals. The generation of information and reuse by contents in the system is facilitated through AM that can be stored permanently and retrieved same as the human brain. It is based on advance cognitive processing that is generally and particularly for learning. The main important goal of AM is employed as repositories for knowledge that is related to the interaction with the world and individual regarding series of activities. The repositories can be improve using autonomous learning process when the robot interact with the environment and improve their actions accordingly. Both memory and learning are the important aspects of cognition, in which the development of intelligent skills like deliberative reasoning, planning and self-regulation. ICAHR architecture used the operations of AM that update the memory's content include the entity, entity relationship, pattern, action and goals. When the robot interacts with the real world, it searches the element of perceptual memory that contain the current partial view of relevant surrounding entities. According to the observe

entities, it makes the relationships between these entities and only the useful actions store in memory that exactly match to achieve a desire level of goals. All other entities are only beneficial for the specific moment. The observe entities are recognized like voice, face, object, obstacle and gesture are types of percept that are utilized to process actions, which is basically delivered into the AM. Every repository in AM has individual relationships with the DP module. The information will be store in AM autonomously based on the observation from environment. The knowledge of AM is used to facilitate prediction of future outcomes, imagination regarding new circumstances, and explanation of observed outcomes like the human.

When the robot observes all the real-world objects then these objects consider as entity. Every object that exists in entity repository appears to indicate distinction with the other entities or existence. ICAHR robot will store all these entities or objects in their memory and used to processes and modify the actions. In entity relationship repository, different entities relate to each other through a relationship and it define memory process systematically. The process is illustrated as components that are connected with each other by relationships, it expresses requirement and dependencies between them. Entity relationship is utilized in AM to figure out the entities, their attributes and relationship between the entities.

Pattern repositories are basically the arrangement of entities and their relationships that guide the robot about the angle of the movements and directions of its body parts or as a whole. It also observes the pattern and placement of all components (entities) in its surroundings. Finally, it plans valid scheme of its angle of movements on which follow to

move its parts to achieve a successful action. ICAHR robot can be categorized into naive and learned pattern behavior. When the AM receives first time of stimuli of any entities and their relationships, it has no experience whether the proposed selection of pattern will succeed the goal or not. If the proposed pattern is succeeded, then it will store this successful pattern into memory for future use and if the attempt isn't successful, it will revise its pattern to accomplish the goals. Learned behavior of pattern appears when the AM find similar entities and their relationships again, its learned pattern for execution makes the previous attempt successful. AM will constantly update its patterns based on the experience and memorize only the successful attempt of patterns.

Action repository is generated by the combination of patterns. A step by step movement of body parts that makes the attempt successful will specify an overall action. However, it is necessary that all correct selection of patterns may lead to successful actions. Some cases of the angle and movement of robot lead to correct pattern but it performs action at the wrong target. So, it needs to perform the action again to achieve its right selection of targets. Due to that action can also be categorize into naive and learned behavior. For the first attempt of action, it has no experience regarding which action perform more better result and achieve its targets. Once the successful action performed with specific set of patterns, it will remember these actions in memory for future use. Based on experience, it shows learned behavior to perform action with better precision. Sometimes the learn repository provides considerable assistance to the robot by selecting the appropriate action to attempt. The ICAHR architecture for robot contains strong learning ability to focus on those potentially successful actions to achieve valid task like the human intelligence.

The combination of pattern may not ensure a correct action but the correct set of actions may lead towards the successful accomplishment of the desired goal. The robot will memorize all valid set of patterns and actions that makes the successful goal achievement and based on the achievement, it will update their repository. The proposed architecture emphasizes on the role of goals based on performance guidance. The ICAHR robot is able to compute goals according to its plans and retrieve them when it gets signal from working environment that it can be achievable.

B. Decision processor

Decision processor is used to process the information that is employed to perform action and to take attempt for the movement of robot towards the destination. This module includes the perception, feeling processor, execution manager and validator that are basically used for input and output to perform actions to achieve targets. When the robot get input from action analyzer, it will move to DP for further process. The robot figures out numerous perceptions that is used for the learning of action. Perceptions process continuously engages with different activities. It needs to know about the target and which object is presented before the movement of manipulator towards the engagement. The learning of robot is more manageable, if the robot perception is limited to the specific related task. ICAHR robot can figure out the feeling through facial or postural, vocal and emotional expression show a

strong message that share between the individuals in shared space. Robot feelings begin in sequence or with fluctuating time intensity like the human as naturally occurring emotions that often lead the stimuli to produce them. The important feature of feeling processor is used to coordinate the shifting behavior hierarchies, behavior response, control, promoting physiology support, interacting and establishing position related to other objects. Perception and feeling processor in DP that are mutually work together to perform collaborative task. Sometimes the robot observes the environment than show feelings based on the perception, in some cases the robot first feels something and make perception according to the situation and perform actions. Robot feelings are efficiently and effectively work with the human interaction, it means that the consideration of internal significance of the feeling processor is important for the robot performance. So, that the set of signals are prepare through the perceivable plans for the human-robot association. Perception and feeling processor play an important role as well as attention control, dialogue management, learning, communication elements and complex task planning.

The execution manager is basically responsible to determine the behavior of ICAHR robot throughout the interaction with the real world. This unit is considered as the fundamental component of decision maker that has capabilities to handle the complex environment. Execution manager will formulate a strategy based on the feeling and processor to perform actions for accomplishing goals and this formulated scheme will further move to the validator. Finally, the validator is the superior task to take the final decision and re-confirm the actions. The scheme that is taken from execution manager contains the sequence of actions in which every unit shows the different environmental intentions that contains one correct action for each unit. The correct selection of action will give positive feedback that enhance the reward function and it learns from right selection of commands. If the validator didn't accept the scheme than it will return to the execution manager for improving actions. Due to that the execution manager is more cautious for right selection of actions and strongly motivated to examine multiple times before send for the final decisions. Moreover, validator contains the set of queries regarding actions in the sort of verification question same as the human brain that can be ejaculated in some situation when it commands additional information.

The overview of the working of DP module can be defined as the processing of actions to achieve goals. Perception and feeling processor are basically a two-way process. Either the robot will develop a perception first and will show feelings about the environment and its prospect goals, or it may feel anything first and afterward it will make a perception. The expression of the perception and feeling will be represented in the form of gestures on the facial part of the humanoid robot. Execution manager will formulate a scheme based on the feelings and perception determined to perform an action for the accomplishment of goals. This formulated scheme passes the execution check in validation unit, then it will be implemented, but if the validation check rejects the scheme will be sent back for updating action to the execution

manager. This scheme will only implement if the validation check passes the execution manager's proposed scheme.

C. Sensor Listener

Sensor listener module basically take sense from external world using action analyzer and sent these commands to the AM and DP module for processing. Similarly, it conveys the set of processed commands using action performer that taken from AM and DP to the external environment. Sensors are utilized for extracting appropriate views of the module of engagement. Based on the cognitive approach, this module is responsible to control the incoming and outgoing actions. Action performer takes commands from other modules to perform a set of actions in the form of interaction and communication with the real world and get feedback for further processing. SL module allows the robot to coordinate using the movement of eyes and hands for grasping and manipulating the objects having reasonable size and appearance, sit up, crawling with the arms and legs. It will allow robot to interact and explore with the world, not only through the manipulation of objects but also by locomotion. Moreover, the SL module of the robot will enable the evaluation and testing of the sensing abilities and display their significant role in the development of cognitive abilities.

The overall summary of proposed ICAHR architecture is said to be a complete framework of cognitive architecture for humanoid robots. This research is trying to develop a cognitive architecture that has broadly same as human intelligence as well as powerful self-controllable and sensible motivational system than human. ICAHR robot consist of strong cognitive abilities like the human memory and has better abilities to face multiple challenges when interacting with the surrounding. Robots are engaged with the real world by sharing activities during interaction in which every action of robot is influence with the actions of real world that results in some mutual constructed patterns of shared behavior. It has capabilities of self-learning which is based on feelings and perceptions as well as learning from experience. It is capable to modify actions and performs more valid actions when the action analyzer give command to take attempt. Humanoid robot stores all successful actions in the memory. Also, when similar situation occurs in future than their self-learning behavior motives to performs more effectively and efficiently based on experience. It has thinking capabilities to plan new strategies for right decision of actions to achieve targets. The cognitive abilities of ICAHR robot includes creativity, thinking, awareness, imagination, feelings, perception, decision making, reasoning, desires, ideas and self-examining. The proposed architecture develops knowledge through interaction with the real world and it captures the invariance and consistency that appears from the dynamic self-organization in the aspect of environmental connection. The capability of cognition and artificial intelligence make this architecture same as the human brain.

The comparison of conventional robots and proposed architecture are illustrated in Table 1. The issues with the conventional robots are deficiency of effective information processing having inadequate capabilities to learning from environment and perform limited tasks that already programmed. As compared with the conventional humanoid

robots, proposed architecture are strong abilities of learning and thinking that display versatile behavior. It performs valid actions consisting of cognitive abilities and artificial intelligence. The comparison table explains different situations which revealed that the proposed architecture is more efficient and flexible than the conventional robots.

TABLE I. COMPARISON OF CONVENTIONAL ROBOTS AND ICAHR ARCHITECTURE

Conventional robots	ICAHR architecture
Conventional robots consist on memory that stores only already programmed instructions about the environment.	Proposed ICAHR architecture contain active memory that is capable to store programmed instructions as well as empirical information regarding environment.
They are unable to distinguish between relationships of different entities.	It can be identifying the relationship between different entities and make patterns for performing actions according to the entity relationship.
They show insufficient mobility that already programmed.	It shows efficient mobility based on the situation's experience.
They present limited reaction according to any stimuli.	It is capable to display flexible reactions based on any stimuli.
If same situation happened, they perform similar actions repeatedly.	It has capabilities to show different reactions towards the similar circumstances based on learned behavior.
They are unable to store instructions temporarily.	It can be store instructions temporarily as well as permanently.
They are unable to recognize feelings and emotions.	It is capable to recognize the feelings by using feelings processor in DP module.
They are unable to learn actions from real world.	It can be able to learn actions from environment and store set of actions in the memory.

III. RESULT AND DISCUSSION

To evaluate the cognitive abilities of self-learning of ICAHR robot by using the case studies and apply these case studies into conventional robots and proposed ICAHR architecture is demonstrated on it. Firstly, considering the case studies that robot directs the people on the way to IT lecture room. Secondly, the robot shares its experience to the peoples about assisting people on the way to the target location. More details of these case studies are demonstrated as follows:

A. Case A: Robot Directs the People on the Way to IT Lecture Room

This case study can be divided into smaller activities and each activity is implemented in the most effectively and efficiently manner such as the self-autonomous working of the robot toward the target, step by step navigation of the robot for the direction of IT lecture room, finds the specific area of IT lecture room, reached the lecture room and finish its task, after that it come back to the front of the building. Navigation of robot is done with different sensors when its functioning toward the target. ICAHR robot need assistance with different sensors to properly navigate by floor plan without disturbing

and touching any surrounding objects. Sensors can locate the target position for the robot, so it can assist people toward specified location. Firstly, applying this case study into the conventional robots and then implement it into the proposed ICAHR architecture.

1) *OpenCog*: When the robot gets the commands about assisting the people those who want to go to IT lecture room and help them toward the right location, it store these instruction signals into the memory called atom space. These atom spaces are linked with each other and communicate with mind agent. Mind agent performs some cognitive actions because it has no capability to do the variety of decisions towards the targets and to learn the new behaviors due to its less intelligence. It performed actions through the mind agents that to locate the specified area of room.

2) *iCub*: When the robot gets instructions by sensing that someone want to go to the IT lecture room and request the robot to help him on the way to the right location. iCub robot will sense these signals by agents and save these instructions into the episodic memory. It will verify that these instructions already exist in episodic memory or not, if robot identify that pre-guided instructions then it will executes its effective state and passes these instruction signals to the selection of actions that navigate of robot toward the faculty building and search the already programmed location and reached the specified area of lecture room.

3) *Nimbro*: When this robot gets instructions to direct the people on the way to IT lecture room, it is not capable to perform these actions because it is only develop for basic soccer skills and has specifically programmed to play only football game. If this robot is utilized for the purpose to assist peoples toward lecture room, it needs to be programmed to perform these kinds of actions that how to sense the assistance instructions from the people and perform specified actions.

4) *Proposed CASLHR*: When ICAHR robot gets the commands to assist people on the way to IT lecture room then first of all, it will sense the lecture room area through the action analyzer operation in SL module and send these commands for processing to the AC and DP module. The process of each module is explained in following steps:

Step 1: Active memory module includes some operations that the ICAHR robot is initially containing a set of actions to accomplish ultimate goals and after that successful actions are utilized to store into the memory for future use similar as human intelligence. Entity repository in AM includes as follows: building, corridor, offices, stairs, peoples, lecture rooms, doors, chairs and tables. Similarly, the entity relationship repository contains the following set of relations with the entities: the building has relation with offices and stairs, offices are used by the staff or people, lecture rooms have a door and it contain chairs and tables, other things have direct or indirect relation with building offices, and people, lecture room.

The ICAHR robot will start moving after getting instructions from people. The movement of the robot towards

the building is done through information provided by sensors, find the targeted location after passing from stair and offices. Robot detects the lecture room title on the door by using sensor. The robot performs several actions like starts automatically when getting instructions from people, moving toward the destination by sensing lecture room titles on the door, after assisting the people on targeted location, it will return in front of the building. The ultimate goal of humanoid robot is to find the targeted lecture room and performs necessary actions that try to save time, tiredness, stress and give motivation for studies.

Step 2: This step discusses how DP module work with robot when getting call for assistance from surroundings. Firstly, the feeling processor takes instruction signals from people and automatically starts the robot. It gives the sense to the robot what he will do after listening the voice of peoples and change his behavior for specific action. Perception operation makes sense to the robot that someone call for assistance and the execution manager directs him for the movement toward target. The validator is to re-confirm the execution manager action and humanoid robot started his movement when the validator guides the SL to proceed this action. Secondly, feeling processor detects the lecture room title on the door with the help of sensor. Humanoid robot makes the perception that still targeted goal is not achieved. The execution manager gives the directions to the robot to keep moving until reached at the target. The validator operation step by step verified the execution manager actions and guide the robot to keep moving towards the target when the validator send signal to the SL to proceed these actions. Finally, in feeling processor the sensor continues to detect the lecture room door by door and the robot makes the perception that he reached to the targeted area of the building. The execution manager wants to stop the robot in front of the IT lecture room and makes them decision. The validator operation validates the actions of execution manager and robot asks the people this is your destination when the validator sends instructions to the SL module.

Step 3: The action performer in SL module gets a set of commands from validators that to perform these actions. This module is used to sense the set of commands using action analyser and after processes these commands from AM and DP to get a set of more valid actions for executing specified tasks such as listen the people voice, robot start its movement, continue its movement until it sense to reach the specified location.

B. Case B: Robot Share its Experience with the Peoples

When the robot meets with the real world, communicate with the peoples about how he can gain experience on the way to target location of IT lecture room and explain the hurdles that he faces on the way. Similarly, applying this case study into the conventional robots gives implementation on the proposed ICAHR architecture.

1) *OpenCog*: When the OpenCog robot receives set of instruction from the peoples to sharing his experience how to direct people on the way to IT lecture room, then it stores these set of instructions into the memory. Atom spaces are

utilized as memory in the OpenCog robot which are connected with each other that ensure the set of instructions already exist in atom space or not and after it communicate with the mind agent. The mind agent executes some cognitive process due to less intelligent. The mind agent performs the set of actions that share experiences how he finds the building that contain IT lecture room and reached the specified lecture room.

2) *iCub*: When the *iCub* robot gets commands from the peoples to share his experience that how you accomplish your specified target to reached at the destination lecture room. Robot will sense these set of commands by agent and store into the episodic memory. It will determine that these commands already happened in episodic memory or not. If the robot find out these pre-programmed commands exist then it will run its effective state process. Further, it pass these commands to execute actions that share main steps how effectively identify the building, how to sense the specified lecture room in the building.

3) *NimRo*: When the *NimRo* robot gets instructions from the peoples to share its experience that how he successfully reached the destination. This robot is not capable to perform any actions regarding sensing the target location in the building because it have limited programmed that is only used for the football game. For the purpose of searching specified location in building, this robot needs to be programmed for executing location searching task then it will be able to share any experience.

4) *Proposed ICAHR*: When *ICAHR* robot communicates with the peoples in the real world, then the sensor gets set of commands about how he gain experience on the way to assisting people toward the specified lecture room. Robot

shares its experience that how fast he sense the location of lecture room in the building, how efficiently and effectively reached the specified target in the building, how many obstacle faces toward the IT lecture room. So, the *ICAHR* robot examined the working pattern of human-like intelligence and it shows a clear example of experience based learning. At first attempt, the robot was naive to the set of patterns and actions about the location search in the building, whether the goal is achievable or not. Robot arbitrarily moving on the way to the lecture room and one by one observe all the room titles until the destination location. Robot initially set the patterns and actions were constantly revised to make a successful attempt. After finalizing the perception and feeling of the robot, it finally makes the strategy to continue moving on the particular area. This strategy will be validate by the validator after examining all parameters and then this strategy implemented by action performer in SL module. Validator forward only those commands that ensures successful attempt when executed otherwise it will return to the execution manager for modification in their decision. If the same situation occurs in second time, a learned behavior of robot was observed based on the past experience. Thus, the previous experience of the humanoid robot brings the more valid actions and having conscious intellectual capability to reduce empirical factual knowledge. The robot strategy for attempt was verified by the validator and it tries to improve its actions and behaviors constantly that provide a similarity to a human-like behavior. It senses the people instruction in the real world and responds it accordingly. Many time experiences in particular situation makes the overall output as a successful attempt.

TABLE II. SUMMARY OF CASE STUDIES BETWEEN CONVENTIONAL ROBOTS AND PROPOSED ICAHR ARCHITECTURE

Humanoid Robot	OpenCog	iCub	NimRo	Proposed ICAHR
Information Storage	Atom space	Episodic memory	RBDL library	Active memory
Information Processing	Mind agents	YARP libraries	Nodes	Decision processor
Observe Instructions	DeSTIN scalable deep learning architecture	iCub Interface	ROS software framework	Sensor Listener
Valid and intelligent actions	Low level capability	Low level capability	Low level capability	High level capability
Self-learning ability	Only pre-programmed commands	Only pre-programmed commands	Limited programming for soccer skills	Powerful ability of self-learning
Case A	It will guide the people toward IT lecture room by receiving pre-defined signals	It will direct the people on the way to the IT lecture room by sensing the pre-programmed instructions	Its performance is limited to football games, so it is unable to assist people toward the targeted location	It will assist the people on the way to the IT lecture room using the valid set of actions to perform a particular task. Its self-learning behavior allowed to store all valid actions into the memory for future use
Case B	Atom space are utilized to share some experience about the hurdles that he faces during this journey	Episodic memory is used to share only major steps about experience that how he directs the people in the building without touching obstacle	Limited functionality for football game.	ICAHR robot have ability to explore, analyze and then execute more valid strategy. At first attempt, the robot was naive to share its experience about the target location search in the building by performing valid actions that verify through validator and save these set of actions in memory for future use. Learned behavior appears with more valid actions after analyzing past experience and exploring new situation

The results of above discussed case studies show that the conventional robot take actions based on the pre-programmed commands given to them and display lack of learning behavior. They used stored memory to get some instructions from the environment that is not enough to work accurately and efficiently. The summary of these case studies are mentioned in Table 2. The proposed ICAHR architecture overcome the deficiencies of conventional robots in much better way same as the human.

IV. CONCLUSION AND FUTURE WORK

An approach for a cognitive architecture for an intelligent humanoid robotic system has been presented in this paper. The purpose of this research is to create a humanoid robot that can be capable of understanding the environment by communicate using gestures and simple expressions. This intelligence can achieve through strong manipulation that is based on exploration, imitation and social interaction. In this paper, we are presenting an innovative architecture to automate a robot which contains rich cognitive ability and their behaviour will be very closely related to the human. Prediction is involving in every action and each action will change the perceptual world to some extent. Similarly, every span of perception is basically associated or linked with an action. The ICAHR robot is compared with conventional humanoid robots, proposed architecture significantly improves the learning behaviour of robot from experience. It provides more effective and valid decisions in the real world and its features are almost same as a human. Validation check in DP module that can precisely validate the strategies and implemented actions more specifically for the successful achievement of goals. The key characteristic of ICAHR robot is to develop perception and feeling and to display a learned behaviour after validating actions. AM is capable to store temporary and permanent set of command and most important feature that it constantly learns from its experience and based on experience, it can continually update its actions in the AM. So, it gives rise to an ICAHR architecture is innovative approach to experience based learning. The future work of this research is to develop and programmed the knowledge-based learning system for the ICAHR architecture to make its functionality closer to the human cognition. If we upgrade the databases of AM and DP module then it will increase the learning ability and store information more accurately and systematically that can reduce the chances to unsuccessful attempt. This research can be further analysed to explore on emergent embodied systems which can develop strong cognitive skills because it will perform the actions in the real world and figure out the strong consequences clearly to adopt this stance.

REFERENCES

- [1] D. Vernon, C. Hofsten, and L. Fadiga, A roadmap for cognitive development in Humanoid Robots, Springer-Verlag Berlin Heidelberg, 2011.
- [2] L. Montesano, M. Lopes, A. Bernardino, and J. Santos-Victor, "Learning object affordances: from Sensory-Motor coordination to imitation," IEEE Trans. Robot., vol. 24, no. 1, pp. 15–26, 2008.
- [3] T. Asfour, K. Yokoi, C. G. Lee, and J. Kuffner, "Humanoid Robotics," IEEE Robot. Autom. Mag., pp. 108–118, 2012.
- [4] T. Asfour, J. Schill, H. Peters, C. Klas, B. Jens, C. Sander, S. Schulz, A. Kargov, T. Werner, and V. Bartenbach, "ARMAR-4: A 63 DOF torque controlled Humanoid Robot," IEEE-RAS Int. Conf. Humanoid Robot., pp. 390–396, 2013.
- [5] J. S. Albus and A. J. Barbera, "RCS: A cognitive architecture for intelligent Multi-Agent systems," Annu. Rev. Control, vol. 29, no. 1, pp. 87–99, 2005.
- [6] M. D. Byrne, "Cognitive architectures in HCI: present work and future directions," in Proc. 11th Int. Conf. Hum. Comput. Interact., 2005.
- [7] W. Duch, R. Oentaryo, and M. Pasquier, "Cognitive Architectures: Where do we go from here?," Front. Artif. Intell. Appl., vol. 171, pp. 122–136, 2008.
- [8] M. Frank and N. Franklin, Computational Cognitive Neuroscience, Wiki Book, 1st Edition, 2013.
- [9] C. Green and J. E. Hummel, "Relational perception and cognition: Implications for Cognitive Architecture and the perceptual-cognitive interface," Psychol. Learn. Motiv. - Adv. Res. Theory, vol. 44, pp. 201–226, 2003.
- [10] Y. Wang, "On Cognitive Computing," Int. J. Softw. Sci. Comput. Intell., vol. 1, no. 3, pp. 1–15, 2009.
- [11] R. S. Lakshmi, "Renovating Robots," Int. J. Emerg. Technol. Eng. Res., vol. 3, no. 2, pp. 70–75, 2015.
- [12] J. Pierezan, R. Zanetti, L. Weihmann, and G. Reynoso-meza, "Static force capability optimization of humanoids robots based on modified self-adaptive differential evolution," Comput. Oper. Res., pp. 1–11, 2016.
- [13] K. Kim, D. Choi, J. Y. Lee, J. M. Park, and B. J. You, "Controlling a humanoid robot in home environment with a cognitive architecture," in Proc. IEEE Int. Conf. Robot. Biomimetics, pp. 1754–1759, 2011.
- [14] R. Sun, P. Langley, J. E. Laird, and S. Rogers, "Cognitive architectures: Research issues and challenges," Cogn. Syst. Res., vol. 10, no. 2, pp. 141–160, 2009.
- [15] J. R. Anderson, D. Bothell, M. Byrne, S. Douglass, C. Lebiere, and Y. Qin, "An integrated theory of the mind," Psychol. Rev., vol. 111, no. 4, pp. 1036–1060, 2004.
- [16] P. Langley, "An adaptive architecture for physical agents," in Proc. - IEEE/WIC/ACM Int. Conf. web Intell., pp. 18–25, 2005.
- [17] P. Langley, "Intelligent behavior in humans and machines," Am. Assoc. Artif. Intell., pp. 3–12, 2006.
- [18] M. Schwarz, J. Pastrana, P. Allgeuer, M. Schreiber, S. Schueller, M. Missura, and S. Behnke, "Humanoid teenSize open platform NimRo-OP," Lect. Notes Comput. Sci. Rob. 2013 Robot World Cup XVII, vol. 8371 LNAI, pp. 568–575, 2013.
- [19] Mushtaq, M. F., Khan, D. M., Akram, U., Ullah, S., & Tariq, A., "A cognitive architecture for self learning in Humanoid Robots," Int. J. of Comp. Sci. and Net. Sec., vol. 17, no. 5, pp. 26–36, 2017.
- [20] B. Goertzel, C. Pennachin, and N. Geisweiller, The CogPrime Architecture for Integrative, Embodied AGI, 2014.
- [21] D. Hart and B. Goertzel, "OpenCog: A software framework for integrative artificial general intelligence," Front. Artif. Intell. Appl., vol. 171, no. 1, pp. 468–472, 2008.
- [22] G. Metta, L. Natale, F. Nori, G. Sandini, D. Vernon, L. Fadiga, C. von Hofsten, K. Rosander, M. Lopes, J. Santos-Victor, A. Bernardino, and L. Montesano, "The iCub humanoid robot: An open-systems platform for research in cognitive development," Neural Networks, vol. 23, no. 8–9, pp. 1125–1134, 2010.
- [23] N. G. Tsagarakis, G. Metta, G. Sandini, D. Vernon, R. Beira, F. Becchi, L. Righetti, J. Santos-Victor, a. J. Ijspeert, M. C. Carrozza, and D. G. Caldwell, "iCub: The design and realization of an open humanoid platform for cognitive and neuroscience research," Adv. Robot., vol. 21, no. 10, pp. 1151–1175, 2007.

Shadow Identification in Food Images using Extreme Learning Machine

^{1,2}SALWA KHALID ABDULATEEF

¹Universiti Utara Malaysia, School of Computing, College of Art and Sciences, MALAYSIA

²Tikrit University, College of Computer and Mathematics Sciences, Department of computer science, IRAQ

MASSUDI MAHMUDDIN

Universiti Utara Malaysia, School of Computing,
College of Art and Sciences MALAYSIA

NOR HAZLYNA HARUN

Universiti Utara Malaysia, School of Computing,
College of Art and Sciences MALAYSIA

Abstract—Shadow identification is important for food images. Different applications require an accurate shadow identification or removal. A shadow varies from one image to another based on different factors such as lighting, colors, shape of objects, and their arrangement. This makes shadow identification complex problem and lacking systematic approach. Machine learning has high potential to be used for shadow recognition if it is used to train algorithms on wide number of scenarios. In this article, Extreme Learning Machine (ELM) has been used to identify shadow in shadow mask area. This shadow mask area was determined in the image based on edge detection, and morphological operations. ELM has been compared with Support Vector Machine (SVM) for shadow identification and shown better performance.

Keywords—Extreme machine learning; shadow identification; food images; support vector machine, edge detection; color spaces

I. INTRODUCTION

Shadow is defined as a result light blocking from its sources by an object, which causes the shadow to appear in another object. The shadow appears in the area that does not receive the light directly from its source [1], [2]. Removing shadows from image is effective in simplifying the tasks of image processing and computer vision algorithms. However, removing the shadow has to maintain the information in the original image and the other details except the shadow [3].

In various applications, shadow provides important information regarding the scene, which as a result makes it useful to preserve the shadow information. However, in certain type of applications such as object recognition based on shape and color shadows becomes disturbing factor when they interfere with the object, which makes crucial to develop shadow identification and removal algorithm. This is why shadow identification and removal is considered to be essential [4], [5]. Examples of the applications that require shadow identification and removal are segmentation, scene analysis, tracking, and object detection [6].

Machine learning is a fast emerging field and is receiving high recognition from the researchers as an effective tool in wide range of applications that involve data analysis, learning from high amount of data or features [7], [8]. Computer vision is one important application of machine learning and it has

been applied in image segmentation [9], [10], vision based learning, autonomous car [11]. The appealing aspect of machine learning is providing the machine with the capability of learning from scenarios using implicit mathematical models where there is difficulty in providing explicit mathematical models due to the unbounded number of cases with the high complexity levels. One of the examples of such applications is the phenomenon of shadows. Yet shadows appear because of lighting conditions, there are unlimited scenarios of shadows shape, distribution, variations, and there is no explicit rules to create a boundary between pixels that pertain to shadows and others that belong to the original color of the object [12]. This argument creates a motivation to develop machine-learning model for identifying shadows using plenty of training scenarios. To the best of our knowledge, this article is the first in developing and applying shadow identification algorithm based on extreme leaning machine that is one of the most efficient and recognized learning algorithms in the literature.

The organization of the article is as follows. The next section is introduced related work. In Section III, materials and proposed method are provided. Results and discussion are given in Section IV. Finally, conclusion and future work are provided in Section V.

II. RELATED WORK

The literature contains different shadow identification approaches. From a taxonomy perspective, shadow-identifying approaches are categorized under two classes: model based and property based. The former detects the shadow based on pre-defined geometrical, or illumination shape, while the latter describes the shadow based on features such as geometry, brightness, or color [1]. Some shadow removal work requires no prior knowledge regarding the scene. Levine and Bhattacharyya [13] used boundary information to identify shadow regions in the image based on support vector machine (SVM) and then assign them the color of non-shadow neighbors of the same material. Learning based approaches for removal of shadows have been used in challenging type of applications such as identifying shadows in monochromatic images where the authors [14] have used a Boosted Decision Tree integrated into a Conditional Random Field (CRF) based model to identify the shadow based on extracted features:

shadow variant features, shadow invariant features, and near black features. Regardless the useful results of this work, it focuses on challenges of monochromatic images, which is not used anymore in recent devices. Other researchers have identified shadow based on region. Guo, Dai and Hoiem [5] have predicted relative illumination conditions between segmented regions from their appearance and performed pairwise classification based on such information. This work is based on an assumption that all surfaces that contain shadows are planar and parallel to each other, which is not met in call cases. Also, their shadow detection might fail in case of multiple light sources.

Some researchers have incorporated near infrared features with color information to define the shadow based on the framework of [15]. In [3] the authors have developed a method to remove the shadows from real images based on probability shadow map. The probability shadow map identifies the amount of shadow that is affecting the surface.

Identifying shadows has been tackled in different applications, determining shadows in food type of images is among them. Patel, Jain and Joshi [16] have aimed at locating the fruit on the tree for harvesting purposes. Removing shadow is important to easy the segmentation process. The authors have applied Gaussian filter to remove shadows, which is not effective. Other researchers have applied shadow removing for specific food items such as banana as the work in [17] which shadows were reduced by arranging the distribution of illumination but no full removal of shadow was accomplished.

Unsupervised machine learning has been used also in shadow identification. In [12] the characteristic of the derivative difference of the brightness and light invariant have been used to automatically cluster pixels to generate shadow mask. This approach has been used to solve the shortcoming of the work in [18], however it does not work for vague shadow boundary.

After reviewing the previous approaches, it can be concluded that most of the shadow identification works are based on simplifying assumptions or easy testing scenarios where the shadow is created because of single light source or the object is single or non-connected to nearby objects. Building a more practical shadow identifier requires developing trained model based on effective machine learning.

III. MATERIALS AND PROPOSED METHOD

The experiments of this article have been done on dataset combined of 300 images. In this dataset, images are acquired by smartphone with 8 mega pixels with different lighting conditions; each image contains fork, knife and plate within food.

In order to detect the shadow in the image, a supervised model has been built based on Extreme Learning Machine (ELM). ELM was used for three reasons: firstly it is a supersized learning approach, which enables us to train the method based on examples of shadows. Secondly, this approach is effective in avoidance of local minima. Thirdly, this approach does not suffer from over-fitting similar to other supervised approaches like SVM. In the next subsections, the following points are presented. Firstly, the extracted features

are given in subsection A. Secondly, the training of the model is provided in subsection B. Thirdly, running the classifier of the shadow identifier is presented in C.

A. Feature Extraction

The image was decomposed into set of blocks or windows that have to be classified as shadow or shadow free blocks. In order to do so, set of features has been extracted from each block. Statistical features are extracted such as mean: average or mean value. For a random variable vector a made up of N scalar observations, the mean is defined as:

$$\mu = \frac{1}{N} \sum_{i=1}^N A_i \quad (1)$$

Also, the variance for a random variable vector a made up of N scalar observations, the variance is defined as

$$V = \frac{1}{N-1} \sum_{i=1}^N |A_i - \mu|^2 \quad (2)$$

Where, μ is the mean of A .

Also, the skewness has been extracted which is a measure of the asymmetry of the data around the sample mean. The skewness of a distribution is defined as:

$$s = \frac{E(x - \mu)^3}{\sigma^3} \quad (3)$$

Where μ is the mean of x , σ is the standard deviation of x , and $E(t)$ represents the expected value of the quantity t . Skewness computes a sample version of this population value.

Besides the statistical features, the windows are converted into different color spaces, and also calculate descriptive statistics for them too. Color spaces are: YUV, HSV, I1213, YCbCr, La*b* and Gray scale.

B. Training the ELM Model

Dataset has been built from wide range of images contain different arrangement of food items with some shadows in some parts. Total size of the dataset is 800 records. Half of the records (400 is chosen) from the dataset are used as a training data, and the other half (400) are used records are as testing data. This percentage of decomposition has been chosen because it is the most suitable one for avoiding over-fitting. Brute force approach has been used to find out the best number of neurons in the ELM for better performance. The best testing accuracy was accomplished for sigmoid function and for 50 neurons.

C. Shadow Identifier

In order to identify the shadow in the image, the typical approach is to apply the trained ELM on all the image blocks. However, this will result in identifying shadows inside the item of food. This might lead to removing parts inside the food item and will cause degradability in the shape of the item. The other way is to identify the shadow at the borders between the food items, which is useful to avoid over and under segmentation and to maintain the quality of the shape of the food item. The approach for identifying the shadows at the borders was by building the shadow mask. The shadow mask can be defined as

the area in the image in which the ELM shadow detector will be applied. As shown in the pseudo-code of Fig. 1, the procedure is combined of sequence of steps. Firstly, the Region Of Interest (ROI) is extracted to represent the actual food items in which the algorithm will be applied. Next, edge detection (Prewitt) is applied to extract the borders of the item. This approach has been used because it is gradient based easy to implement comparing with other edge detection. The only problem of Prewitt is its sensitivity to noise, which is not an issue in this stage as the processed image is simply a binary image and is not subject to noise comparing with an original raw image.

These borders are the region in which the morphological operations are applied for performing thickness and closing in order to add nearby area to the border and to maintain continuity. The result represents the mask that is provided to ELM shadow to identify in which windows shadows exist.

```
Input: Binary_ROI,Original Image
Begin
1- Edge_Image=Prewitt(Binary_ROI)
2- Thickened_Edge_Image =Thicken(Filtered_Image)
3- Image_Mask=Close(Thickened_Edge_Image)
4- Shadow_Detected_Image=ELM(Image_Mask,Original Image)
Output: Shadow_Detected_Image
```

Fig. 1. Pseudo code of the proposed method.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

For evaluating the developed approach of shadow detection, different types of food images were used. Fig. 2 shows an image with two food items; all the intermediate steps are shown. The shadow mask was not applied on the whole image, instead it has been applied only on a part of the image where the shadow is expected to appear on the surrounding part of the items as it is shown on Fig. 3. This region is used for testing the shadow algorithm to identify the shadow part and the non-shadow part.

Results of shadow identification were generated for different sizes, colors and number of food items with the developed approach as it is shown in Table 1. The algorithm was able to identify the shadow (red color) and the non-shadow (blue color). It is observed that some parts were identified falsely as shadows. However, this does not degrade the performance because it happens only on the surrounding part of the item and it can only lead to removing small parts in the borders as what it happens in item 3. For further evaluation, ELM classification has been compared with SVM and visual results are shown in Fig. 4. Obviously, ELM was better in identifying shadows than SVM which has failed in some parts in the borders and led to non-smooth results of shadow removing.

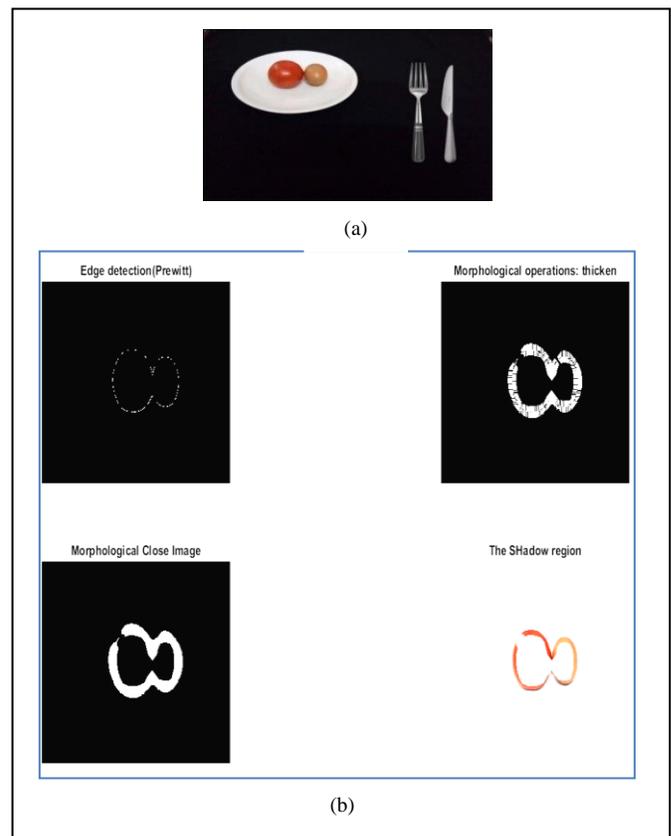


Fig. 2. An example to shows steps of the proposed method (a) Original image (b) Edge detection (Prewitt), thickness and close morphological operation.

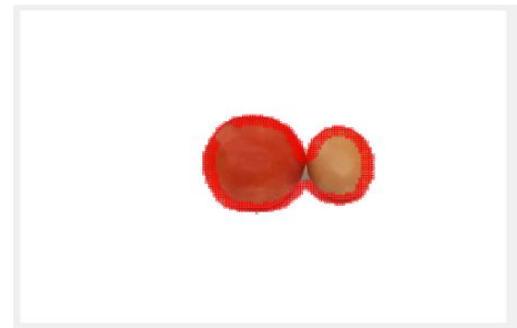


Fig. 3. Overlapping the shadow mask over the original food image.

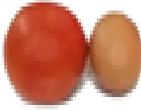
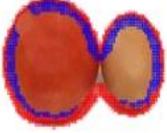
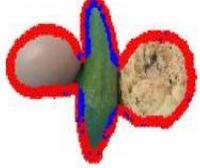
V. CONCLUSION AND FUTURE WORK

Shadow was identified based on combination between shadow mask approach and machine learning approach. For shadow mask edge detection and morphological operations were used while for machine learning ELM was used. The developed approach has been evaluated on different number, shape, and color of food items with different lighting and arrangement of food items. Results have shown good performance. ELM has been compared with SVM and results of ELM have outperformed SVM. Future work is to evaluate this approach as a part of object recognition approaches such as

items identification and calories estimation. In addition to that, the developed approach has to consider adding more features for shadows identification. This might have a role in increasing

the accuracy of shadow identification and decreasing the rate of positive false.

TABLE I. RESULTS IDENTIFYING SHADOWS OF FOOD ITEMS

No.	Original image	Image without background	Mask of shadow	Result of classification
1				
2				
3				
4				

REFERENCES

[1] Salvador, E., Cavallaro, A., & Ebrahimi, T. Cast shadow segmentation using invariant color features. *Computer Vision and Image Understanding*, 95(2), 238–259, 2004.

[2] Vincent, N., & Mathew, S. Shadow Detection: A Review of Various Approaches to Enhance Image Quality. *International Journal of Computer Sciences and Engineering*, 2(4), 49–54, 2014.

[3] Salamati, N., Germain, A., & Siisstrunk, S. Removing shadows from images using color and near-infrared. In *2011 18th IEEE International Conference on Image Processing*, (pp. 1713–1716), 2011.

[4] Zhu, J., Samuel, K. G. G., Masood, S. Z., & Tappen, M. F. Learning to recognize shadows in monochromatic natural images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 223–230), 2010.

[5] Guo, R., Dai, Q., & Hoiem, D. Paired regions for shadow detection and removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12), 2956–2967, 2013.

[6] Blajovici, C., Kiss, P. J., Bonus, Z., & Varga, L. Shadow detection and removal from a single image. *Szeged, Hungary: SSIP, 19th Summer School on Image Processing*, 2011.

[7] Huang, G.-B., Zhu, Q.-Y., & Siew, C.-K. Extreme learning machine: theory and applications. *Neurocomputing*, 70(1), 489–501, 2006.

[8] Singh, L., & Chetty, G. A comparative study of MRI data using various machine learning and pattern recognition algorithms to detect brain abnormalities. In *Proceedings of the Tenth Australasian Data Mining Conference-Volume 134*, (pp. 157–165). Australian Computer Society, Inc, 2012.

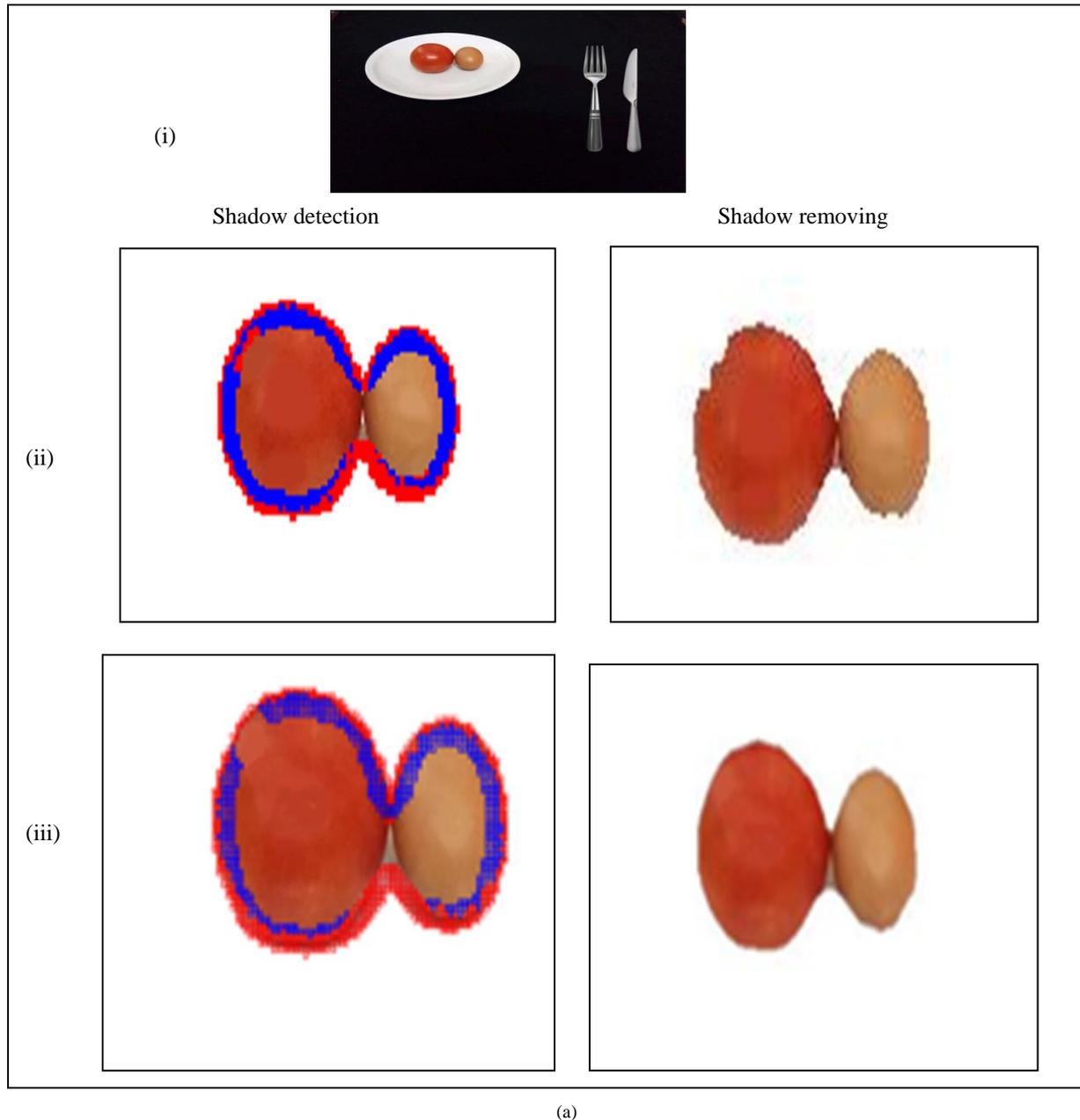
[9] Pan, C., Park, D. S., Yang, Y., & Yoo, H. M. Leukocyte image segmentation by visual attention and extreme learning machine. *Neural Computing and Applications*, 21(6), 1217–1227, 2012.

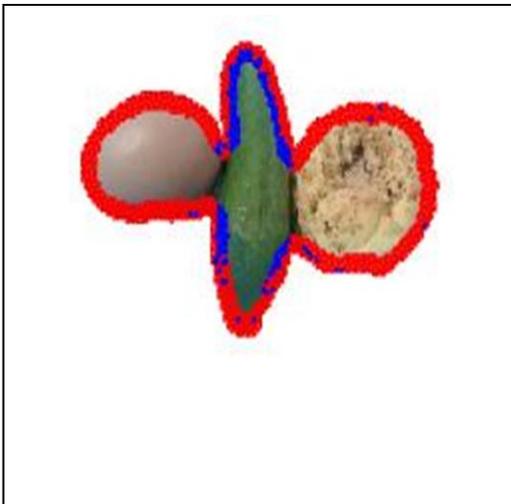
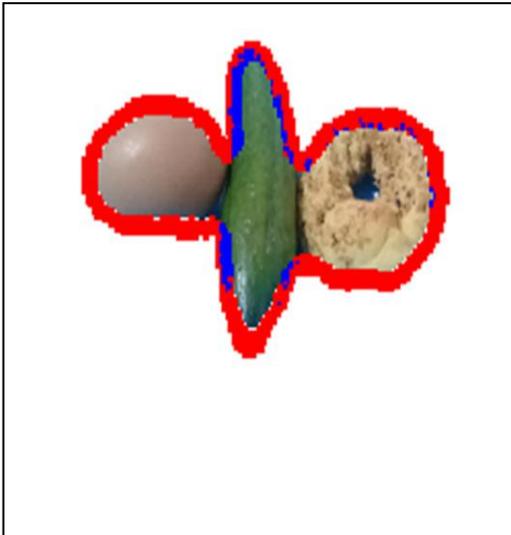
[10] Pratondo, A., Chui, C.-K., & Ong, S.-H. Integrating machine learning with region-based active contour models in medical image segmentation. *Journal of Visual Communication and Image Representation*, 43, 1–9, 2016.

[11] Lee, H., Grosse, R., Ranganath, R., & Ng, A. Y. Unsupervised learning of hierarchical representations with convolutional deep belief networks. *Communications of the ACM*, 54(10), 95–103, 2011.

[12] Shiting, W., & Hong, Z. Clustering-based shadow edge detection in a single color image. In *Proceedings 2013 International Conference on*

- Mechatronic Sciences, Electric Engineering and Computer (MEC), (pp. 1038–1041), 2013.
- [13] [Levine, M. D., & Bhattacharyya, J. Removing shadows. Pattern Recognition Letters, 26(3), 251–265, 2005.
- [14] Xu, M., Zhu, J., Lv, P., Zhou, B., Tappen, M. F., Ji, R., & Member, S. Learning-based Shadow Recognition and Removal from Monochromatic Natural Images, 1–14, 2016.
- [15] Fredembach C, S. Sustrunk S. Automatic and accurate shadow detection from (potentially) a single image using near-infrared information [R]. EPFL Technical Report 165527, 1–12, 2010.
- [16] Patel, H. N., Jain, R. K., & Joshi, M. V. Automatic segmentation and yield measurement of fruit using shape analysis. International Journal of Computer Applications, 45(7), 19–24, 2012.
- [17] Hu, M. H., Dong, Q. L., Liu, B. L., & Malakar, P. K. The potential of double k-means clustering for banana image segmentation. Journal of Food Process Engineering, 37(1), 10–18, 2014.
- [18] Finlayson, G. D., Hordley, S. D., Lu, C., & Drew, M. S. On the removal of shadows from images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(1), 59–68, 2006.





(b)

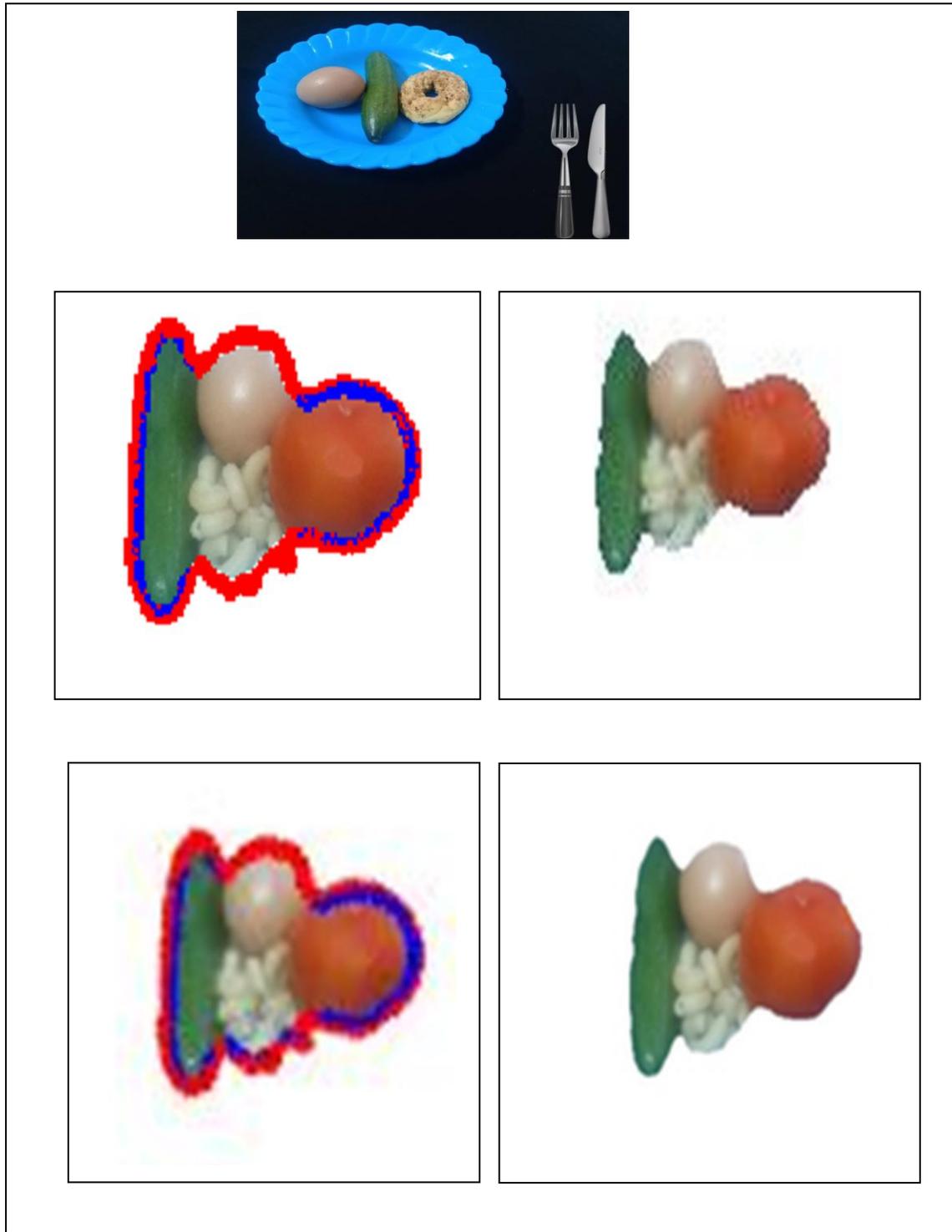


Fig. 4. (a-b) Comparison detection and removing shadow between proposed (i) raw image (ii) detection shadow by SVM and (iii) detection shadow by ELM.

PCA based Optimization using Conjugate Gradient Descent Algorithm

Subhas A. Meti

Department of Electronics and Communication Engineering,
Research Scholar, VTU Regional Research Center,
Belgaum, India

V.G. Sangam

Department of Electronics and Instrumentation,
Dayanand Sagar College of Engineering,
Bangalore, India

Abstract—The energy dissipation in Wireless Body Area Network (WBAN) systems is the biggest concern as it proportionally affects the system longevity. This energy dissipation in the WBAN system mainly takes place due to the signal interference from other networks causing reduction on the dimensionality. The data prediction in WBAN is also a considerable concern corresponding to misinterpretations and faults in the signals. In this paper a novel combination of Principle Component Analysis (PCA) pre-processing along with optimization using the conjugate gradient descent algorithm is proposed. Experimental observations show an improvement in the mean square error and the regression based correlation coefficient when compared to other standard techniques.

Keywords—Associative neural network (AANN); conjugate gradient descent; Non-Linear Principle Component Analysis (NLPCA); Principle Component Analysis (PCA); Wireless Body Area Network (WBAN)

I. INTRODUCTION

The effective improvement in the wireless communication area corresponding to the wireless sensor network (WSN) providing the wide range applications in different areas like military, medical, etc. A kind of WSN is named as Wireless Body Area Network (WBAN) helps to connect the different medical sensor within and outside the body. These WBANs offers the significant mobility for the patients by portable monitoring gadgets. The monitoring ability of the WBAN is area independent and can access the data network to transfer the patient's data. Thus, WBAN framework likewise could get to the information systems (e.g. 3G, 4G) to transfer the patients data. The prime concern of WBAN framework is the productivity regarding energy which demonstrates the system lifetime. The energy in the WBAN could be influenced by numerous variables in light of the area of the observing gadgets which produces clamor/obstruction in the sign. The conventional rarities created from other similar gadgets could be because of variables, for example, state of the checking gadgets, interference from other medicinal sensors, and so on. The conventional methods delivered from inside restorative gadgets incorporate impedence of different signs created because of inadvertent physiological criteria. Subsequently, the restorative gadgets of WBAN system produces signal where antiquities may exist. This kind of interference misjudges the sign that injects the errors and along these lines prompting inappropriate signal forecast. The use of neural networks for WBAN systems helps to enhance the system efficiency, signal

prediction and artifact reduction. In order to describe the problem of energy dissipation in WBAN, An enhancement was done on the WBAN framework in view of the arrangement of the facilitator utilizing neural system strategies [1]. A learning algorithm was used with Kohonen neural system (KNN) to analyze and classify the biomedical signals in the WBAN framework [2]. By utilizing learning based techniques as a part of the neural systems the general energy utilization was lessened to 90%. One of the issues tended to in the WBAN systems is the planning of numerous WBAN in a specific region. The work of [3] considered the same issue of non-linearity to achieve the system high throughput. If a WBAN system exists in a system of multiple WBAN then dimensionality issue may take place, due to which the communication performance may vary because of channel interface among the WBAN systems [4]. Thus, there is a need of a method which can reduce the dimensions in multiple WBAN systems. The analysis and fault detection of biomedical sensors can be done through the modular neural network consisting of associative neural network (AANN). The significant feature of AANN is that it interprets the obtained outcomes and it can be analyzed through data spaces correlation in space modes, which adversely helps in improvement of prediction capability in WBAN system.

This paper is planned as per the sections, where Section II represents the existing research work highlighting the advantage of learning based AANN algorithm. The Section III explains the problem of interest while the next Section IV explains the general modules used in the proposed system. The Section V is subjected to describe the research methodology of the proposed model. The Section VI illustrates the analysis of the outcomes of the system. Finally the Section VII briefs about the conclusion of the proposed system.

II. RELATED WORK

In past various learning based mechanisms were introduced for different application needs. The AANN is a method which falls under the associated neural network (ASNN). This section discussed few existing researches pertaining to ASNN.

The work of Guo-Jian et al. [5] expressed a self-restoration mechanism for the intelligent sensors that implements the AANN to monitor the online insulator contamination status by performing learning. The outcomes of the study suggested that

the use of the AANN based learning based mechanism can locate two faulty transducers synchronously and also the method was able to estimate and recover the drift failures in <5 seconds.

In combined work of Gupta and Singh [6] a bidirectional associative memory (BAM) NN is used to develop a string based NN system for recognition. The study lags with the storage and retrieval of pattern issue as per the BAM efficiency is concerned.

The method of Ang et al. [7] uses a Spiking neural network method to interconnect the delays through FPGA implementation. This study adopted the AANN based memory structure. The outcomes suggests that the method was able to perform the data recalling and learning up to four input pattern by implementing the temporal coding. The obtained neuron response was reduced about 1 to 2 ms by reconfiguring the 60 ns pulse width.

The collaborative work of Lemma and Hasim [8] explained the fault recognition using gas turbine recognition system by introducing a combination model of both AANN and wavelet transformation. The outcomes of the study found bias of ~10% with average detection rate of 95%.

The work of Ravi et al. [9] considered an ANN based banking application where training and classification methods are used for bankruptcy prediction datasets. The implemented AANA method in this study was an optimized method and is named as “Bio-inspired swarm intelligent” model with ASNN. The component analysis was done to reduce the data dimensionality. This [9] study has acquired the higher accuracy than the previous methods.

The study introduced by Gerimella et al. [10] gives the Regularized AANN model based “speaker verification model”. Here the noise generates due to the outliers are considered for regularized AANN model and the parameters of this models are used for Speaker verification model which implements the Linear Discriminate Analysis (LDA). The final outcomes suggested that the EER is improved.

The significant work of Chakraborty [11] introduces a Cardiac Arrhythmia based classification model by using the AANN method. The classifier used in this extracts the non-required features and the study outcome with the high accuracy of >97% and relative gain of >90%.

In Krstulovic et al. [12] a local AANN based power estimation model was introduced. In this Kirchoff’s law is considered as solution for topology constraints and helps to form the specific pattern. This study suggests that under topology constraints the AANN based system is more useful.

The work of Zin et al. [13] gives a multidimensional data visualization and compression system by considering ANN to interpret the multidimensional data. The final outcomes of [13] research gives that the multidimensional data can only be compressed to few non-linear principle components.

In Ito et al. [14] have described a AANN based face recognition system to generate a learning scheme for various components representing facial variation of expressions. The study found accuracy of 77%.

The work addressed in Wang and Yanying [15] states the biometric system for appearance based face recognition. In this, the AANN is considered over one person. The research outcome with the increased recognition rate under partial occlusion and nose of image.

A work addressed in Zhang and Zhou [16] represents a face recognition system for video data by using AANN, in which two facial images were compared for recognition. The accuracy of 90% was obtained under temporal aspects of face image during its recognition.

In the next section, the identified problem is illustrated.

III. PROBLEM IDENTIFICATION

This paper addresses multiple issues in the domain of WBAN. The energy dissipation in the WBAN exhibits factors like dimensionality, signal prediction, cross interference, computational time, etc. All these factors contribute to the energy dissipation which in turn leads to reduced network lifetime. Some of the issues addressed in this paper are as follows:

- *Reduction of dimensionality:* The interference of multiple WBAN in some areas. The dimensionality issue signifies the pervasive in this scenario, there is a necessity of dimensionality mapping (from higher dimensional to lower dimensional subspace. In such case the data may be of non-linear and hence data decomposition is need to be performed or dimensionality reduction.
- *Gradient Descent in node training:* For nonlinear data to reduce the dimensionality the Multi Perception Training (MLP) can be implemented. The MLP overhauls just the weights and totally overlooks the status of the inputs. This makes the framework to consider additionally preparing tests which makes the framework redundant. Thus there is a prerequisite to consider upgrading weights alongside its inputs. For this reason a Non-Linear Principle Component Analysis (NLPCA) is performed. In the NLPCA both the weights and the inputs are overhauled unlike other MLP techniques.
- *Machine learning application to WBAN:* The mapping methods implemented for feature extraction is a preparatory step with regards to machine learning. The following are the important stage in the proposed recognition system which is acknowledged in the mapped data. This requires the utilization of pattern recognition calculations. The example pattern recognition techniques ranges from unsupervised learning mechanisms i.e., Multi-linear Principal Component Analysis (MPCA) to supervised learning (regression based) methods, and for linear regression, Gaussian process regression, neural networks and Deep learning techniques.
- *Non-Linear Optimization in WBAN:* The problem associated with non-linear optimization which is difficult to provide optimal solution given the objective function as to maximize throughput. The work in Li et

al. [1] details this problem by introducing a system called “Fairness-based Throughput Maximization Heuristic (FTMH)” which gives “suboptimal solution” with reduced complexity. Considering the WBAN system of linear equations, enhancement could likewise be performed using techniques, like Conjugate Gradient Algorithms.

IV. PROPOSED SYSTEM

This gives the modules used in the proposed system to tackle the energy dissipation issue. Fig. 1 represents the block diagram of proposed system.

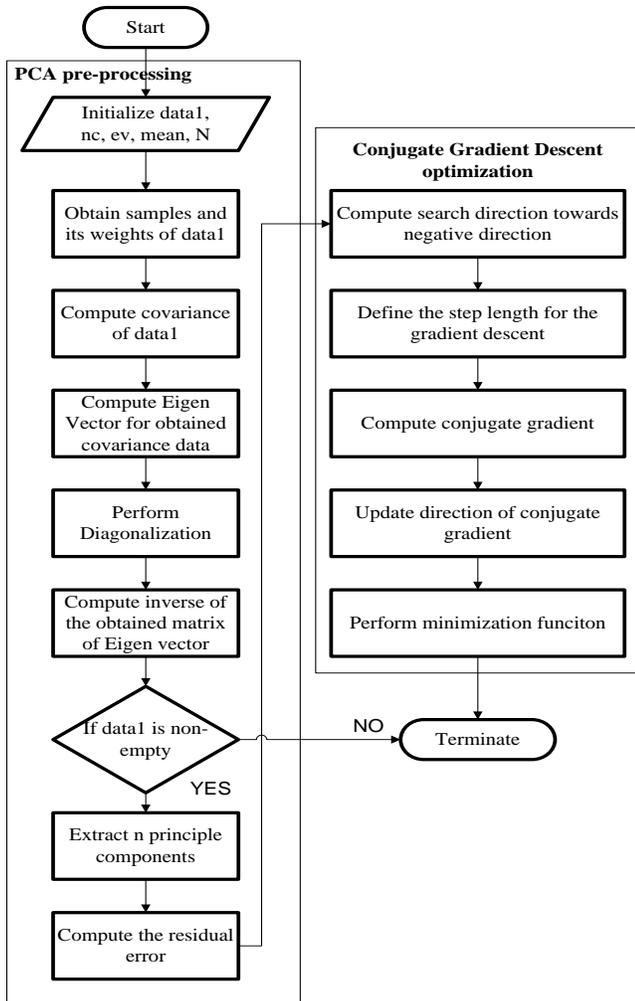


Fig. 1. Proposed system.

- The system contains the preprocessing stage which performs the PCA extraction. In the system, the component analysis (nonlinear principle) method is used for dimensionality reduction.
- The next stage consists of optimization method based on conjugate gradient. In this the supervised (regression) learning mechanism of ANN implemented. Also a modular (based on associative memory) learning mechanism of ANN implemented.

- For training phase AANN method is used which automatically generates the desired training samples for prediction.

V. RESEARCH METHODOLOGY

A WBAN system is mainly composed of more medical devices interconnection which forms a network. Some of the challenges associated with WBAN system are highlighted in Table 1. Practically, the application of WBAN associated with medical devices used for inner body (internal) gives the issues occurred due to signal interference (from same devices) or unfamiliar physical behavior. The electroencephalogram (EEG) is a signal which measures the brain activities by using electrical signals generated by brains where noise may be generated (unfamiliar signal interference). The similar thing can be found in electrocardiography (ECG) which helps in heart rate measurement.

The main cause of energy dissipation in the WBAN system is signal interference from system to system of WBAN, WBAN scheduling, WBAN fault detection, WBAN analysis, etc. Hence the energy dissipation is more crucial factor in WBAN system as it impacts the longevity of network directly. In recent years various techniques were introduced to tackle the energy dissipation issue. The scheduling of WBAN is indicated as optimization problem and is non-linear in nature where obtaining the maximum throughput is main objective [1]. Also, the reduction of dimension in WBAN is indicated with the techniques like component reduction and extraction performed by component analysis mechanism. The modular NN based associative memory is presented to detect the fault and analyze the WBAN system.

Also, with all the addressed above problems of WBAN, to make the capable system for real time applications computational efficiency need to be considered. This computational efficiency is mainly has impact on data training and prediction for WBAN system.

In this section, the implementation of the system is discussed. First the data for training is considered which is represented as data1 which consists of samples along its columns and its respective weights along its rows. Data1 is sent for the preprocessing stage which extracts principle components and is further performed an optimization with respect to conjugate gradient descent algorithm. The following section explains steps involved in the process in detail:

- *Pre-processing of Principle Component Analysis:* Prior to performing the computation for the gradient descent algorithm. The considered data is first applied a pre-processing module whose primary function is to extract the first m number of principle components using Algorithm 1. The data considered in this process consists of samples which are represented along the columns (x) and their respective weights along its rows (y). This data is represented as data 1.

The covariance is calculated with respect to x and y of data 1, which is represented as shown in (1):

$$cov(x, y) = E [(X - E[X])(Y - E[Y])] \quad (1)$$

Where, X, Y → rows and columns of data, respectively

E[x], E[y] → mean of X and Y, respectively

Following the computation of the covariance of the matrix, the Eigen vectors for the obtained covariance is computed which is given in (2):

$$T(v) = \lambda (cov(x, y)) \quad (2)$$

Where, v → Eigen vector

λ → Eigen value

Equation (2) is alternatively represented as given in (3),

$$(T - \lambda I)v = 0 \quad (3)$$

Where, I → Identity matrix

T → transformation matrix

Considering the obtained Eigen vectors which are linearly independent, A square matrix which consists of n linearly independent vectors is defined which is given in (4) as shown below:

$$Q = [v_1 v_2 \dots v_n] \quad (4)$$

Where, Q → square matrix consisting of n linear independent Eigen vectors

$v_1 v_2 \dots v_n$ → Eigen vectors

The matrix containing the linearly independent Eigen vectors is multiplied with their corresponding Eigen value which is given in (5) as shown below:

$$AQ = [\lambda_1 v_1 \lambda_2 v_2 \dots \lambda_n v_n] \quad (5)$$

The diagonal elements along the matrix AQ are obtained which considers the ith column of the matrix defined in (4). The same is represented in (6) as follows:

$$AQ = Q \Lambda \quad (6)$$

Where, Λ → diagonal matrix consisting of Eigen value associated with ith column of Q.

The matrix Q is invertible due the consideration that the columns present in the matrix in (5) is linearly independent. Hence, by multiplying Q^{-1} on both sides, we get,

$$A = Q \Lambda Q^{-1} \quad (7)$$

TABLE I. CHALLENGES OF WBAN SYSTEMS

Sl. No	Issues	Methods Applied
1	Energy dissipation	<ul style="list-style-type: none"> • Learning based algorithm [1] • Efficient design of routing protocols • Methods to be applied to increase the network lifetime.
2	Multiple WBAN scheduling	<ul style="list-style-type: none"> • Non-linear optimization with an objective norm of maximum throughput [1]. • Clique base WBAN scheduling [17] • Coloring based scheduling [18] • Intra WBAN based scheduling [18]
3	Inference between multiple WBAN systems (Dimensionality issue)	<ul style="list-style-type: none"> • Component extraction, component identification and analysis such as PCA. • Channel estimation for efficient dimensionality reduction [19] • Design of WBAN system at physical layer [19]
4	Fault detection and analysis	<ul style="list-style-type: none"> • Modular neural network based associative memory for training and prediction. • SVM based models [20] • Linear regression models [20]

The above (7) represents the Eigen decomposition which is considered from a similarity transform. The n principle components are then extracted using the following equation:

$$pca_{data} = v \times (data1 - \sum_{i=1}^N [d_{mean}(i)]) \quad (8)$$

Where, pca_{data} → principle components

d_{mean} → Mean of the principle components

The residual error post subsequent extraction of the principle components are given as shown in (7).

$$res_{error} = \sum_{i=1}^N \frac{(v \times (data1 - \sum_{i=1}^N [d_{mean}(i)]))^2}{d_{org}} \quad (9)$$

Where, res_{error} → residual error post extraction of principle components.

The next stage in the proposed system is the learning stage where the respective machine learning algorithm is employed. Optimization is to be performed, in this case a particular type of optimization called as the convex optimization is employed which is possible to the assumption that the system concerning the WBAN is a system of linear equations. The optimization algorithm used in this case is the Conjugate Gradient Descent based optimization as stated in Algorithm 2.

Algorithm 1: Preprocessing of PCA

1. Start
2. Initialize $data1, nc, v, mean, N$
3. Initialized_mean, inv_v
4. $N \leftarrow$ obtain column of data1
5. $d_org \leftarrow$ obtain row of data1
6. $v \leftarrow$ compute eigen vector for covariance of data1
7. $v \leftarrow$ extract diagonal of v
8. $inv_v \leftarrow$ compute inverse of v
9. **if** data1 \cong 0 **Then**
 - a. $pca_{data} = v \times (data1 - \sum_{i=1}^N [d_{mean}(i)])$
 - b. $res_{error} = \sum_{i=1}^N \frac{(v \times (data1 - \sum_{i=1}^N [d_{mean}(i)]))^2}{d_org}$
- End
10. End

- **Conjugate Gradient Descent of non-linearity:** In the context of non-linearity in the optimization process, the appropriate method considered for this condition would be conjugate gradient descent algorithm which is a non-linear optimization in nature. For the function of quadratic nature given as f(x) which is given in (10) below:

$$f(x) = \|Ax - b\|^2 \tag{10}$$

Where, A, b \rightarrow constraints

At the point when the gradient is 0, the local minimum is achieved which is given in (11) as shown below:

$$\nabla_x f = 2 \cdot A^T (Ax - b) = 0 \tag{11}$$

Where, $\nabla_x f \rightarrow$ gradient function with respect to x.

In order to obtain the local minimum, an iterative process called as *line search* is performed which moves in a particular direction (along the line) in an iterative process until a local minimum is achieved. Here, $\nabla_x f$ indicates the direction in which there is an increase in the maximum value. Hence to begin the line search, the initial iterative point begins in the direction opposite (or the steepest descent) to $\nabla_x f$. This is represented by the following equation as shown below:

$$\Delta x_0 = -\nabla_x f(x_0) \tag{12}$$

To minimize the above mentioned objective function, the linear constraint β_n is obtained by performing one of the following methods [21], [22]:

$$\beta_n^{FR} = \frac{\Delta x_n^T \Delta x_n}{\Delta x_{n-1}^T \Delta x_{n-1}} \tag{13}$$

Algorithm 2: Conjugate gradient Algorithm

1. **Start**
2. **Initialize** $x_n, \alpha_n, \beta_n, S_n$
3. $x_n \leftarrow$ compute for search direction towards negative gradient
4. $\alpha \leftarrow$ step length
5. **if** x_n is non_empty, **Then**
 - a. **Compute** conjugate gradient
 - b. $S_n \rightarrow$ conjugate gradient direction update
 - c. $\alpha_n = \text{argmin}_\alpha f(x_n + \alpha S_n)$
 - d. $x_{n+1} \rightarrow$ update position of conjugate gradient
- End**
6. **End**

Where, $\beta_n^{FR} \rightarrow$ Fletcher- Reeves method for function minimization.

Another method used for function minimization using the conjugate gradient method is the Polak – Ribiere method which is given in (14) as follows:

$$\beta_n^{PR} = \frac{\Delta x_n^T (\Delta x_n - \Delta x_{n-1})}{\Delta x_{n-1}^T \Delta x_{n-1}} \tag{14}$$

Where, $\beta_n^{PR} \rightarrow$ Polak – Ribiere method for function minimization.

The direction corresponding to the conjugate is updated using the equation given below:

$$s_n = \Delta x_n + \beta_n s_{n-1} \tag{15}$$

Where, $s_n \rightarrow$ conjugate direction

A line search is performed which computes the local minima along the direction of steepest descent. This is represented in x as shown below:

$$\alpha_n = \text{arg min}_\alpha f(x_n + \alpha S_n) \tag{16}$$

Where, $\alpha_n \rightarrow$ objective function

The position required for the consecutive iteration is updated by the following equation as given below:

$$x_{n+1} = x_n + \alpha_n S_n \tag{17}$$

Where, $x_{n+1} \rightarrow$ updated position along the conjugate direction.

The variables used in the development process are represented in the following table:

TABLE II. PARAMETER DESCRIPTION

Sl.No	Parameter	Description
1.	data1	Samples considered for training
2.	x	Samples present in the data1
3.	y	Respective weights present in data1
4.	v	Eigen vector
5.	λ	Eigen value
6.	Q	Matrix of linear independent Eigen vectors
7.	Λ	Diagonal matrix consisting of Eigen value
8.	pca_{data}	Principle components
9.	d_{mean}	Mean of principle components
10.	res_{error}	Residual errors
11.	$f(x)$	Quadratic function
12.	A, b	constraints
13.	$\nabla_x f$	Gradient function w.r.t x
14.	Δx_0	Initial iterative point
15.	β_n^{FR}	Function minimization using Fletcher-Reeves method
16.	β_n^{PR}	Function minimization using Polak – Ribiere method
17.	s_n	Conjugate direction
18.	α_n	Objective function
19.	x_{n+1}	Updated position along conjugate direction

VI. RESULTS AND DISCUSSIONS

In this section the system is analyzed by measuring two parameters which are the mean square error and the regression based correlation coefficient. The input data considered in this experiment is a matrix consisting of 1000 samples with 10 sensors which is given in Table 2.

- **Mean Square Error:** The estimation of the prediction error is computed with respect to risk estimation which corresponds to squared loss also sometime called as the quadratic loss. The characteristic of this error is mainly due to the factor of randomness which is present in the estimate. To ensure the quality of the measure with to the estimation the mean square error is used for the same. The MSE is of second order moment which constitutes the variance and its estimation with respect to bias (Fig. 2). The MSE estimation considering the predictor system can be defined as shown in (18) below:

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{Y}_i - Y_i)^2 \quad (18)$$

Where, $Y_i \rightarrow$ vector of observed values

$\hat{Y}_i \rightarrow$ Predicted vector

The estimator of the MSE is defined by a variable called θ , which is given as,

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] \quad (19)$$

The above defined equation is mainly based on the unknown parameter which in sense is considered to be the property of the estimator of the MSE.

The MSE is considered as the average sum considering the estimators variance and the square bias.

The MSE along with its equivalent variance is given as shown in (20) as follows:

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + Bias(\hat{\theta}, \theta)^2 \quad (20)$$

- **Regression based Correlation coefficient:** The coefficient of determination which is considered in this particular metrics is defined as the proportion with respect to variance corresponding to the dependent variable which is obtained from the respective independent variable (Fig. 3).

Considering the mean of the observed data which is represented as \bar{y} , its variability is given as,

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \quad (21)$$

Where, $y_i \rightarrow$ modeled vector

The variability of data is computed by using the following methods:

$$SS_{tot} = \sum_i (y_i - \bar{y})^2 \quad (22)$$

Where, $SS_{reg} \rightarrow$ sum of squares of total variable

$$SS_{reg} = \sum_i (f_i - \bar{y})^2 \quad (23)$$

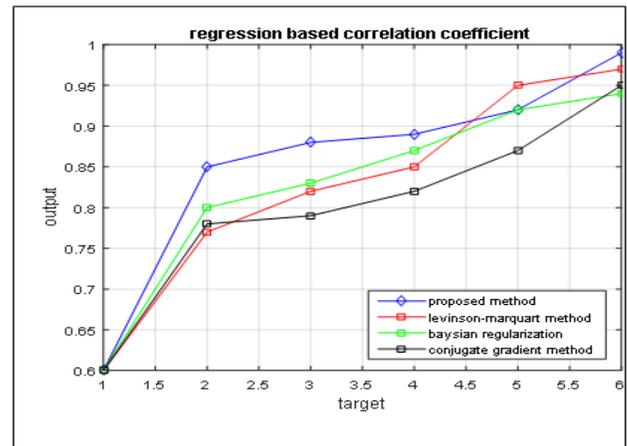


Fig. 2. Comparative analysis for regression based correlation coefficient.

Where, $SS_{reg} \rightarrow$ regression based sum of squares of total variable

$$SS_{res} = \sum_i (y_i - f_i)^2 \quad (24)$$

Where, $SS_{res} \rightarrow$ residual based sum of squares

The overall general definition of the regression based correlation coefficient is,

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \quad (25)$$

Where, $R^2 \rightarrow$ coefficient of determination

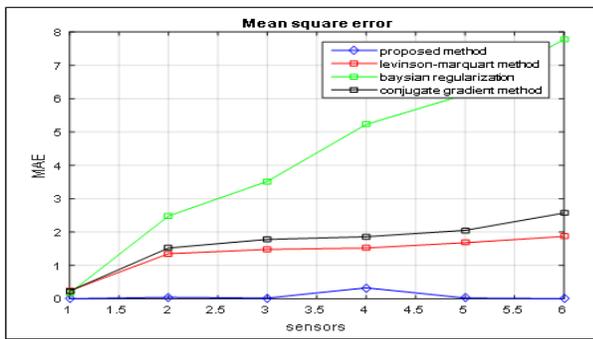


Fig. 3. Comparative analysis with respect to MSE.

The training sample data considered for this experiment is represented in Table 3 as shown below. The network structure of the proposed auto associative neural network considering the bottleneck configuration is represented in Fig. 4.

In Fig. 4 of the AANN architecture, the network configuration considered is here is [6 12 6 12 6],

where a network consisting of 6 neurons is mapped to a network consisting of 12 neurons which is again compressed to a network of 6 neurons which is called as the bottleneck as shown in the above figure. The DE mapping process again consists of a network having 12 neurons which are in turn connected to 6 neurons which are then applied for the process of prediction.

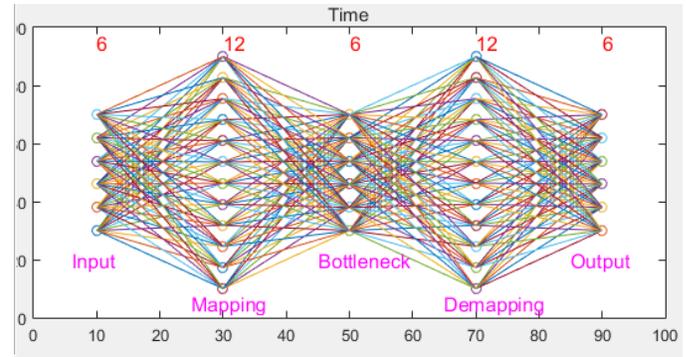


Fig. 4. Network configuration for the NLPCA based AANN system.

TABLE III. DATABASE FOR TRAINING SAMPLES CONSIDERED

Number of samples	200 samples	400 samples	600 samples	800 samples	1000 samples
Sensor 3	0.003344783	0.001562586	0.002399367	0.008482183	0.003442298
Sensor 4	0.046968715	0.043513381	0.4472326	0.046887739	0.046652258
Sensor 5	0.036573744	0.024047742	0.021381542	0.018151447	0.016981448
Sensor 6	0.336240087	0.326675964	0.321492001	0.304267503	0.305928877
Sensor 8	0.023718774	0.02296842	0.02344671	0.02329735	0.023237142
Sensor 10	0.009956085	0.005232597	0.004373233	0.005903767	0.006529151

VII. CONCLUSION

Experimental results show that the combination of principle component extraction along with conjugate based optimization provides improved results of energy dissipation in WBAN that takes place due to inaccurate prediction, dimensionality reduction, and nonlinearity. A comparative analysis is performed considering the neural network methods such as Baysian regression, Conjugate gradient method, and Levinson-Marquart method. The proposed method shows a significant reduction in the MSE as compared to other methods (Fig. 2). The regression based correlation coefficient has improves in the proposed method as compared to other standard methods (Fig. 3).

REFERENCES

- [1] Li M, Liu J, Ma Z, Yuan C, Yuan B. Throughput optimization with fairness consideration for coexisting WBANs. In 2015 IEEE International Conference on Communications (ICC) 2015 Jun 8 (pp. 6418-6423). IEEE.
- [2] Kolasa M. Fast and energy efficient learning algorithm for Kohonen Neural Network realized in hardware. *acta mechanica et automatica*. 2012;6(3):52-7.
- [3] Egbogah EE, Fapojuwo AO. Achieving Energy Efficient Transmission in Wireless Body Area Networks for the Physiological Monitoring of Military Soldiers. In MILCOM 2013-2013 IEEE Military Communications Conference 2013 Nov 18 (pp. 1371-1376). IEEE.
- [4] Khan JY, Yuce MR, Bulger G, Harding B. Wireless body area network (WBAN) design techniques and performance evaluation. *Journal of medical systems*. 2012 Jun 1;36(3):1441-57.

- [5] Guo-Jian H, Gui-xiong L, Geng-xin C, Tie-qun C. Self-recovery method based on auto-associative neural network for intelligent sensors. In Intelligent Control and Automation (WCICA), 2010 8th World Congress on 2010 Jul 7 (pp. 6918-6922). IEEE.
- [6] Gupta AK, Singh YP. Analysis of Bidirectional Associative Memory of Neural Network Method in the Strid Recognition. In Computational Intelligence and Communication Networks (CICN), 2011 International Conference on 2011 Oct 7 (pp. 172-176). IEEE.
- [7] Ang CH, Jin C, Leong PH, van Schaik A. Spiking neural network-based auto-associative memory using FPGA interconnect delays. In Field-Programmable Technology (FPT), 2011 International Conference on 2011 Dec 12 (pp. 1-4). IEEE.
- [8] Lemma TA, Hashim FM. Wavelet analysis and auto-associative neural network based fault detection and diagnosis in an industrial gas turbine. In Business Engineering and Industrial Applications Colloquium (BEIAC), 2012 IEEE 2012 Apr 7 (pp. 103-108). IEEE.
- [9] Ravi V, Naveen N. Hybrid classifier based on particle swarm optimization trained auto associative neural networks as non-linear principal component analyzer: Application to banking. In 2012 12th International Conference on Intelligent Systems Design and Applications (ISDA) 2012 Nov 27 (pp. 77-82). IEEE.
- [10] Garimella S, Mallidi SH, Hermansky H. Regularized auto-associative neural networks for speaker verification. *IEEE Signal Processing Letters*. 2012 Dec;19(12):841-4.
- [11] Chakraborty S. Accurate Arrhythmia classification using auto-associative neural network. In 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) 2013 Jul 3 (pp. 4247-4250). IEEE.
- [12] Krstulović J, Miranda V, Costa AS, Pereira J. Towards an auto-associative topology state estimator. *IEEE transactions on power systems*. 2013 Jan 1;28(3):3311-8.

- [13] Zin ZM. Using auto-associative neural networks to compress and visualize multidimensional data. In *Ubiquitous Robots and Ambient Intelligence (URAI)*, 2014 11th International Conference on 2014 Nov 12 (pp. 408-412). IEEE.
- [14] Ito M, Ohyama W, Wakabayashi T, Kimura F. Rotated face recognition by manifold learning with auto-associative neural network. In *Frontiers of Computer Vision (FCV)*, 2015 21st Korea-Japan Joint Workshop on 2015 Jan 28 (pp. 1-4). IEEE.
- [15] Wang C, Yang Y. Robust face recognition from single training image per person via auto-associative memory neural network. In *Electrical and Control Engineering (ICECE)*, 2011 International Conference on 2011 Sep 16 (pp. 4947-4950). IEEE.
- [16] Zhang B, Zhou J. Video-based face recognition by Auto-Associative Elman Neural network. In *Image and Signal Processing (CISP)*, 2014 7th International Congress on 2014 Oct 14 (pp. 89-93). IEEE.
- [17] Xie Z, Huang G, He J, Zhang Y. A clique-based wban scheduling for mobile wireless body area networks. *Procedia Computer Science*. 2014 Dec 31;31:1092-101.
- [18] Seo S, Bang H, Lee H. Coloring-based scheduling for interactive game application with wireless body area networks. *The Journal of Supercomputing*. 2016 Jan 1;72(1):185-95.
- [19] Islam SR, Kwak KS. A comprehensive study of channel estimation for WBAN-based healthcare systems: feasibility of using multiband UWB. *Journal of medical systems*. 2012 Jun 1;36(3):1553-67.
- [20] Salem O, Guerassimov A, Mehaoua A, Marcus A, Furht B. Anomaly Detection in Medical Wireless Sensor Networks using SVM and Linear Regression Models.
- [21] R. Fletcher and C. M. Reeves, "Function minimization by conjugate gradients", *Comput. J.* 7 (1964), 149-154.
- [22] E. Polak and G. Ribière, "Note sur la convergence de directions conjuguées", *Rev. Française Informat Recherche Operationelle*, 3e Ann´ee 16 (1969), 35-43.

Improvement of Radial basis Function Interpolation Performance on Cranial Implant Design

Ferhat Atasoy

Computer Engineering Department
Karabuk University, Karabuk, Turkey

Baha Sen

Computer Engineering Department
Ankara Yildirim Beyazıt University, Ankara, Turkey

Fatih Nar

Computer Engineering Department
Konya Food and Agriculture University, Konya, Turkey

Ismail Bozkurt

Neurosurgery Department
Cankiri State Hospital, Cankiri, Turkey

Abstract—Cranioplasty is a neurosurgical operation for repairing cranial defects that have occurred in a previous operation or trauma. Various methods have been presented for cranioplasty from past to present. In computer-aided design based methods, quality of an implant depends on operator's talent. In mathematical model based methods, such as curve-fitting and various interpolations, healthy parts of a skull are used to generate implant model. Researchers have studied to improve performance of mathematical models which are independent from operators' talent. In this study, improvement of radial basis function (RBF) interpolation performance using symmetrical data is presented. Since we focused on the improvement of RBF interpolation performance on cranial implant design, results were compared with previous studies involving the same technique. In comparison with previously presented results, difference between the computed implant model and the original skull was reduced from 7 mm to 2 mm using newly proposed approach.

Keywords—Cranioplasty; interpolation on medical images; radial basis function interpolation; symmetrical data

I. INTRODUCTION

Cranioplasty is a neurosurgical operation for repairing cranial defects that have occurred in a previous operation or trauma. This operation is important for both aesthetics and health [1]. Encephalitis, cerebritis, trauma, malignancy, hydrocephalus, epilepsy, mental or psychological disorders are associated with cranial bone defects [2], [3]. The main goals of cranioplasty are protection of intracranial contents and providing normal development and growth of the brain in children [4].

Various metals, ceramics, synthetic materials can be used for cranioplasty. The task is to complete the damaged skull bone with the selected material. Cranioplasty operations are performed on frontal bone, parietal bone, occipital bone, sphenoid bone, and portion of the temporal bone [1].

The oldest cranial operations dates back to 7000 B.C. in ancient Egypt [1], [5]. Archaeological finds indicate that inorganic materials have been used much earlier than organic materials. Bones were used for cranioplasty from a wide population of donor groups such as rib bone and tibia, in the

19th century. Although many different materials and methods have been described up to now, there is no consensus on which method is better [1].

An ideal implant material must have following features for cranioplasty applications [1], [5]:

- It must close and fit the defected part of the skull completely
- Not dilated with heat
- Resistance to infections
- Radiolucency
- Lightweight and compatible with tissues (biocompatibility)
- Easy to shape
- Ready to use
- Not expensive
- Resistant to biomechanical procedures
- Easily sterilized
- Non-inflammatory and non-carcinogenic

Thickness of implant varies according to the material. Therefore, implant mold should be specially created for implant material. While surface interpolation may be a good choice to manufacture titanium implant, it may not be right choice for cement-based materials such as methacrylate.

Computer-aided manufacturing of cranial implants have come into use with increasing processing speed of computers and development on imaging and modeling. In previous studies, implants were created with mathematical model or using solid modeling software.

Carr et al. designed cranial implants with radial basis function (RBF) based surface interpolation method on computed tomography (CT) images. In the study, they began with detection of defected part of skull and a height map was created for the defected part and nearby. Unknown areas (greater values on height map) were computed with RBF by

This study was supported by Ankara Yildirim Beyazıt University as pre-scientific research project with 587 project number.

using known values of neighbors (obtained from non-damaged part of the skull). The results were computed with various radial kernels and thin-plate spline was specified as the optimal kernel for cranial implant design [6].

Heissler et al. designed titanium implants for the defected part by using CAD/CAM according to the anatomical structure on the healthy (symmetrical) side of a skull. In this study, healthy side of the skull was mirrored and the mirrored data was applied for defected part of the skull. 12 male and 3 female patients between 21 and 35 years old were treated clinically and only one of the implants was removed due to premature infection. Reason of infection was interpreted as a corner of the implant may not be fully placed [7].

Lee et al. made manipulations on CT images to perform simulation, segmentation and planning processes. They used polymethylmethacrylate material in their work and developed a rapid prototyping device. The subject of the study was an 8-year-old boy. There is a large defect on his skull's left side. In the study, mirrored image of healthy side of the skull was used and a device was developed which can quickly produce implants as a result of the obtained data [8].

Fu et al. used multi-point forming and reverse engineering techniques for implant design. They used arithmetical profile curve blending method based on a well-proportioned point cloud data acquired with analyzing the patients' CT images. They produced titanium implant model for new points by using multiple point forming pressure machine [9].

Gerber et al. designed patient-specific cranial implants with low-cost material polymethylmethacrylate by using computer aided design method. Despite low-cost and widespread use of polymethylmethacrylate material, the technique is a time-consuming method. Thus, surgery time and risk of infection increase. In this study, implant manufacturing and operation time was shortened by using computer aided implant design method. Primarily, a mold was created with the patient's CT data by using 3D printer. Then, the implant was produced by using the mold and it was implanted to the patient. Three patients were treated successfully using the proposed technique [10].

Yusoff et al. created a 3D model from 2D CT images. In this study, they called the technique as biomedical model and used the model in pre-surgery planning. The manufactured model was produced in 45 hours and 19 minutes depending on the size and complexity. Entire skull was modeled with the manufactured bio-model as the manufactured bio-model provided educational usage and testing before surgical operations [11].

Kun et al. designed implants by using OpenCV and OpenGL object viewer interactively. A user selected symmetry plane on 2D CT images and defected region completed with symmetrical data. Then, OpenCV filled the defected region with symmetrical contour. At the end, the user modified the implant because of asymmetry of the skull [12].

Castelan et al. studied modeling and manufacturing technique of an implant. They manufactured a bio-model of a skull by a 3D printer and modeled an implant using Solidworks software. Titanium was preferred as the implant material in the

study since its high-strength characteristic. In the study, symmetrical data and CAD software were used to create the implant [13].

In proposed method by Rudek et al., a missing region of a skull was defined by curvature descriptors. They applied optimization technique of artificial bee colony to estimate descriptor parameters. The estimated descriptors were searched in database to replace defected region. Thus, an implant was modeled from similar images automatically [14].

Van der Meer described a technique to design a cranial implant for all sort of defects. In the study; process, material selection, design, and production were fully controlled by a user. He used Geomatic Studio 12 for data conversion from dicom to surface model and filled holes by curvature-based filling on the software [15].

In general, an operator uses CAD/CAM software to design an implant and symmetrical data is used to complete defected region of a skull. If the operator is skilled, implant will be well-fitted and aesthetically successful. As a conclusion, success of implant manufacturing depends on the operator's talent. Mathematical model based methods are robust but they must be finely-modeled. However, since skull is not completely symmetric and symmetrical data is not always available, studies that use only symmetrical data are not robust. RBF is one of the mathematical models using known neighbors of data-free regions of a skull.

In this study, previously presented scattered data estimation for cranioplasty applications with RBF [16] has been improved. For this purpose, implant shape was created by a mathematical model successfully for high-strength materials such as titanium. Obtained results are smoother and outer surface of the implant is similar to original. This is important in terms of the social and psychological status of the patients. The model has still disadvantages such as inner surface interpolation error and quality of results depends on symmetrical data, symmetry plane position and calculation of normal vector of scattered data.

II. METHOD

A. RBF Based Interpolation

Completion process of cranial defect is obviously a 3D curve fitting or scattered data estimation problem. The estimated data must be in accordance with non-damaged neighbors' bone data geometrically. RBF is one of the most frequently used modern approach to complete data-free regions. This approach is convenient when the problem depends on the multivariate or multi-parameters of scattered data. Besides, it is an appropriate estimation of scattered data in high dimensional space [6], [16], [17].

RBF approach is generally defined as [6], [12]:

$$f: \mathbb{R}^d \rightarrow \mathbb{R} \quad (1)$$

The real-valued function f of d variables is approximated by $s: \mathbb{R}^d \rightarrow \mathbb{R}$. Here, given values $\{f(x_i): 1 = 1, 2, \dots, n\}$, where $\{x_i: 1, 2, \dots, n\}$ is a set of discrete points in \mathbb{R}^d , are called the nodes of interpolation. Thus, approximation form of interpolation is achieved as:

$$s(x) = p_m(x) + \sum_{i=1}^n \lambda_i \phi(\|x - x_i\|), \quad (2)$$

$$x \in \mathbb{R}^d, \lambda_i \in \mathbb{R}$$

where, p_m is low-degree polynomial, or does not exist, $\|\cdot\|$ is Euclidean norm, and ϕ stands for a fixed function from \mathbb{R}^+ to \mathbb{R} . Radial basis function s is a linear combination of translates of the single radially symmetric function $\phi(\|\cdot\|)$, plus a low-degree polynomial. Space of all polynomials of degree at most m in d will be indicated by π_m^d . Afterwards, the coefficients λ_i of the approximation s are calculated by assuming that s fulfills the interpolation conditions as

$$s(x_j) = f(x_j), \quad j = 1, 2, \dots, n \quad (3)$$

with the side conditions:

$$\sum_{j=1}^n \lambda_j q(x_j) = 0, \text{ for all } q \in \pi_m^d \quad (4)$$

Some popular RBF kernels are given as:

$$\phi(r) = r \quad \text{Linear} \quad (5)$$

$$\phi(r) = r^2 \log r \quad \text{Thin-plate spline}$$

$$\phi(r) = e^{-\alpha r^2} \quad \text{Gaussian}$$

$$\phi(r) = \sqrt{r^2 + c^2} \quad \text{Multi-quadratic}$$

$$\phi(r) = (1 - \epsilon r)^4 (4\epsilon r + 1) \quad \text{Wendland}$$

where, α and c are positive constants and $r \geq 0$.

Let given values of known centers are (-3, -0.5, 1.8, 15, 16), the known values of unknown function are (5, 3, -6, 12, 40). Unknown values of the function in the working space (-32, 32) were computed in one-dimensional space to show the behaviors of the kernel functions. According to RBF kernels, obtained results are given in Fig. 1.

Here, linear kernel result is not given since its behavior is well-known. When Fig. 1 is examined carefully, thin-plate spline kernel generates smooth curve. As mentioned in previous published studies, thin-plate spline kernel is well-suited for cranial defects [6], [16].

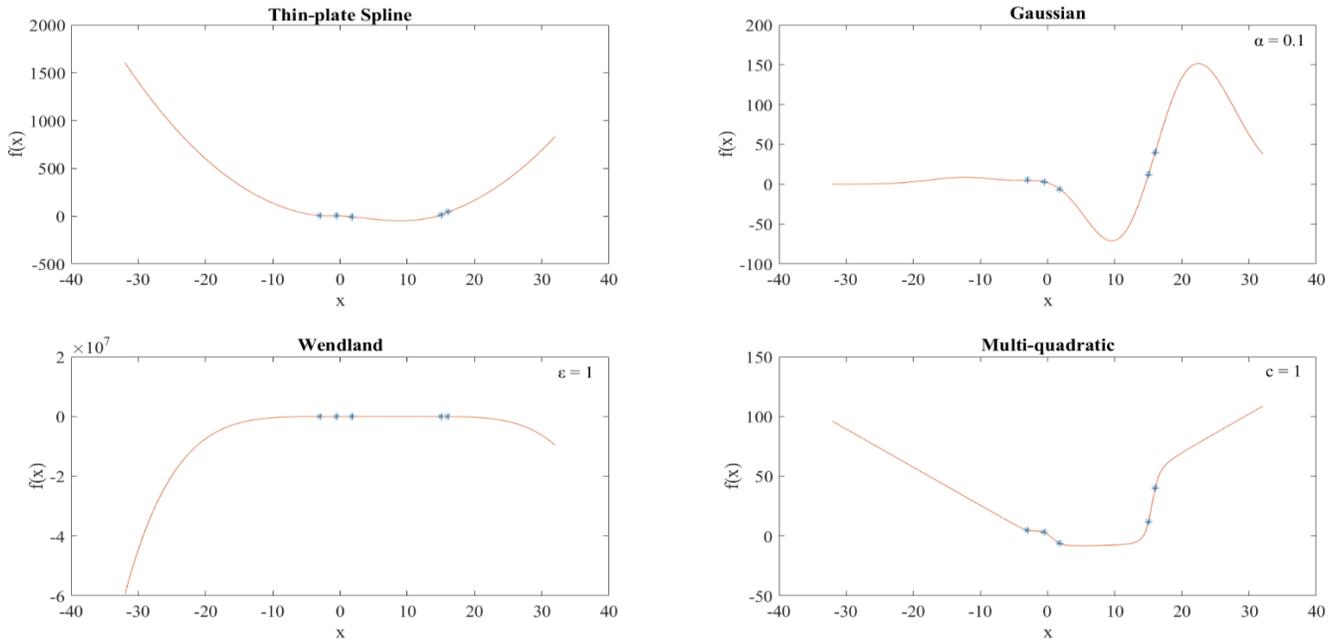


Fig. 1. The obtained curves of various RBF kernels.

In 2D space, RBF method is applied as the following [6], [16], [17]:

$$\begin{bmatrix} A & Q \\ Q^T & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ c \end{bmatrix} = \begin{bmatrix} f \\ 0 \end{bmatrix} \quad (6)$$

where

$$A = (a_{ij}) = (\phi(\|x_i - x_j\|)) \quad (7)$$

$$Q = \begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & y_n \end{bmatrix} \quad (8)$$

$$\lambda = (\lambda_1, \lambda_2, \dots, \lambda_n)^T \quad (9)$$

$$c = (c_0, c_1, c_2)^T \quad (10)$$

$$p_m(x) = c_0 + c_1 x + c_2 y \quad (11)$$

$$f = (f_1, f_2, \dots, f_n)^T \quad (12)$$

B. Proposed Improvement

RBF method has been used successfully in previous studies for scattered data interpolation [6], [16], removing an object from image [18] and image inpainting applications [18], [19].

Carr et al. published their study about RBF interpolation for cranioplasty applications. According to the study, RBF interpolation method is superior than parametric spline interpolants, tensor product spline interpolants, etc. Even with large missing region, RBF has variable characterizations that make them well-suited for scattered data interpolation. However, we stated in our previous study [16] that scattered data interpolation generated good results by using RBF if large defected region is not on elliptical regions of a skull. The elliptical region of the skull is given in Fig. 2. When the diameter of defected region on elliptical region grows (bigger than 13 mm), generated results deteriorate. Thereby, we propose that RBF interpolation performance in cranioplasty applications can be improved with using symmetrical data information.

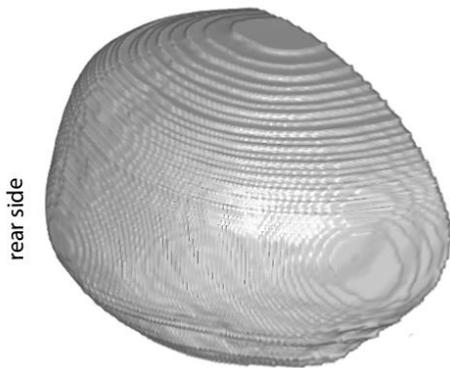


Fig. 2. Elliptical region of a skull.

The method which we implemented in our previous study is described below:

- Scattered data is obtained from CT images.
- Skull is segmented (soft tissue and background is removed).
- Normal vector of defected region is calculated by using its neighbors.
- A plane is created d mm away from the defected region in the normal vector direction (green plane in Fig. 3).
- A height map is created by sending rays (red lines in Fig. 3) from the green plane to the skull.
- Distances of the rays that touch the skull surface (blue region in Fig. 3) are determined as RBF known values of center points.
- Distances of the rays that do not touch the skull surface are determined as defected region.
- λ and polynomial coefficients are calculated for known center points.
- The distances of unknown points (defected region) are computed by RBF method with λ and polynomial coefficients.
- For the defected region, voxel values are updated according to the computed distances.

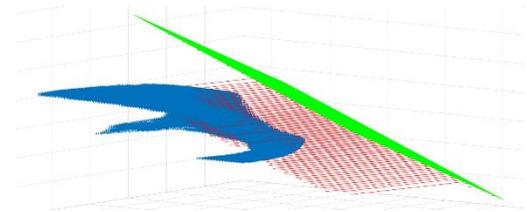


Fig. 3. Calculating the distances between plane and scattered data.

For thickness of the cement based implants, rays that touch the skull is continued along the bone. Thus, thickness map of the implant is calculated.

Fig. 4 shows 3D calculated height map of the defected region and its neighbors. Non-significant data on the height map was deleted before visualization.

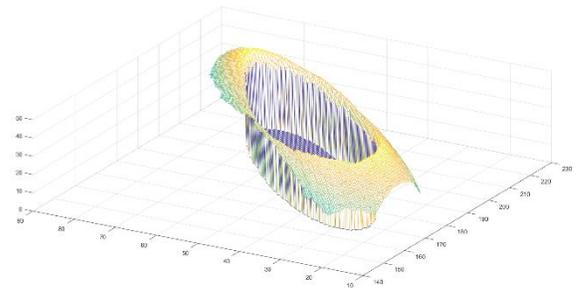


Fig. 4. 3D calculated height map of the defected region and its neighbours.

Weakness of the defined method comes up with large defects, because healthy neighbors of the defected region remain far away from the center of large defects. Therefore, computed distances become non-acceptable for the defected region. Hereby, symmetrical data of the defected region can be used to improve RBF performance. Axial view of the ineffective surface interpolation example is shown in Fig. 5. Bold red region in Fig. 5(a) denotes original skull while bold green region denotes the symmetrical data in Fig. 5(b) and thin white line demonstrates surface interpolation result in Fig. 5(a), (b) and (c).

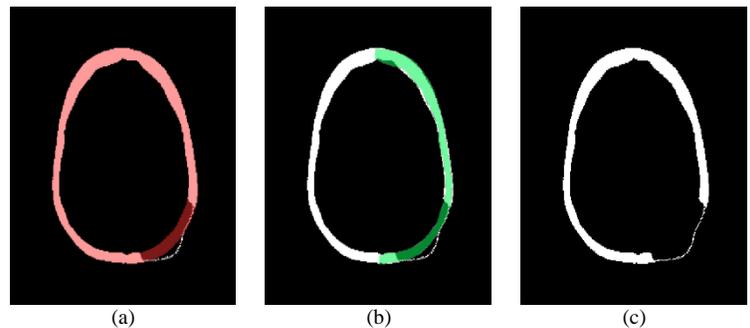


Fig. 5. Surface interpolation results (a) compared with original skull, (b) compared with symmetrical data, (c) RBF interpolation result.

For this improvement, we suggest that symmetrical data of the defected region can be used as healthy neighbors. Thus, maximum error is limited by using this new approach. In this approach, computed height map is updated with symmetrical data. In Fig. 6, black regions show non-significant data for scattered data interpolation, white region is the defected part of

the skull, grey region represents height values of healthy neighbors, and grey dots in white region demonstrates height values sampled from symmetrical side.

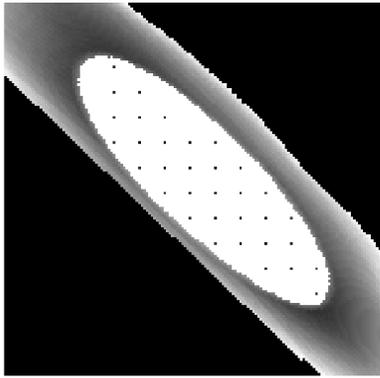


Fig. 6. The 2D projection of the height map.

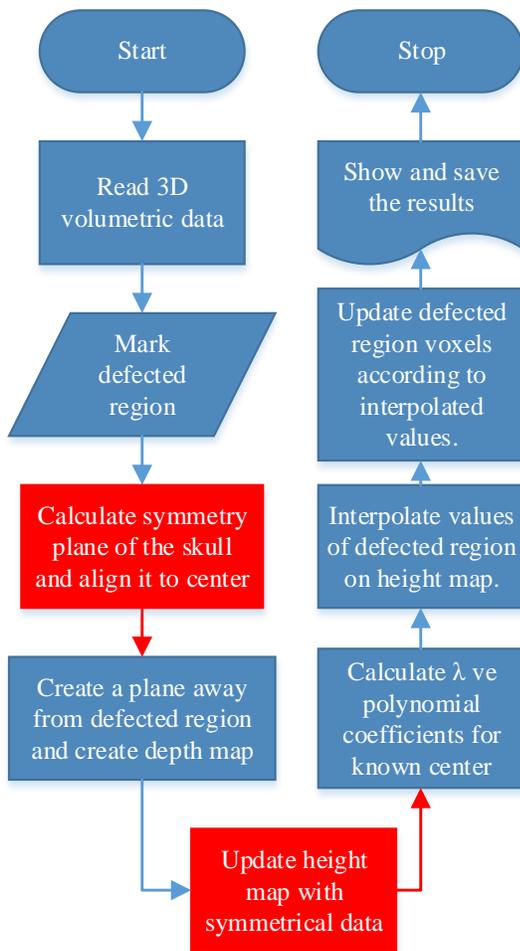


Fig. 7. Flowchart of the previously proposed method (blue boxes) and suggested improvements (red boxes).

The mirrored symmetrical side of the skull is replaced with the defected side. Then, rays are sent periodically from the green plane (Fig. 3) to the skull for the defected region. Thus, sampled distances of the symmetrical data are observed. At the

end, the sampled distances are 1mm shortened for the implant to stay outside of the defected region.

Modified flowchart of the new approach is shown in Fig. 7. In the figure, previously proposed method (blue boxes) and suggested improvement (red boxes) are given.

III. RESULTS AND DISCUSSIONS

The CT image used in this study is the same in our previous study with 1 mm slice spacing and 512 x 512 resolution. The difference of 7 mm between the computed implant model and the original skull in the previous study was reduced to 2 mm with the proposed new approach. Diameter of the defected region was 13 mm in the tests. Intensity-based mutual information was used for symmetry plane computation. Defected region was marked in 3D Slicer and the marked region was used as a mask. Defected region and its symmetry were not considered during computation.

Results of the proposed method were compared with previous studies [6], [16]. Achievement of scattered data interpolation with RBF for cranioplasty applications was improved by using the symmetrical data. The differences between previously proposed method and new approach are shown in detail in Fig. 8 and 9. If Fig. 8 is compared carefully, it can be seen that new approach improved the success of scattered data interpolation. The computed data on the outer surface is close to the original data. Thus, this achievement will provide a great advantage in terms of aesthetics, when the social and psychological conditions of the patients are considered, although the new approach has a disability on inner surface as shown in Fig. 9(c).

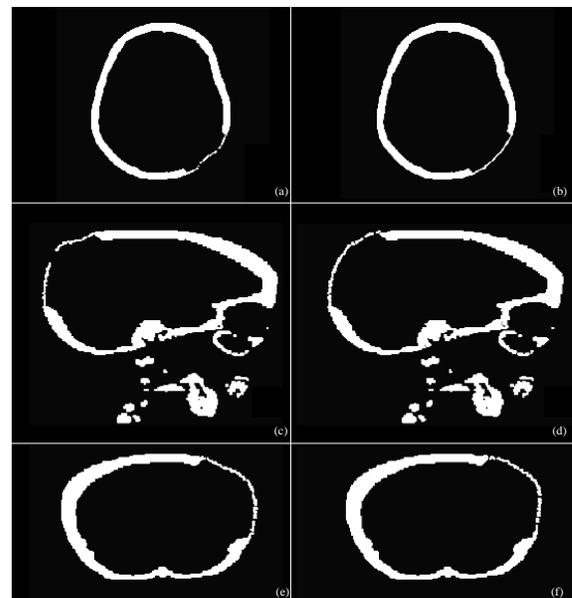


Fig. 8. Result comparison. The left side shows previously presented study results and the right side shows new approach results. (a) and (b) are axial slices, (c) and (d) are sagittal slices, (e) and (f) are coronal slices.

Since the thickness of the implant can be disregarded for high-strength materials such as titanium, the results of new approach will be feasible. When cement-based materials are used, the thickness of the implant becomes important.

Therefore, the inner surface of the results should be improved for the implants which are manufactured by 3D printers. Obtained results for inner surface is given in Fig. 9.

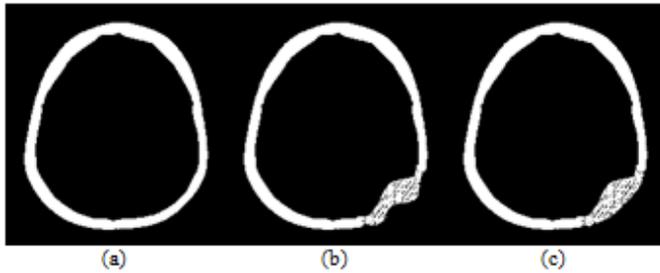


Fig. 9. Axial slices (a) Original data, (b) previously proposed method, (c) new approach.

It is already known that the previous presented method [16] is successful in tests which the defect size is small and thin-plate spline kernel is the optimal RBF kernel as given in the previous study. If upside of a skull is defected, even linear kernel results are successful. Therefore, large defects on elliptical region (parietal-occipital bones) are focused in this study. The 3D results of previously proposed method and new approach are shown in Fig. 10.

The new approach searches symmetrical data to improve RBF performance. If the defected region does not have symmetrical data, it is not crucial, scattered data interpolation is realized by previously proposed method.

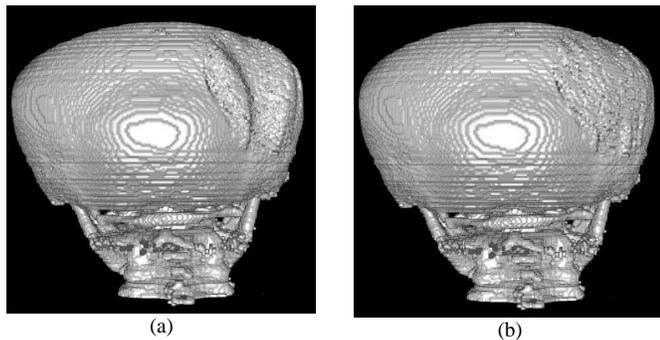


Fig. 10. The computed results, (a) previously proposed method, (b) new approach.

IV. CONCLUSION

In this study, RBF interpolation was improved for cranial implant design. Defected part on outer surface of a skull can be completed using RBF and symmetrical data for cranioplasty. As a result, difference between the computed implant model in previous study [16] and the original skull was reduced to 2 mm with the newly-proposed approach. In addition, removal of aesthetic concerns is a great success when the sociological and psychological conditions of patients are considered. Nevertheless, there is a problem for inner surfaces of skulls to minimize errors. Therefore, a new approach should be enhanced for cement-based implants which can be manufactured with 3D printers. In next study, inner surface

interpolation errors will be minimized and the thickness of cement based implants will be similar to original.

REFERENCES

- [1] S. Aydin, B. Kucukyuruk, B. Abuzayed, S. Aydin, and G. Z. Sanus, "Cranioplasty: Review of materials and techniques.," *J. Neurosci. Rural Pract.*, vol. 2, no. 2, pp. 162–167, 2011.
- [2] S. Chibbaro et al., "Decompressive craniectomy and early cranioplasty for the management of severe head injury: A prospective multicenter study on 147 patients," *World Neurosurg.*, vol. 75, no. 3–4, pp. 558–562, 2011.
- [3] S. Honeybul and K. M. Ho, "Long-term complications of decompressive craniectomy for head injury.," *J. Neurotrauma*, vol. 28, no. 6, pp. 929–935, 2011.
- [4] R. M. Redfern and H. Pülhorn, "Cranioplasty," *Adv Clin Neurosci Rehabil.*, vol. 7, no. 5, pp. 32–34, 2007.
- [5] Q. Yu et al., "Skull repair materials applied in cranioplasty: History and progress," *Transl. Neurosci. Clin.*, vol. 3, no. 1, pp. 48–57, 2017.
- [6] J. C. Carr, W. R. Fright, and R. K. Beatson, "Surface interpolation with radial basis functions for medical imaging," *IEEE Trans. Med. Imaging*, vol. 16, no. 1, pp. 96–107, 1997.
- [7] E. Heissler et al., "Custom-made cast titanium implants produced with CAD/CAM for the reconstruction of cranium defects.," *Int. J. Oral Maxillofac. Surg.*, vol. 27, no. 5, pp. 334–338, 1998.
- [8] M. Lee, C. Chang, C. Lin, L. Lo, and Y. Chen, "Three-Dimensional Image Reconstruction and Rapid Prototyping Models Improve Defect Evaluation , Treatment Planning , Implant Design , and Surgeon Accuracy," *Eng. Med. Biol.*, no. April, 2002.
- [9] H. Fu, L. Gao, L. Ju, and Y. Liu, "Personalized Cranium Defects Restoration Technique Based on Reverse Engineering," *Tsinghua Sci. Technol.*, vol. 14, no. SUPPL. 1, pp. 82–88, 2009.
- [10] N. Gerber, L. Stieglitz, M. Peterhans, L.-P. Nolte, A. Raabe, and S. Weber, "Using rapid prototyping molds to create patient specific polymethylmethacrylate implants in cranioplasty," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, 2010, pp. 3357–3360.
- [11] W. A. Y. W. Yusoff, H. H. M. Ali, and M. A. H. M. Shukri, "Fabrication of surgical cranioplasty biomodel using fused deposition modeling," *ICIMTR 2012 - 2012 Int. Conf. Innov. Manag. Technol. Res.*, pp. 550–554, 2012.
- [12] W.-M. Kung, S.-T. Chen, C.-H. Lin, Y.-M. Lu, T.-H. Chen, and M.-S. Lin, "Verifying Three-Dimensional Skull Model Reconstruction Using Cranial Index of Symmetry," *PLoS One*, vol. 8, no. 10, p. e74267, Oct. 2013.
- [13] J. Castelan, L. Schaeffer, A. Daleffe, D. Fritzen, V. Salvaro, and F. Pinto, "Manufacture of custom-made cranial implants from DICOM @ images using 3D printing , CAD / CAM technology and incremental sheet forming," vol. 30, no. 1996, pp. 265–273, 2014.
- [14] M. Rudek, Y. B. Gumiel, and O. Canciglieri, "Autonomous ct replacement method for the skull prosthesis modelling," *Facta Univ. Ser. Mech. Eng.*, vol. 13, no. 3, pp. 283–294, 2015.
- [15] W. J. Van Der Meer, "3D Workflows in Orthodontics , Maxillofacial Surgery and Prosthodontics," 2016.
- [16] F. Atasoy, F. Nar, B. Sen, and M. Ferat, "Scattered data estimation on medical images for cranioplasty applications," in *22nd Signal Processing and Communications Applications Conference (SIU)*, 2014, pp. 1682–1685.
- [17] V. Skala, "Fast reconstruction of corrupted images and videos by radial basis functions," 2013 *Int. Conf. Control. Autom. Inf. Sci. ICCAIS 2013*, pp. 267–271, 2013.
- [18] W. Wang, "An Image Inpainting Algorithm Based on CSRBF Interpolation," *Int. J. Inf. Technol.*, vol. 28, pp. 112–119, 2006.
- [19] M. Alsalamah and S. Amin, "Medical Image Inpainting with RBF Interpolation Technique," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 8, pp. 112–119, 2016.

Performance Evaluation of Transmission Line Protection Characteristics with DSTATCOM Implementation

Yasar Khan

Department of Electrical CECOS
University of IT and Emerging
Sciences, Peshawar, Pakistan

Khalid Mahmood

Department of Electrical CECOS
University of IT and Emerging
Sciences, Peshawar, Pakistan

Sanaullah Ahmad

Department of Electrical
Engineering, IQRA National
University (INU), Peshawar,
Pakistan

Abstract—To meet with the ever-enhancing load demands, new transmission lines should be bolted-on in the existing power system but the economic and environmental concerns are major constraints to this addition. Hence utilities have to rely on the existent power system infrastructure with some modifications. To enhance controllability and boost power transfer potential of the existing power system the use of Flexible Alternating Current Transmission System (FACTS) device is the most viable modification. FACTS devices include Static VAR Compensator (SVC), thyristor controlled series capacitor (TCSC), Thyristor Controlled Reactor (TCR), Thyristor Switched Capacitor (TSC) and Self Commutated VAR compensators i.e. Static Synchronous Compensator (DSTATCOM). Among the FACTS devices, DSTATCOM is the most feasible choice because of its capability to furnish both leading and lagging reactive power, faster response time in comparison with others, smaller harmonic content, inrush current generation is minimum and the dynamic performance with variations of voltage is quite good. DSTATCOM has the ability to have effective control over various issues concerning AC power transmission. However, the Parameters of the protection devices in the present power system are set without taking into account the reaction of these FACTS devices. So in order to ascertain stability and reliability of power system, reaction of FACTS devices with the existent protection schemes must be thoroughly investigated. This paper aims to explore the deviations in the performance characteristics of transmission line protection due to installation of DSTATCOM on a 220KV EHV transmission line using theoretical as well as MATLAB/SIMULINK simulation models. The dynamic performance of a DSTATCOM connected to an existing transmission line system is evaluated when large industrial induction motor is started and voltage sags are introduced.

Keywords—Power system analysis; DSTATCOM; transmission line loss minimization; distribution dynamic compensation; transmission losses and efficiency

I. INTRODUCTION

Industrial and domestic development has led to increased electrical energy demand day to day. To avoid power system instability, the existing power networks are being growingly interconnected. To cope with the ever-growing electrical energy demands, the existing transmission systems are often run at or more than their rated capacity which leads to problems in maintaining effective power flow distribution.

The intuitive and obvious solution to this problem is to construct new transmission lines but environmental and economic issues are major hurdles to this solution. So, we have to rely on the existing power system infrastructure with requisite modifications to fully utilize the capability of transmission system. These modifications necessarily involve “reactive power compensation” and initial solutions were capacitor banks and shunt reactors, but these proved to be very rigid and inefficient way outs. To achieve better controllability and escalated capability of transferring electrical power of present power system, FACTS devices were developed.

While we have achieved improved efficiency of power system employing the FACTS devices, it is altogether very important that the power system must be “Reliable” and “Dependable” to ensure stability of Power system [17]. These requirements of dependability and reliability are intemperately influenced by the installed power system protection. A dependable protection system must give tripping for faults within its defined zone and a reliable protection system must not fail to operate when it is called for its duty/function. Both of these factors lead to desirable Stable Power System. Different attributes of power system lead to various protective Relays e.g. over current relay, differential relay, Over voltage relay, etc. The major protection scheme for transmission systems is furnished with the distance/impedance relays. The principle on which the distance relay is modeled, when FACTS devices are installed within their defined protection zone, has led to the challenges that were not taken into account during development of protection devices and their parameter settings. Among FACTS devices, DSTATCOM is the most feasible choice for power flow control due to its desirable features and is the subject of this Paper.

II. FACTS DEVICES

The detailed literature survey depicts the increasing trends in the field of reactive power compensation incorporating FACTS devices to furnish greater controllability and escalated transfer capability of power. Researches also grabbed attention of concerned quarter about the influence of FACTS devices on dynamics of power system. Extensive studies pointed out the side effects of trending FACTS devices because these devices greatly influenced the dynamics of power system. Due to this

disturbance, many subsystems of power system are affected and the Protection system is the one that is highly impacted by this disturbance. In fact the question of disturbance in the protection system would have been raised since the introduction of capacitor banks or Shunt reactor as compensating devices but their dynamic behavior and response time were quite slower to actually affect the fast responding protection system.

Some research works have already been carried out on the influence of different FACTS devices on protection system behavior and a lot of work is underway to get comprehensive analysis and suggest practical solutions to the challenges offered by new trending FACTS devices. The distance relay operation problems employing shunt compensator at different locations of transmission line and found that mid-point compensation affected the relay behavior the most [1]. Both inductive and leading compensations are considered to show mal-operation of relay in the form of reach discrepancy but the effects of different faults and operating time issues are not addressed.

The performance of Impedance relay when power swing occurs on a transmission line [2]. Relay behavior is observed both for an uncompensated case the line compensated by UPFC. Different modes of operation of UPFC and the corresponding response of relay is monitored and found that relay performance is subjected to faulty measurements with UPFC inclusion.

In [3], the authors described the distance relay performance on a 400KV transmission line employing TCSC and DSTATCOM located at midpoint. Quadrilateral characteristics are chosen as case study using S-transform to show deviations of measured impedance with and without the inclusion of mentioned devices but effects of fault location, operating times variations are not taken into account to get proper insight.

In [4], the authors investigated the behavior of generator loss of excitation protection (LOE) installed at a hydro generator station having mid-point DSTATCOM installed on the transmission line. Using PSCAD simulations, results depicted that presence of DSTATCOM affected the performance of LOE causing relay delay time phenomenon and upsetting GUEC and relay coordination. It is also pointed out that for heavily loaded generators, DSTATCOM impact extends to healthy generators in parallel by prolonging its armature overloading time. Alternative methods/modifications in the LOE protection are also proposed.

The model of DSTATCOM and distance relay in PSCAD showed the Impedance trajectories for a single phase fault after the placement of DSTATCOM. Effect of level of compensation and errors in calculations are discussed but the effects of location of fault and the concept of critical location of fault are not considered [5].

The impact of location of DSTATCOM is on Impedance measurements. A phase to ground fault is introduced with DSTATCOM at start, at mid-point and at the far end of a 400kV transmission line and variation of tripping characteristics is noted but the effects of type of faults, various

fault locations and tripping time variations are not discussed [6]. The behavior of Distance protection on a transmission line of 400KV with GCSC compensates TCSC. Both devices are connected at the center of line. The author also discusses the impact of controlling angle variation on the total impedance measured by the relay [7]. MPSO technique is used to study the fluctuations of impedance relay behavior in the presence of mid-point TCSC compensation. The impact of firing angle using MPSO approach was also taken into consideration. Suggestions are also given at the conclusions to somehow improve the performance as per desired results [8].

The effects are on impedance relay performance in the presence of SSSC series FACTS device and DSTATCOM. Faults are considered at various locations on the line and behavior of relay is observed to be inappropriate. Effect of fault resistance is also considered to be a contributing factor of erroneous performance. Operating time delays are not highlighted as an effect of compensation [9]. A variation of measured impedance is due to installation of TCSR at midpoint of a transmission line rated at 400KV commencing phase to earth fault. Different ratings of TCSR are employed to get detailed insight of relay characteristics deviations. Study is concluded by proposing adaptive methods to overcome the mal-operation of relay [10].

Multiline VSC-based type FACTS controllers and showed that these have noticeable effect on the relay performance. Impact of IPFC, UPFC and GUPFC were analyzed and found that measured impedance was higher than expected leading to false operation of relay. It was also found that GUPFC has the most severe influence than others with IPFC having the least impact [11].

Mathematical approach is to visualize distance relay issues on a transmission line having DSTATCOM. Effect of load angle, symmetrical and unsymmetrical faults is proposed using mathematical results. One important point grabbing attention towards a resonance condition when DSTATCOM impedance equals line impedance between DSTATCOM to fault is also raised but its effect on system is not explained [12]. The behavior of distance relays with in-feed impact and out-feed impact of compensation on a 500KV transmission line. Different fault cases and different placement of compensating device are considered to comprehensively analyze the situation. The authors concluded the study by proposing setting rules for relay to achieve proper working. The issues are of distance relay erroneous behavior in a system including series and shunt FACTS devices. The series device included is SSSC and the shunt device is DSTATCOM. It is shown that due to system short circuit level, voltage level and load angle the compensation severely affected the calculation of impedance made by distance relay [13], [14], [19].

III. REACTIVE POWER COMPENSATION

To have a clear understanding of reactive power compensation consider a simplified prototype of electrical power transmission system as shown in Fig. 1. Fig. 1(a) shows the sample system and 1(b) its phasor diagram. A grid station at Bus1 (Sending end) with phasor voltage $V_1 = V_1 < \delta_1$ is connected to another Grid system at Bus2 (Receiving end)

with phasor voltage $V_2 = V_2 < \delta_2$ through a transmission line of length L having transmission line reactance of X_L . The transmission angle δ is defined as:

$$\delta = \delta_1 - \delta_2 \quad (1)$$

The voltage that is dropped across the line is defined to be the phasor difference between sending end and receiving end voltages as given:

$$V_L = V_1 - V_2 = V_1 < \delta_1 - V_2 < \delta_2 \quad (2)$$

The current flowing through the transmission line has the magnitude given as:

$$I_L = \frac{V_L}{X_L} = \frac{|V_1 < \delta_1 - V_2 < \delta_2|}{X_L} \quad (3)$$

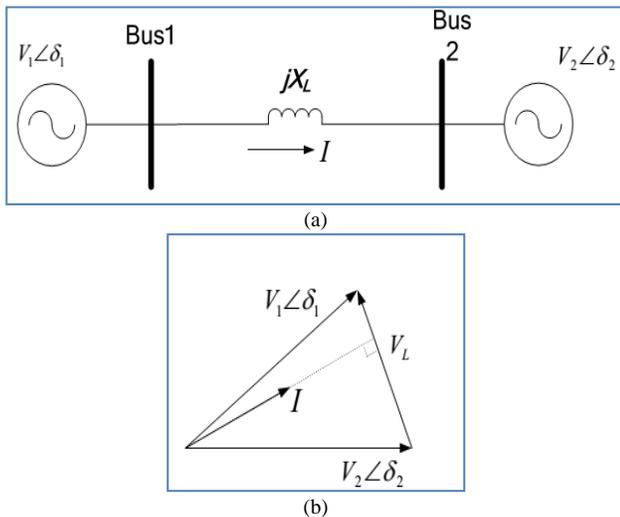


Fig. 1. Transmission system (a) Simplified (b) Phasor Diagram.

At Bus1 the Active part of current is

$$I_{P1} = \frac{V_2 \sin \delta}{X_L} \quad (4)$$

Reactive counterpart of current at Bus1 is

$$I_{Q1} = \frac{(V_1 V_2 \cos \delta)}{X_L} \quad (5)$$

Hence the Active Power flow from Bus1 is

$$P_1 = \frac{V_1 - V_2 \sin \delta}{X_L} \quad (6)$$

And the Reactive Power flow from Bus1 is

$$Q_1 = \frac{V_1(V_1 - V_2 \cos \delta)}{X_L} \quad (7)$$

Likewise the active part of current at Bus2 is

$$= \frac{V_2 \sin \delta}{X_L}$$

And the Reactive counterpart of current at Bus2 is

$$I_{Q2} = \frac{(V_2 - V_1 \cos \delta)}{X_L} \quad (8)$$

The Active Power at receiving side Bus2 is

$$P_2 = \frac{V_1 V_2 \sin \delta}{X_L} \quad (9)$$

And the Reactive Power at receiving side is

$$Q_2 = \frac{V_2(V_2 - V_1 \cos \delta)}{X_L} \quad (10)$$

Equations (4) to (7) suggest that we can regulate the flow of active and reactive power or current by having control over:

- Voltages at sending and receiving ends (V_1 & V_2),
- Angle (δ) of transmission line, and
- Reactance (X_L) of transmission line [15].

IV. TYPES OF FACTS DEVICES

In general, the FACTS devices are classified into generations as depicted in the flow chart. The first generation consists of typical rigid devices including phase varying and taps changing transformers, fixed capacitor combinations and synchronous condensers and are usually controlled at the generating end of the power grid [18]. These are rigid and costly solutions having minimum control over desired parameters. As type of interest the 2nd generation Static type compensators are superior to first generation due to fast responding nature and amelioration in the transient and dynamic functioning of power system and are divided into:

- 1) Conventional thyristor based devices &
- 2) Voltage Source based devices

The flow chart depicting types of FACTS devices is shown in Fig. 2.

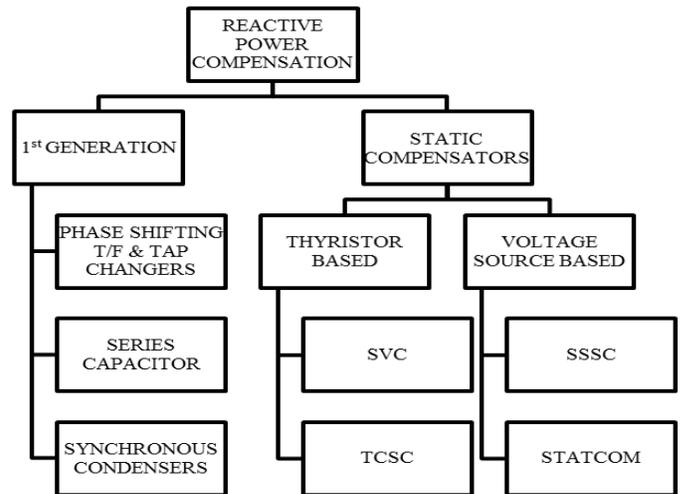


Fig. 2. Types of FACTS Devices Flow Chart.

A. Thyristor Based Conventional FACTS Devices

These devices employ thyristor as switch to insert proper combination of capacitive and/or inductive elements. They are not fully controlled devices because thyristor does not possess gate turn-off function as it has the ability to switch on but does not cut-off by itself. This category of devices includes SVC as shunt type and TSSC (Thyristor Switched Series Capacitors) as series type of compensators.

B. Static VAR Compensator (Static Shunt Compensators)

SVC is a shunt connected thyristor based compensator which yields reactive power i.e. exchanges capacitive current

or absorbs reactive power i.e. exchanges inductive current to keep certain parameters of the power system within defined range (usually bus voltage to which it is shunt connected) [16]. SVC is comprised of four basic devices: TSC (Thyristor Switched Capacitor), TCR (Thyristor Controlled Reactor), TSR (Thyristor Switched Reactor) & FC (Fixed Capacitor) and their desired combination. Typical configurations of SVC are:

- TSC-TCR type SVC
- TCR-FC type SVC
- TSC-TSR type SVC

C. Thyristor Controlled Series Capacitor (Static Series Compensators)

The basic arrangement of a TCSC is comprised of a compensating capacitor which is shunted by a TCR as shown in Fig. 3. In a practical application to achieve the required voltage rating and desirable operating characteristics, several such compensators are connected in cascade. It can be observed that if the reactance of reactor is very small it is equivalent to TSSC scheme. The presence of TCR in shunt with the capacitor provides the effect as that of a variable capacitor where TCR tends to partially cancel the compensating effect of capacitance. As TCR is equivalent to a variable reactance controlled by delay angle, the net impedance of TCSC in steady state is the parallel combination of X_L and X_C given as:

$$X_{TCSC} = \frac{X_C X_L}{X_C + X_L} \quad (11)$$

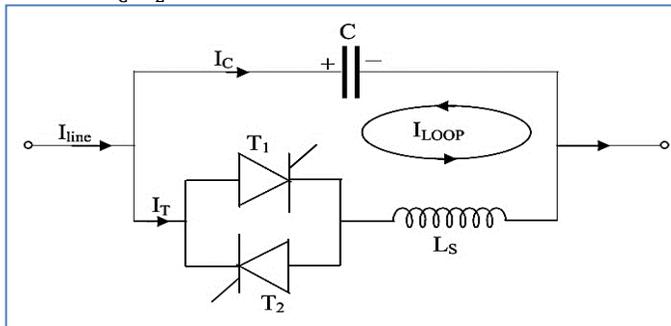


Fig. 3. Basic arrangement of TCSC.

D. Switching Converter Based FACTS Devices

The thyristor based FACTS devices discussed so far absorb or produce the governable reactive power by switching capacitor and reactor modules “in” and “out” of the system in a synchronous manner. The objective of this technique is to generate adjustable reactive impedance either in a continuous or in a discrete manner to compensate the transmission system to which these devices are connected. The generation of adjustable reactive power without involvement of capacitors or inductors is the basis for Converter based FACTS devices. This is achieved by incorporating Voltage source based and/or Current source based converters which operate by flowing ac current among the ac system phases. From the reactive power production viewpoint, these FACTS devices are analogous to synchronous machines which produce the reactive power by

controlling their excitation. Voltage source converters (VSC) are preferred over current source converters (CSC) because

- CSCs involve power semiconductors having two-way voltage blocking ability. The available electronic devices e.g. GTOs, IGBTs are either unable to impede reverse voltage or able to do it with higher conduction losses.
- CSCs are terminated at dc terminals with a reactor charged by current and hence have more losses than VSCs which are terminated by capacitor charged by voltage.
- The dc side termination of VSCs with a high rated dc capacitor furnishes automatic protection to power semiconductors against HV side system transients [16].

E. Static Series Synchronous Compensator (Voltage Source Based)

The SSSC is a cascaded connected voltage source based synchronous converter that is capable of varying the transmission line effective impedance by introducing a voltage having an adequate phase angle relation with line current. Depending upon this phase relation the SSSC can exchange both active/real and reactive power with the connected transmission system. For example, an in phase relation of voltage with the line current corresponds to active power exchange. On the contrary, if the fed voltage is in phase quadrature with the line current, this corresponds to absorbing or generating exchange of reactive power with the system. The SSSC is advantageous to TCSC due to its capability to regulate line reactance as well as line resistance during power swings, thereby providing increased damping for generators imparting power oscillations [20]. The SSSC consists of a multi-phase VSC as key component having dc-energy source and a coupling transformer in cascade with the line as shown in Fig. 4. The modes of operation are shown in Fig. 4(b)

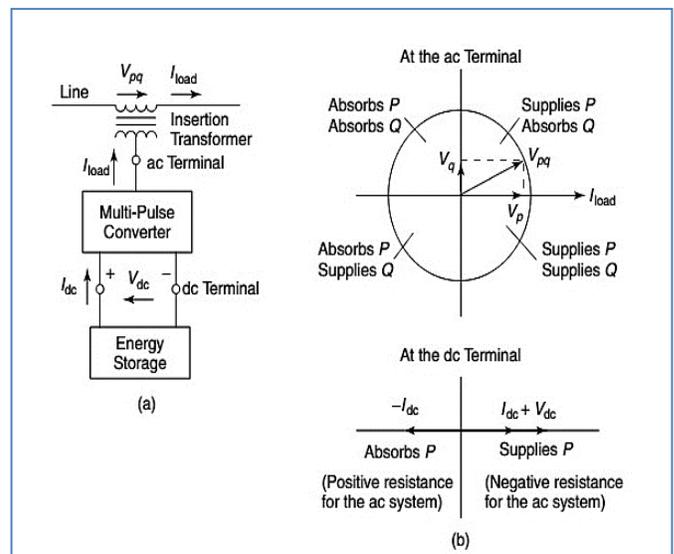


Fig. 4. (a) Basic Arrangement of SSSC, (b) Operating Modes of SSSC.

The principle of working of SSSC may be understood by considering a two machine system with sending and receiving

end voltages as V_S and V_R , respectively. The dropped voltage across line reactance is V_L and line current is I and if capacitive mode of SSSC is considered, V_C is the capacitor voltage as shown in Fig. 5. In this case the t/line inductance is recompensed by the SSSC by furnishing a voltage in lagging quadrature with line current. This voltage works to oppose the voltage across the t/line inductance that is in leading quadrature and the overall effect is equivalent to a reduction in the line inductance. The capacitor voltage may be expressed as

$$V_C = -kX I_L \quad (12)$$

Where k = degree of furnished series compensation and X implies series t/line inductance.

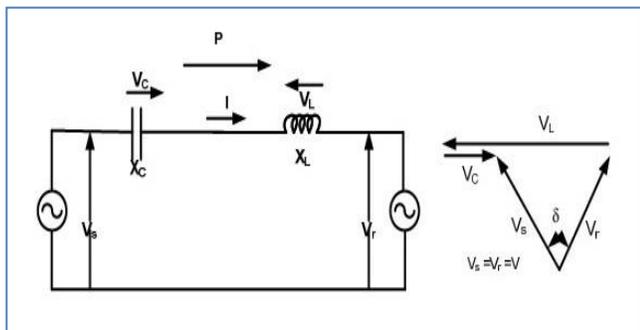


Fig. 5. A Two Machine system explaining SSSC Principle.

1) Static Synchronous Shunt Compensator (VSC Based)

The DSTATCOM is a reactive power compensatory device, shunt connected to the system, having the ability to generate and/or absorb the reactive power and output of which can be altered to control certain parameters of the power system to which it is coupled. Basically it is a solid-state converter being fed from an energy storage element at input terminals and has the ability to produce or absorb the governable reactive and real power at output ports. As explained before, VSC based DSTATCOM is preferred over CSC based compensators, which being fed from dc voltage of a dc capacitor, generates a set of 3- Φ ac voltages at the output, each voltage is in phase with the ac system to which DSTATCOM is coupled through a link reactance [20]. The key concept of DSTATCOM is that the connection of two AC sources, having same frequency, through a smaller cascaded inductance causes the active power flow from the leading ac source to the lagging one, while the reactive power flow is from the higher magnitude source to the one with smaller voltage magnitude. The flow of active power is influenced by the difference of phase angle between two ac sources while the flow of reactive power is dependent upon the difference of voltage between the two connected sources. Hence based on this concept, the DSTATCOM is capable of controlling the flow of reactive power by comparing the converter output voltage with the bus bar voltage of the system to which it is connected in shunt. DSTATCOM can be visualized as the static version of an ideal rotating synchronous condenser with no inertia, responding instantaneously to system changes and having the ability to produce reactive power without involving larger inductors or capacitor banks.

The DSTATCOM is named for its basic structural elements; First Static i.e. it does not involve any

moving/rotating parts and is based on static VSCs or CSCs, Second Synchronous i.e. produced voltages are in synchronism with the system and Third Compensator that expresses its compensating ability [21]. The single line circuit depicting basic configuration of DSTATCOM is shown in Fig. 6.

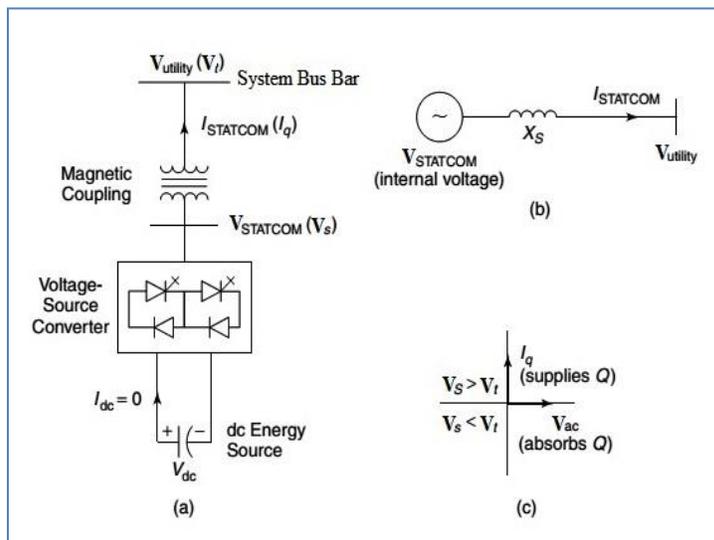


Fig. 6. DSTATCOM (a) basic configuration, (b) equivalent circuit, (c) concept of power exchange.

Fig. 6(a) depicts that VSC being fed from dc storage element (capacitor) is connected to the system bus via magnetic coupling (Coupling Transformer). Fig. 6(b) in the form of an equivalent circuit shows DSTATCOM as an adaptable voltage source with a reactance X_S appreciating the fact that shunt connected reactors and large capacitor banks are not involved for the compensation of reactive power which entails compactness and smaller footprint of the overall system. Fig. 6(c) describes that the reactive power transfer from or to the ac system bus can be monitored by altering the magnitude of 3- Φ output voltage V_S of DSTATCOM.

Capacitive Mode: If the DSTATCOM output voltage V_S is made higher than the utility system voltage V_t ($V_S > V_t$), it causes a leading current to flow from the DSTATCOM to the coupled ac system via link reactance (coupling transformer) and the DSTATCOM behaves as a source of capacitive reactive power.

Inductive Mode: On the contrary if V_S is made smaller than V_t ($V_S < V_t$), it results in a current flow from the coupled ac system to DSTATCOM via reactance and inductive reactive power is absorbed by the DSTATCOM.

Floating Mode: If the DSTATCOM output voltage is equal to utility ac system voltage ($V_S = V_t$) then DSTATCOM is in floating mode and no reactive power exchange follows.

The active power transfer/exchange between DSTATCOM and ac utility system can be monitored by adapting the phase shift between DSTATCOM output voltage and ac utility system voltage. If the DSTATCOM output voltage leads the system voltage, then DSTATCOM supplies active power to utility system. On the contrary, if its voltage is lagging behind

the utility system voltage then it absorbs the real/active power from its coupled ac system. The requirement of active/real power exchange arose due to the fact that the stored dc energy in the storage element (capacitor) may be used to overcome the VSC internal losses in the semiconductor switches and this is usually attained by making VSC output to lag behind the utility voltage by a smaller angle in the range of 0.1° – 0.2° [20]. It enables the converter to assimilate some real power from the coupled ac utility system thus maintaining capacitor voltage high enough to achieve the desired operation. The exchange of real and reactive power between utility ac system and DSTATCOM can be achieved irrespective of each other. If the DSTATCOM is provided with an appropriately rated energy storage element, any combination of reactive power assimilation or generation with active power assimilation or generation can be achieved as described in Fig. 7 and 8. This effective flexibility enables the utility systems to contrive efficacious control schemes to achieve improved transient and dynamic stability boundaries.

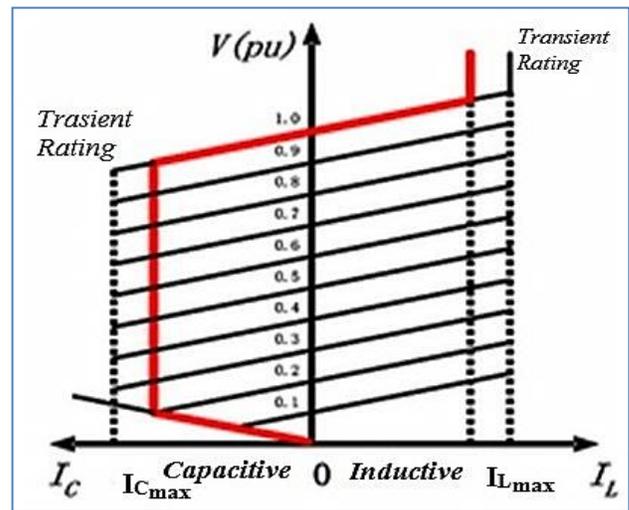


Fig. 8. Distinctive V-I Characteristics of DSTATCOM.

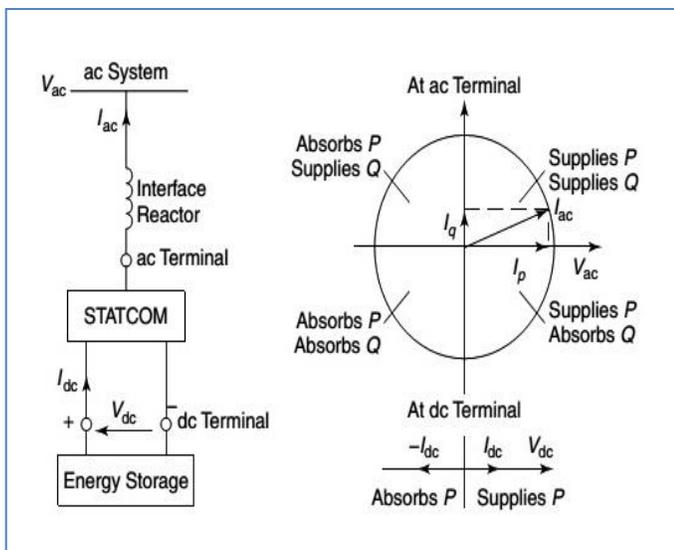


Fig. 7. Active, reactive power exchange b/w utility ac system and DSTATCOM.

The distinctive V-I characteristics of a typical DSTATCOM is described in Fig. 4 to 7 and 9. The characteristics reveal that the DSTATCOM has the ability to furnish capacitive as well as inductive compensation and is capable of controlling its output independently over the defined capacitive or inductive ranges regardless of the utility ac system voltage. It is claimed that the DSTATCOM can effectively furnish maximum capacitive VARs even at very low system voltages as 0.15 p.u [22]. The V-I characteristics also disclose the potential of DSTATCOM to generate maximum capacitive output irrespective of the system voltage showing its preference over SVCs. This characteristic is very desirable for the situations where DSTATCOM is employed to sustain the system voltage on the occurrence of faults and after faults; otherwise voltage collapse will be a limiting factor to the system performance [20].

This characteristic is very desirable for the applications where DSTATCOM is employed to sustain the system voltage on the occurrence of faults and after faults; otherwise voltage collapse will be a limiting factor to the system performance [20]. The characteristics also reveal the enhanced transient rating of DSTATCOM in both modes rendering it to be the most suitable option for compensation.

F. Why DSTATCOM?

The DSTATCOM is a preferable choice over other compensators due to its superior dynamic performance compared with others like SVC. The major distinguishing features of DSTATCOM that render it as the most suitable choice are:

- a) Quicker dynamic reaction to faults than other compensators
- b) Faster in ameliorating the transient response than others like SVC
- c) Capable of providing reactive as well as real power compensation
- d) Unlike SVC, the DSTATCOM can effectively furnish maximum capacitive VARs even at very low system voltages as 0.15 p.u.
- e) Unlike SVC, the DSTATCOM has the quality of controlling its generated current over maximum leading (capacitive) or lagging (inductive) range irrespective of coupled system voltage
- f) The generated current has low harmonics
- g) The DSTATCOM does not produce inrush currents
- h) The SVC behaves as *controllable reactive admittance* connected in shunt while the DSTATCOM functions as synchronous voltage source connected in shunt with the system.

These distinguishing features prove the DSTATCOM as the most favorable choice with greater flexibility and improved performance compared to other conventional options like SVC, etc.

V. VOLTAGE FAULTS IN STUDY

Different types of faults that might occur in a transmission line are briefly discussed below to gradually build up our problem scenario.

A. Voltage DIP

A voltage dip or sag is defined as an abnormality in the form of drop in the nominal rms voltage of a transmission system. The drop is normally between 10% and 90% of the total real power as shown in Fig.9. This drop usually occurs in time duration taken by one complete cycle of the AC power as a minimum up to a maximum of 1 minute. This fault occurs due to consumer side overloading of the transmission network by installations of medium voltage induction motors due to the fact that they draw about 9 to 10 times the nominal operating current of the motor at startup. As a result, power trips and failures occur. The following figure shows what a voltage sag looks like in the V-t axis.

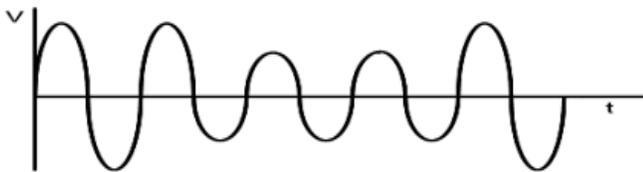


Fig. 9. Voltage sag.

The role of the DSTATCOM device is to provide the compensatory amount of power to the transmission line to act as an intermediate temporary solution to voltage sag and to avoid tripping of the transmission system.

B. Voltage Swell

A voltage swell is also a transmission fault due to the increase in the voltage provided by the transmission system to the network and consumers out of the tolerance values set on the line as shown in Fig.10. The time domain of such fault is usually some 1 to 3 seconds. A voltage swell is shown in the following figure:

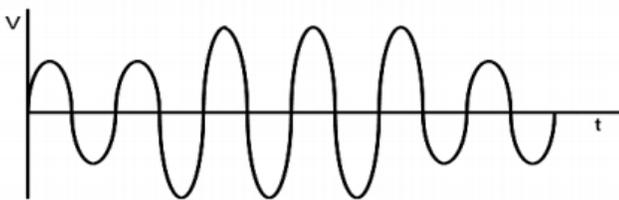


Fig. 10. Voltage swell.

C. Voltage Unbalance

In any three-phase voltage system, an unbalance is defined as a difference in the magnitude or phase angle for a large time. By large time we mean greater than 3 seconds. This is caused by large single-phase loads introduced to the network suddenly. A three phase voltage unbalance is harmful for all three phase machines especially induction motors due to the introduction of a negative sequence voltage. The waveform of an unbalance is shown in Fig. 11.

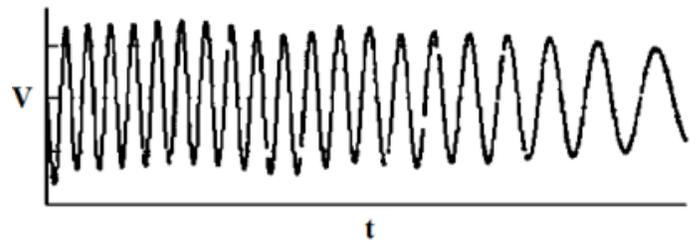


Fig. 11. A voltage unbalance.

There are other numerous faults in the T/L such as Interruptions, noise and transients but those are negligible as compared to the voltage sag, unbalance and swell as shown in Fig. 10 and 11. DSTATCOM is an ideal device for balancing and compensating these faults especially the power factor, neutral current elimination and current regulation. It regulates the voltage and improves power quality at PCC (Point of Common Coupling).

VI. SYSTEM MODEL

The theoretical and analytical concepts discussed so far in the previous sections paved the way to substantiate the expected outcomes. A model of T/line is implemented covering Peshawar City to Daud Kheil and then from Daud Kheil to Bannu, a total of 337km. The chosen line viz. 220KV T/line under the transmission system of NTDC (National Transmission & Dispatch Company Ltd.) is taken under study. A theoretical model of 100MVAR D-DSTATCOM after making certain changes to get it implemented on a 220KV system is accomplished herein. After carefully creating the overall model with different faults, detailed simulations will be performed in the next chapter to verify the expected outcomes.

1) Details of 220KVPeshawar-Bannu Road T/Line under Study

The technical details of 220KV T/L under the study employed for parameter settings used as reference and modeling of the system in MATLAB are as follows:

Length of Transmission line AB (Line under study) =189 km;

Length of Transmission line BC (for settings purpose) = 74 km

Length of Transmission line CD (for settings purpose) = 74 km

Conductor: Lynx

Capacity: 920 Amps

CCVT Ratio: 220kV/110V

C.T Ratio: 1200A/1A

K-Factor=CCVTRatio/CT Ratio= (220kV/110V) / (1200A/1A) = 1.67

Positive Sequence Impedance = $Z^+ = (0.21 + j1.24) \Omega/\text{km}$

Positive Sequence Impedance = $Z^+ = 1.26 / 80.437^\circ \Omega/\text{km}$

Positive Sequence Resistance = $R^+ = 0.21 \Omega/\text{km}$

Positive Sequence Inductance = $L^+ = 3.9 \text{ mH}/\text{km}$

Positive Sequence Capacitance = $C^+ = 2.5 \text{ nF}/\text{km}$

2) System Modeling in MATLAB/SIMUINK

The simplest model of the system under study is designed in Simulink 2016 and is shown below in Fig. 12. The model consists of a 220KV Peshawar to Daud Khel and then Bannu T/Line using Π -model with aforementioned technical details emanating from 500KV Grid Station NTDCL, ShahiBagh and terminating at 220KV Grid station at Peshawar City. The distance relay is located at the end near Bannu Grid Station for essentially detecting any faults whatsoever. The 111MVAR D-DSTATCOM is situated at the center of the line to compensate the voltage to maintain the system voltage by injecting or absorbing reactive power. The Fault Selector is used to involve various types of faults at different locations of the line to observe the behavior of impedance relay with and without DSTATCOM.

A Distribution Static Synchronous Compensator (D-DSTATCOM) is used to regulate voltage on a 222-kV distribution network as shown. Two feeders, namely, Peshawar to Daud Khel 189 km and Daud Khel to Bannu and 148 km) transmit power to loads connected at buses B2 and B3. The 220-V load connected to bus B3 through a 222kV/220V transformer represents a water motor drawing a large amount of power from the source. The load power factor stays at 0.9 for the setup.

The D-DSTATCOM regulates bus B3 voltage by absorbing or generating reactive power. This reactive power transfer is done through the leakage reactance of the coupling transformer by generating a secondary voltage in phase with the primary voltage (network side).

The D-DSTATCOM consists of the following components:

- Voltage Source Converter,
- Energy Storage,
- L-C Passive Filter,
- Control Block, and
- Coupling Transformer.

3) Voltage Source Model

A VSC converts the DC voltage across storage device into a three phase AC output voltage wave. It can be a 3-phase-3-wire or 3-phase-4-wire. The VSC is shown in Fig. 12.

4) Energy Storage

The DC Storage is essentially a capacitor of 10,000 Farads value. Two level conversions are used to convert the DC source to an equivalent 3 phase AC voltage.

5) LC Passive filter

LC filters as shown below are utilized to compensate for the harmonic distortion generated on the line as shown in Fig. 13. The use is necessary for matching the output impedance of the line.

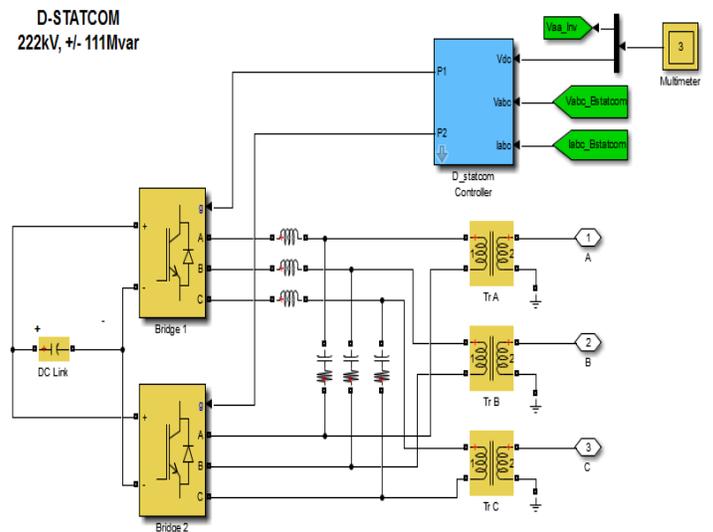


Fig. 12. VSC D-DSTATCOM, 2-level conversion using IGBTs.

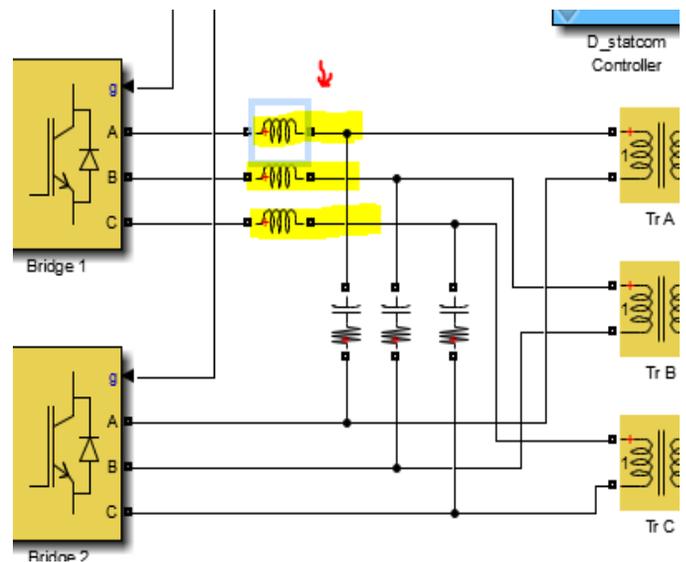


Fig. 13. RL Branches for harmonics distortion compensation.

6) Coupling Transformer

The output from the D-DSTATCOM is linked with that of the main lines via coupling transformers.

7) Control Block

The control block detects faults, voltage sags and swells, generates trigger pulses for the PWM inverter on occurrence of faults and stops them after event of fault has cleared. The control block is designed on the basis of DQO theory.

D-DSTATCOM working is illustrated in Fig. 14.

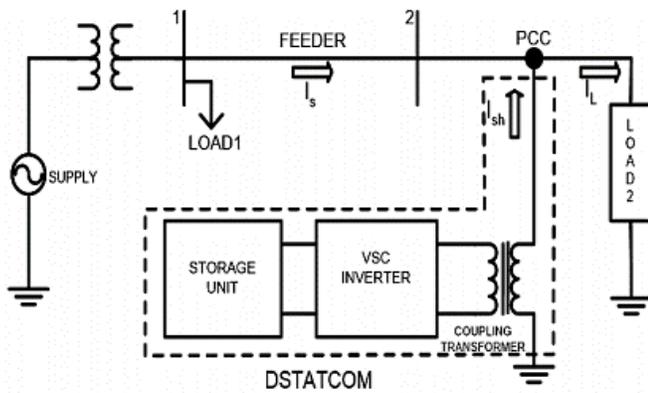


Fig. 14. D-DSTATCOM working illustration.

The voltage generated is injected at the targeted T/L according to the fault type so that the load supply is not affected.

VII. SIMULATION RESULTS AND DISCUSSION

After detailed modeling of the system in MatLab/Simulink under study in the previous chapter, this chapter presents the results obtained after comprehensive simulations by introducing various faults such as sag and unbalance when an induction motor is started, and evaluating the performance of T/line protection with and in the absence of D-DSTATCOM. The simulations are carried out by creating faults at a specified location in the transmission line and the follow up of the DSTATCOM is observed. Based on the simulation results, conclusions about deviations in the performance characteristics of T/Line protection with mid-point compensating D-DSTATCOM and other important results are presented. By analyzing these deviations, the recommendations to improve the discrepancies are discussed as closing remarks.

A. Simulation Results With 3- ϕ Faults

Consider the modeled system, a 3- Φ fault is created after 0.2 seconds at various locations along the T/line length and behavior of distance protection is observed for the cases with and in the absence of DSTATCOM at the mid-point. The voltage and current signals before and after fault are shown in Fig. 5. The fault is introduced at 60% (202km) of total line length (337km).

B. Results

Fault has been introduced at start of simulation. The fault is modeled to simulate the behavior of a medium Voltage industrial induction motor being started. As the motor is expected to withdraw a large amount of power, the waveform of the transmission line voltage is introduced with a sag and then an unbalance. The behavior of the system can be seen as follows:

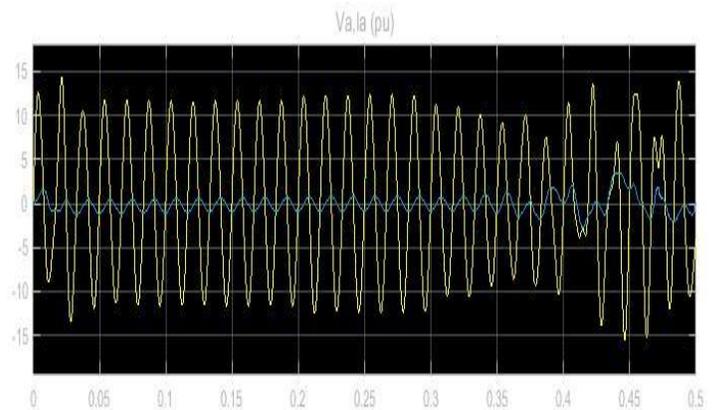


Fig. 15. Va per unit of the 3-phase T/L with current in Blue color.

In Fig. 15, we see the Voltage waveform of the transmission lines when no fault is introduced on it. It is a normal sinusoidal waveform in harmony with the current transmission.

Next a voltage sag is introduced in the system. Due to a single-phase inductive load with high current withdrawal, the waveform becomes distorted and a sag at 0.05 seconds is introduced as can be observed in the following figure (Fig. 16):

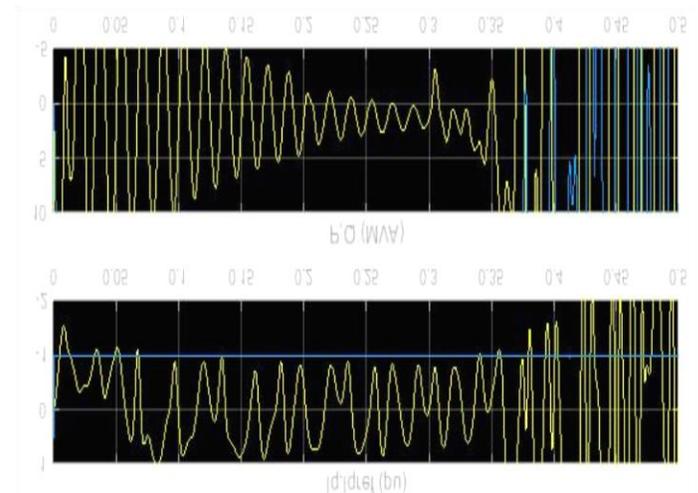


Fig. 16. Voltage Sag as a result of Inductive load.

Also observe that in this figure, we have taken current as unavailable or scarce, and the compensator has to complete the requirement for the given loads in the form of reactive power and current.

The reactive compensated power introduced by the DSTATCOM in the form of Voltage can be seen in Fig. 17.

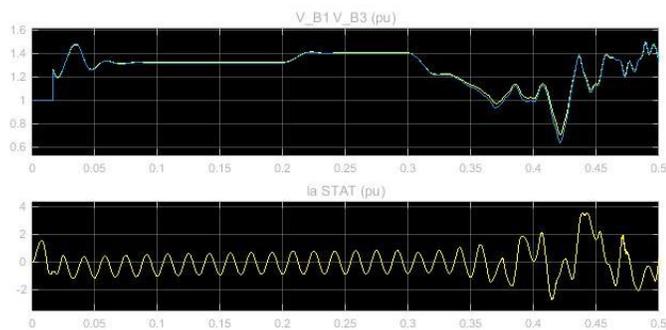


Fig. 17. Injected voltage and current by DSTATCOM for compensation.

Here when the system is attacked by Sag, the Compensator reacts precisely at 0.05 seconds with a few microseconds delay and produces a high amplitude voltage to balance the fault. From Fig. 16, we can see that the fault was introduced for 2 cycles precisely, and as a result, the DSTATCOM can be seen from Fig. 3 to 5 to compensate the fault for 2 cycles only. In the third cycle, it stabilizes the voltage cycle and goes to a standby phase for 8 cycles.

At 0.4 seconds we have introduced unbalance for 0.2 seconds, (Fig. 16) and the compensator can be observed to apply appropriate compensation in the form of a higher amplitude voltage for that cycle after which it again stabilizes.

In the 16th Cycle of the waveform due to the voltage cycles getting out of frequency synchronization with the source comparing signal, the compensator slightly modifies the reactive waveform in the same cycle for one cycle only and then goes to a stable form (idle form).

Similarly, in the reactive power waveform, we can see the same behavior where the reactive power (in blue- lower line in graph) and real power (in yellow- upper line in graph) is shown to be correspondingly relateable.

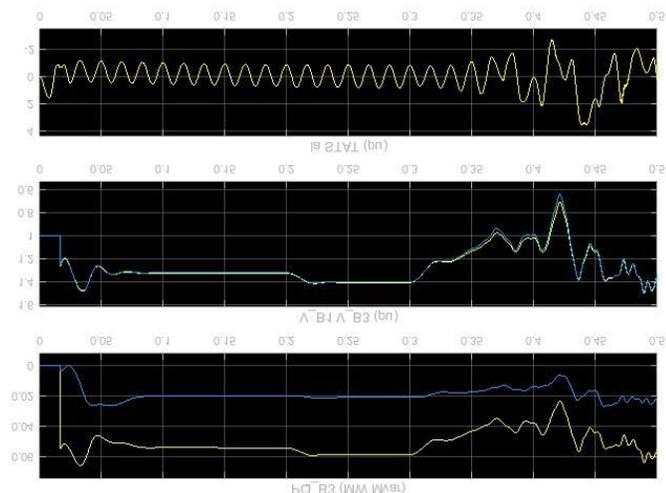


Fig. 18. D-STATCOM compensating for the injected fault.

In Fig. 18, we can see that initially at the introduction of fault, the power goes to a large amount (0.025 MW or 25KW) and the DSATCOM has to swing in to provide for the

remaining amount of required power i.e. surplus on 22KW (as designed in the simulator). The compensator completes the requirement by providing the surplus amount of power 0.009 MVar to the load in the first 0.4 seconds. After that, the induction machine is started and the required power drops lower whereas in compensation, the reactive power production also drops lower. The graphs can be seen to be coherent throughout the 2 seconds of sag introduction and unbalance. The current output of the DSTATCOM can be seen below where at the start of the cycle, when the motor required a large amount of current and we have decreased the available current in the transmission line below 0, the DSTATCOM device is fully operational and provides for the current compensation in each cycle of the waveform both in start of the induction motor as well as during loaded operation as seen in Fig. 19 below:

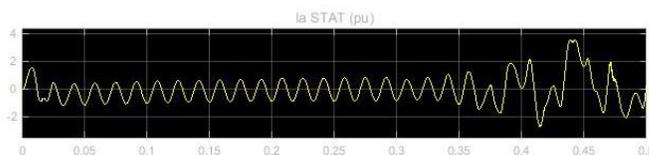


Fig. 19. Current output of DSTATCOM in fault event

Fig. 19 shows the current output of the DSTATCOM in the event of a fault.

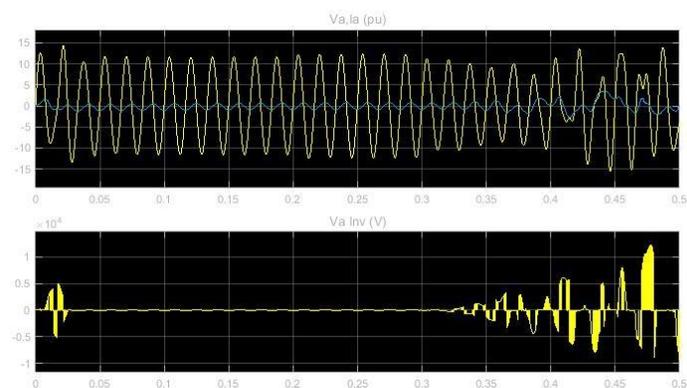


Fig. 20. Power Inverter output for compensating the voltage sags.

The output of the inverter in Fig. 20 shows that the reactive power output from the DSTATCOM has been injected in to the T/L for correcting the fault.

VIII. CONCLSIONS AND FUTURE WORK

Analytical and simulation results lead to following conclusions:

During Unsymmetrical faults, percentage increase in impedance is more due to symmetrical compensation furnished by DSTATCOM (due to its fast responsive settings) to each phase rather to faulty phase only. This contributes to over compensation. The 3- Φ concurrent control strategy for DSTATCOM is desirable because

- This mode contributes to best VA consumption of the converter.

- Harmonic generation is the lowest in this mode.
- Response time is immediate [13].

The major parameters affecting the performance of relay are:

- Fault Location
- Distance of fault after location of compensating device
- Level of compensation furnished by DSTATCOM
- Type of fault.

Simulation also shows that the DSTATCOM is an effective device in compensating for the various types of faults in the transmission line.

The system cannot compensate 100% of the voltage during sag, but the compensation values are in nominal ranges and are best for practical installations and deployments.

The current implemented scheme of Distance Protection for 220KV Peshawar to Bannu T/L is "Basic Distance Protection" which must be revised before the proposed installation of DSTATCOM.

The scheme may be revised to Communication Aided Distance Protection.

Due to the use of IGBTs in the Distribution Static Compensators in the Dynamic Application, there are bound to be many harmonics and distortions, hence a suitable type of filter can be used to filter the noise. For this purpose, LCL filter can be used and introduced in the control structure of the DSTATCOM for further research.

Since DSTATCOM devices use PWM signals, a better alternative can be SWPWM (Sine Wave PWM) where the efficiency of the converted voltage and its power factor increases and a more quality voltage signal with less distortions and harmonics is produced. This will affect the devices performance in a positive way.

ACKNOWLEDGEMENTS

Authors would like to acknowledge the financial support provided by the CECOS University of IT and Emerging Sciences Peshawar Pakistan.

REFERENCES

- [1] Mr. Ajaysing, T. Chandan, K. Venkata Rama Mohan, Santhosh Kampelli, Arvind R. Singh, "Advance Distance Protection of Transmission Line in Presence of Shunt Compensator" International Research Journal of Engineering and Technology (IRJET) June 2015.
- [2] Z. Moravey, M. Pazoki and M. Khederazahe, "Impact of UPFC on Power Swing Characteristic and Distance relay Behavior" IEEE Transactions on Power Delivery, Vol. 29, No. 1, February 2014.
- [3] Naresh Patmana, Swamynadha Sree Venkata Ramana Chakkirala, "Effect of FACTS Controllers on Impedance Relay Characteristics" International Conference on Engineering Trends and Science & Humanities (ICETSH) 2015.

- [4] Mohamed Elsamaly, Sherif Omar Faried and Tarlochan Sidhu, "Impact of Midpoint DSTATCOM on Generator Loss of Excitation Protection" IEEE TRANSACTIONS ON POWER DELIVERY, VOL 29, No. 2, APRIL 2014.
- [5] Danna Hemasunder, Mohan Thakre, V.S Kale, "Impact of DSTATCOM on Distance Relay-Modeling and Simulation using PSCAD/EMTDC" IEEE Students Conference on Electrical, Electronics and Computer Sciences 2014.
- [6] Gorakshanath Abande, M.F.A.R Satarkar, Mohan Thakre, Dr. V.S Kale, Ganesh Patil, "Impact Analysis of DSTATCOM on Distance Relay" International Journal of Innovative Research in Science, Engineering and Technology, Volume 3, Special Issue 3, March 2014.
- [7] Mohammed Zellague, "A Comparative Study of Impact Series FACTS Devices on Distance Relay in 400KV Transmission Line" Journal of Electrical and Electronics Engineering (JEEE).
- [8] Mohammed Zellague, Abdelaziz Chaghi, "Impact of TCSC on Distance Protection Setting Based Modified Particle Swarm Optimization Technique" IJISA Vol. 5, May 2013.
- [9] Gaber El-Saady, Rashad M. Kamel, Essam M. Ali, "Error Analysis in Distance Relay Readings with presence of FACTS Devices" Innovative System Design and Engineering, Vol. 4, No. 14, 2013.
- [10] Mohammed Zellague, Abdelaziz Chaghi, "Impact of Apparent Reactance injected by TCSR on Distance Relay in Presence of Phase to Earth Fault" Power Engineering and Electrical Engineering Volume 11, Number 3, 2013.
- [11] Mojtaba Khaderzadeh and Amir Gharbani, "Impact of VSC-Based Multiline FACTS Controllers on Distance Protection of Transmission Line"
- [12] Sankara Subramanian, Anthony Perks, Sarath B Tennakoon, Noel Shammass, "Protection Issues Associated With The Proliferation Of Static Synchronous Compensator (DSTATCOM) Type Facts Devices In Power Systems".
- [13] Wen-Hao Zhang, Seung-Jae Lee, Myeon-Song Choi and Shigeto Oda, "Considerations on Distance Relay Setting for Transmission Line with DSTATCOM" IEEE 2010.
- [14] Mohammad Javad Farah, Behrooz Vahidi, Hossein Askarian Abyaneh, "Effect of FACTS Devices on Measured Impedance by Distance Relay" SINTE 8, 2013.
- [15] Yongan Deng, "Reactive Power Compensation of Transmission Lines" Concordia University.
- [16] Narain G. Hingorani, Laszlo Gyugyi, "Understanding FACTS" IEEE Power Engineering Society.
- [17] Sanaullah Ahmad, Sana Sardar, Azzam ul Assar, Fazal Wahab Karam, "Reliability Analysis of Distribution System using ETAP" International Journal of Computer Science and Information Security 15.3 (2017).
- [18] Xunchi Wu, "Reactive Power Compensation Based on FACTS Devices" Columbia University.
- [19] Aamir Aman, Sanaullah Ahmad, Khalid Mahmood, Designing and Strategic Cost Estimation of Stand-Alone Hybrid Renewable Energy System, 4th International Conference on Energy, Environment and Sustainable Development 2016.
- [20] R. Mohan Mathur, Rajiv K. Varma, "Thyristor Based FACTS Controllers for Electrical Transmission Systems" IEEE Press.
- [21] Tariq Masood, R.K. Aggarwal, S.A. Qureshi, R.A.J Khan, "DSTATCOM Model against SVC Control Model Performance Analysis Techniques".
- [22] Festo Didactic Canada, Courseware sample on "Static Synchronous Compensator (DSTATCOM)" Library and Archives Canada 12/2014.

Synchronous Authentication Key Management Scheme for Inter-eNB Handover over LTE Networks

Shadi Nashwan

Computer Science and Information Department
Aljouf University
Aljouf, Saudi Arabia

Abstract—Handover process execution without active session termination is considered one of the most important attribute in the Long Term Evolution (LTE) networks. Unfortunately, this service always is suffered from the growing of security threats. In the Inter-eNB handover, an attacker may exploit these threats to violate the user privacy and desynchronize the handover entities. Therefore, the authentication is the main challenge in such issue. This paper proposes a synchronous authentication scheme to enhance the security level of key management during Inter-eNB handover process in LTE networks. The security analysis proves that the proposed scheme is secure against the current security drawbacks with perfect backward/forward secrecy. Furthermore, the performance analysis in terms of operations cost of authentication and bandwidth overhead demonstrates that the proposed scheme achieves high level of security with desirable efficiency.

Keywords—LTE network; X2 handover; horizontal and vertical key derivations; desynchronizing attack

I. INTRODUCTION

In order to enhance the quality of service (QoS) with higher data rate in third generation (3G) networks, the Third generation partnership project (3GPP) has been developed the LTE network [1]. Therefore, the network architecture has been restructured to provide sufficient services by increasing bandwidth, enhancing performance, supporting heterogeneous connections with the other IP-technology and enhancing security level [4].

The main components of the LTE network architecture can be summarized as the following. The User Equipment (UE) connects to the Evolved Packet Core (EPC) through the Evolved Universal Terrestrial Radio Access Network (E-UTRAN). The latter component includes a set of Evolved NodeB (eNBs). The eNB is a base station that modulates and demodulates the signals to perform the radio communications between the UEs and EPC. The latter includes the Home Subscriber Serve (HSS), Mobility Management Entity (MME), Serving Gateway (S-GW), Packet Data Network Gateway (P-GW), Authentication Center (AuC) and Policy Charging Rules Function (PCRF) [7].

Data is transmitted between the eNBs and the P-GW through the S-GW. The P-GW connects the network with the outside IP networks. The PCRF recognizes the policies of QoS and the network resources. The HSS contains the AuC to fetch the user identifier and the pre-loaded shared key as well as to perform the key derivation functions during the authentication

sessions. The MME interacts with HSS for user authentication and mobility management.

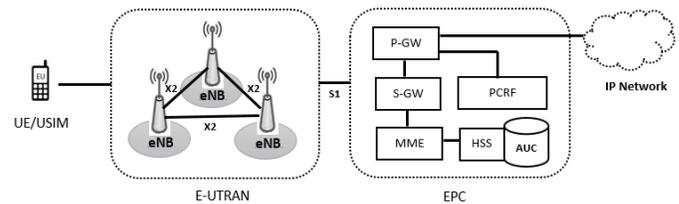


Fig. 1. LTE Network architecture.

The S1 interface connects the eNBs with the MME while the eNBs communicate with each other through X2 interface. Fig. 1 shows the LTE network architecture.

The improvement of mobility management is an essential process in LTE networks especially in the handover service which is requested by subscriber more frequently than other services in LTE networks. The security service is considered the main critical section in such improvement. This paper concentrates on the security drawbacks of the Inter-eNB handover and key management of LTE networks during the handover process.

When the UE moves away from the Serving eNB, the handover process should be performed to connect the UE with the Target eNB during the active session without service termination. In the LTE network, the air interface includes two different handover types, the X2 handover and S1 handover [5].

In the Inter-eNB handover (called X2 handover), both of the Serving eNB and the Target eNB are connected directly though the X2 interface. However, if the X2 interface does not exist between the Serving eNB and the Target eNB, or the Serving eNB initiates the handover process towards a particular Target eNB via the S1 interface, the S1 handover will be executed. In the S1 handover, both of the Serving eNB and the Target eNB are connected indirectly though the MME over the S1 interface.

Considering the Inter-handover, the Serving eNB sends the authentication parameters with session key to the Target eNB though the X2 interface directly, the mutual authentication does not exist between the Serving eNB and Target eNB which will be vulnerable to be attacked [16]. The UE exchanges the authentication parameters with the Serving eNB and Target eNB as clear text. Therefore, an adversary easily can catch

these parameters. This is open the door for several drawbacks, an adversary can masquerade as a legitimate eNB to send an authentication messages by utilizing valid identities and authentication parameters, this is known as a rogue base station attack [2], [22]. Moreover, the MME provides the Serving eNB through S1 interface the recent parameters as clear text to generate a new session key to perform the handover process with the UE [3]. Subsequently, once the adversary acquires these parameters, the adversary using a rogue base station can disrupt and modify the refresh values of the authentication parameters, this is known as the desynchronizing attack [6], [10].

The handover process should be more secured against the current drawbacks in LTE networks. Therefore, the authentication is an important part in the handover process. Considering these security weaknesses, this paper proposes a synchronous authentication scheme to enhance the security level of key management during Inter-handover process in LTE networks.

The proposed scheme can overcome the existing drawbacks such as a rogue base station attacks, desynchronization attacks, replay attack and redirection attacks. Furthermore, the performance analysis in terms of operations cost of authentication and bandwidth overhead demonstrates that the proposed scheme achieves high level of security with desirable efficiency comparing with existing handover key management schemes.

This paper is organized as follows: Section 2 reviews the current handover authentication scheme. In Section 3, the related works is discussed. The proposed scheme is introduced in Section 4. The security and performance analysis of the proposed scheme are demonstrated in Sections 5 and 6, respectively. Finally, this paper will be concluded in Section 7.

II. LTE HANDOVER AUTHENTICATION SCHEME

In this section, the Key hierarchy of LTE networks is illustrated, then the X2 handover process is reviewed. Finally, the security drawbacks of the X2 handover are discussed.

A. Key Hierarchy in the LTE Network

To minimize the security threats, the design consideration of the LTE networks not just separates between signaling and user data traffic but also separates the key management for encryption, integrity and handover protection [1], [8], [13].

TABLE I. KEY HIERARCHY OF THE AKA PROTOCOL OF LTE NETWORK

Authentication entities	Keys
UE, HSS	root key K
UE, HSS	CK, IK
UE, MME, HSS	Local root key KASME
UE, eNB, MME	KeNB
UE, source eNB, target eNB	KeNB*

Table 1 illustrates the key hierarchy of the Extended Authentication and Key Agreement protocol (EPS-AKA) that is deployed in the LTE network [11], [18] and [19]. The root key (K) is used by the UE and the HSS to derive both of the Cipher key (CK) key and the Integrity key (IK) key. When the mutual authentication between the UE and the HSS is

completed, both of the UE and the HSS derive the local root key (KASME) by binding the CK, IK with MME identity to the key derivation function (KDF) function, then HSS forwards the KASME to the MME. Furthermore, the (KeNB) key is derived key from KASME key by the UE and the MME, then the MME sends the KeNB to the eNB. The KeNB key is specified to encrypt the traffic between the UE and the eNB. Finally, based on the KDF function, the KeNB is used by the UE and the eNB to derive KeNB*, the source eNB forwards the KeNB* to the target eNB during the handover process.

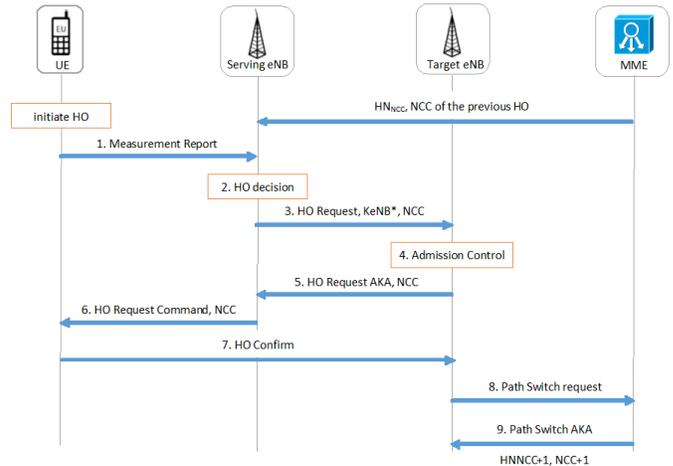


Fig. 2. X2 handover process.

B. X2 handover process

In LTE network, when a UE moves away from the Serving eNB, the handover process should be performed to connect the UE with the Target eNB without interrupting the active session. In the X2 handover, the Serving eNB sends the KeNB* to the Target eNB [2], [20], [21]. This process includes several steps that are shown in Fig. 2.

To initiate the handover process, the UE sends a measurement report to the Serving eNB which includes the information that is related to the neighboring eNBs to specify the Target eNB (Step 1). The Serving eNB analyses the measurement report to decide if the handover is necessary and to choose the best Target eNB. The Serving eNB derives the KeNB*, then transmits the handover request message with the KeNB* and Next Hop Chaining Counter (NCC) value to the Target eNB through X2 interface (Steps 2 and 3). In order to ensure if the resources is available to serve new UE, the Target eNB performs the admission control process, then sends the handover request acknowledgement to the Serving eNB, this message includes the NCC parameter to connect the UE with the target eNB (Steps 4 and 5).

NH0 = initial value of the KeNB

NH1= KDF (KASME, HH0)

NH2= KDF (KASME,NH1)

then

$$\text{NHNCC} = \text{KDF} (\text{KASME}, \text{NHNCC}-1) \quad (1)$$

$$\text{KeNB}^* = \text{KDF} (\text{NHNCC}, \text{PCI}, \text{EARFCN-DL}) \quad (2)$$

$$\text{KeNB}^* = \text{KDF}(\text{KeNB}, \text{PCI}, \text{EARFCN-DL}) \quad (3)$$

The serving eNB sends the handover request command to the UE with the NCC that has been transmitted from the Target, after that the UE sends a confirmation message back to the Target eNB as a new serving eNB (Steps 6 and 7).

To achieve the forward key secrecy during the handover process, the Next Hop key (NH) value can be derived using the KDF function as defined in (1). The Serving eNB has fresh values of NHNCC key and NCC that have been sent from the MME during the previous handover session, the value of NHNCC means that the NH key is refreshed NCC times [1], [12]. Using the Physical Cell Identity (PCI) and E-UTRAN Absolute Radio Frequency Channel Number on the Download (EARFCN-DL), the UE and Serving eNB can derive the KeNB* from the NHNCC or from current KeNB as defined in (2) and (3), respectively.

Subsequently, the UE verifies the value of NCC that have been received from the Target eNB. In case, the received NCC is matched with the current NCC that is association with the previous handover session (i.e., NCC-1), then the UE derives the KeNB* using vertical key derivation as defined in (2), where a new value of NHNCC is derived from the previous value of NHNCC-1 and KASME key as defined in (1).

In case, the received NCC is greater than the current NCC that is association with the previous handover session, the UE using the current value of KeNB performs the horizontal key derivation to derive the KeNB* as defined in (3).

The new serving eNB (Target eNB) sends the S1 path switch request message to the MME through S1 interface (Step 8). Upon receiving the path switch request, the MME derives the fresh NH key and NCC values, then the MME sends S1 path switch request acknowledgement message back to the new Serving eNB, this message includes the NHNCC+1 and NCC+1 next handover (Step 9).

C. Security Drawbacks of the X2 Handover

In despite of the key hierarchy system in LTE network performs more security level by supporting the backward/forward key separation features. The current X2 handover process is suffered from different drawbacks. The session keys and handover parameters are exchanged between handover entities as clear text without protection. The UE and the Serving eNB does not authenticate by the Target, and the user identity is exchanged between the handover entities without concealing.

In order to catch and modify the authentication messages that are exchanged between the handover entities, an adversary can use a rogue base station to masquerade as a legitimate eNB [21]. Subsequently, an adversary can forward modified NCC values between the handover entities by utilizing valid identities.

Therefore, an adversary can leave the Target eNB desynchronized and the session keys of the next handover processes vulnerable to compromise, then the adversary can decrypt all messages between the UE and eNB, this is known as the desynchronizing attack [10], [20].

When the NCC that sent from the MME is modified, an adversary forces the Target eNB to drive the KeNB* based on the current KeNB* using the horizontal key derivation. In the same manner, when the adversary changes the NCC value that sent to the Target eNB from the Serving eNB to be extremely larger than the original NCC value, the KeNB* will be derived using the horizontal key derivation. Consequently, forward key separation feature is disrupted and the future sessions keys of next hops will be compromised until the KASME key is recomputed during the next EPS-AKA execution.

III. RELATED WORK

There have been many researches on the authentication handover scheme of LTE networks.

In 2014, Han and Choi [10] propose a scheme to overcome the desynchronization attack of the handover process in the LTE network. An algorithm to derive the key based on specific minimal interval time has introduced. However, the scheme does not prevent the desynchronization attack and the communication overheads have increased.

Haddad et al. [9] introduce a secure and efficient handover scheme for the LTE-Advanced (LTE-A) network. The scheme classifies the eNB into two types, the eNB that is operated by the subscriber and the eNB that is operated by the network provider. The authors demonstrate different handover scenarios using uniform authentication scheme to thwart well-known attacks. The proposed scheme uses asymmetric key technique to perform the authentication between the communication entities rather than the symmetric key technique that is used in the current used scheme. However, it also cannot provide enough security.

In 2015, Lin et al. [15] pointed out that the X2 handover mechanism of LTE network has some security drawbacks. The first drawback is that the source eNB and UE are not authenticated by the target eNB. The second is that the attacker can modify the NCC and cause the desynchronizing attack. To overcome these vulnerabilities, Lin et al propose a scheme based on pre-loaded shared group key between all eNBs and MME. The scheme, however, does not resolve the current drawback issues in defeating desynchronization attack, replay attack and redirection attack.

In 2016, Khairy et al. [12] propose a new authenticated key management scheme for intra-MME handover. Hence, the MME is used as a third party and the source eNB is keeping out from the key management process to overcome the desynchronization attack. The scheme uses the pre-shared key for each eNB to protect the handover parameters between the eNBs and the MME. Hence, the mutual authentication between handover entities is partially achieved. Unfortunately, the proposed scheme increases the communication overheads of handover process.

Mathi and Dharuman [17] design a scheme to prevent the desynchronization attack due to rogue base station in handover key management of 4G LTE network. The proposed scheme generates a new key for future communication between the Target eNB and UE after the Target eNB is verified by the MME. However, similar to current handover key management scheme, the proposed scheme is not suitable to protect the

handover process due to lack of the backward/forward keys separation and mutual authentication between handover entities.

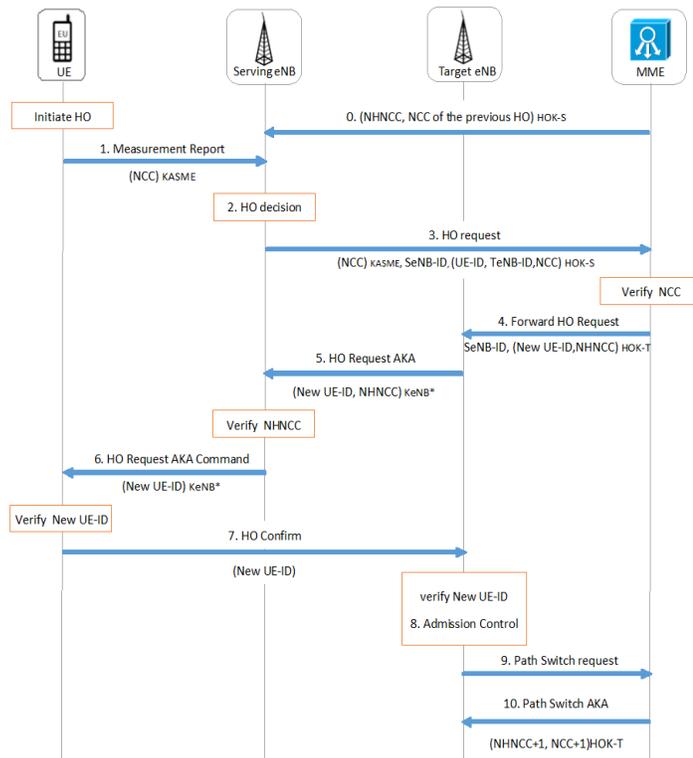


Fig. 3. Proposed X2 handover process.

IV. PROPOSED SCHEME

This section introduces a synchronous authentication key management scheme for Inter-handover over LTE networks. In the current handover key management scheme, an attacker can catch the KeNB key that is sent from the MME to the Serving eNB, then an attacker forces the handover entities to derive the new session key KeNB* using the horizontal key derivation based on the current session key (KeNB). Therefore, the forward Keys separation feature is disrupted.

One of the main goals of the proposed scheme is to keep the forward Key separation feature by continuous the synchronization between all the handover entities. Therefore, just the vertical key derivation will be used with a fresh NHNCC key value to derive KeNB*. The pre-loaded shared key for each eNB with MME (HOK) is used in the proposed scheme to protect the handover parameters that are exchanged between MME and eNBs. Fig. 3 illustrates the proposed scheme according to the following.

The UE initiates the handover process by sending the measurement report to the Serving eNB, this message contains the encrypted NCC value using KASME key (Step 1). Through measurement report, the Serving eNB decides that the handover is necessary or not. The Serving eNB chooses the best Target eNB according to the measurement report (Step 2).

The Serving eNB sends the handover request message to the MME through S1 interface for performing the handover process. The message contains the encrypted NCC that has

been sent by the UE along with the identity number of the Serving eNB (SeNB-ID). The message also contains a set of encrypted authentication parameters using the pre-loaded key of Serving eNB (HOK-S). The encrypted authentication parameters includes the identity number of the UE (UE-ID), identity number of the Target eNB (TeNB-ID) and the NCC value that has been received by the Serving eNB from the MME during last handover session (Step 3).

Upon receiving the handover request message, the MME decrypts the UE-ID, TeNB-ID and NCC that have been added by the Serving eNB, then fetches the KASME key of the UE to decrypt NCC value that has been sent from the UE through the serving eNB. In case, both NCC values are not equal, the handover request will be rejected by the MME. Otherwise, MME fetches the NHNCC that is associated with the NCC value, then calculates the new UE-ID. According to the TeNB-ID, the MME fetches the pre-loaded key of the Target (HOK-T) to encrypt the NHNCC and the new UE-ID. After that, MME forwards the handover request message along with the encrypted values of NHNCC and the new UE-ID over S1 interface to Target eNB (Step 4).

Upon receiving the handover request message, the Target eNB checks that it has a resource for the handover process. The Target eNB decrypts the NHNCC and new UE-ID, then derives the new KeNB* from the received NHNCC as defined in (2). The Target eNB sends the Handover request acknowledgement to the Serving eNB over X2 interface. This acknowledgement includes the encrypted value of the NHNCC and the new UE-ID using the new session key KeNB* (Step 5).

The Serving eNB decrypts the NHNCC using the KeNB*, then authenticates the acknowledgement message by comparing the NHNCC value that has been received from the MME with the NHNCC value that has been received from the Target eNB. In case, both values are not equal, then the Serving rejects the handover process. Otherwise, the Serving eNB sends the handover request acknowledgement command message to the UE along with the encrypted value of a new UE-ID (Step 6).

The UE derives the new session key KeNB* as defined in (2) to decrypt the new UE-ID that has been received from the Target eNB. In order to verify the new UE-ID, the UE calculates the new UE-ID in the same way in the MME. The UE compares between both values, if are not equal, then UE rejects the handover process, else sends handover confirmation message along with the new UE-ID over X2 interface to the Target eNB. This message announces that the handover process has been successful (Step 7).

Upon receiving the handover confirmation message, the Target eNB compares the new UE-ID that has been received from UE with the decrypted UE-ID value that has been received from MME, if the both values are not equal, then the Target rejects the handover process, else the Target eNB is becoming the new Serving eNB (Step 8). Through this step, the Target eNB authenticates the UE and indirectly authenticates the Serving eNB and the MME. After admission control process, the Target eNB sends the S1 Path Switch Request to the MME over S1 interface to the MME. Through this message, the new Serving eNB notifies the MME to change the

UE location and requests the switch path towards the new Serving eNB (Step 9).

Upon receiving the path switch request, the MME calculates the fresh values NHNC and NCC, then the MME sends S1 path switch request acknowledgement message back to the new Serving eNB, this message includes the encrypted fresh values of the (NHNCC+1) that is computed as defined in (1) and NCC+1 by the HOK-T of next handover (Step 10).

V. SECURITY ANALYSIS

In this part, the security analysis is conducted to demonstrate that the proposed scheme has attractive security features during the Inter-handover process. In addition to, explain how the enhancements of the proposed scheme can resist the current drawbacks.

The same security architecture of the current handover key management scheme is used in the proposed scheme. Moreover, the proposed scheme uses the same authentication parameters and functions to enhance the security level of the handover process to be more secure against the current drawbacks as the follows:

A. Mutual Authentication

The mutual authentication feature can be performed between all handover entities in the proposed scheme during the Inter-handover process. More precisely, the MME authenticates the UE and the Serving eNB by verifying the NCC values that have been sent from the UE and the Serving eNB.

Indirectly, the Target eNB authenticates the MME, Serving eNB and UE by verifying the new UE-ID values that has been sent through the serving eNB and has been computed by the MME and UE. The Serving eNB can authenticate both of the MME and Target eNB by comparing the NHNCC value that has been sent from the MME in previous handover session with the NHNCC value that has been sent from the Target eNB. In the same manner, the UE authenticates the MME, Serving eNB and Target eNB by comparing the computed new UE-ID with the new UE-ID that has been encrypted by the Target eNB.

Therefore, the mutual authentication is achieved between all handover entities while in current handover scheme the Target does not authenticates the Serving eNB and UE.

B. Key Backward/Forward Security

The proposed scheme depends on the secrecy of the pre-loaded shared key of the eNB (HOK) with the MME that is used to encrypt the NCC and NHNCC values that are exchanged between the eNBs and the MME. The new session key KeNB* is never sent between the handover entities as in the current scheme. Instead, the KeNB* is derived locally using the vertical key derivation function by the Target eNB and UE.

In the proposed scheme, an adversary cannot reversely deduce the previous session key from the current session key due to using the same vertical key derivation function of the current handover scheme. Consequently, the proposed scheme satisfies one-hop key separation for Backward/Forward

security. In contrast, the current scheme can achieve only two-hop forward security.

C. Anonymity

The proposed scheme changes the identity number of the UE (UE-ID) periodically. In each handover session, a new UE-ID will be computed by the UE and the MME. Subsequently, the user identity will be concealed in all next handover session between the user and network. Thus, through using fresh values of the NCC and UE-ID in all next handover sessions, the anonymity feature is hold in proposed scheme. In contrast, the same UE identity is used for all handover sessions in the current scheme.

D. Resistance to Attacks

In addition to, the attractive security features that are mentioned in previous sections, the proposed scheme can resist different attacks. Supposed an adversary can catch the handover messages between the handover entities, and can use a rogue base station to impersonate and control either the Serving eNB or the Target eNB. In the proposed scheme, an adversary cannot deduce the new session key parameters that are exchanged between the handover entities where all parameters are sent as encrypted messages. The NCC is sent to the Serving eNB as encrypted message from the UE using the KASME key. The NCC value that is sent to the MME from the Serving eNB also is encrypted by HOK-S. In same manner, the NHNCC is sent as encrypted message either through the S1 interface or X2 interface using the HOK-T and KeNB*, respectively. Therefore, the refreshing of the current NCC and NHNCC values cannot be disrupted by manipulating the message between handover entities, any change in the NCC or NHNCC, the handover process will be rejected by receipted entity.

In proposed scheme, handover process is performed through the MME, the Serving eNB sends the identity of the Target eNB as encrypted message to MME, the latter sends the Serving eNB identity to the Target eNB also as encrypted message. In addition to, the User identity is changed in each time the handover process is held. Therefore, if an adversary redirects or replays the handover messages to another eNB then handover process will be rejected. Consequently, the adversary cannot deduce the new session key or disrupt the refreshment of authentication parameters, also cannot reply or redirect the communication messages between the handover entities. Therefore, the drawbacks of the current scheme are eliminated, the proposed scheme can prevent the desynchronization attack, replay attack and redirection attack.

E. Comparisons

Table 2 shows that the proposed scheme achieves the highest security level among the other handover Key management schemes. In contrast, the current scheme achieves the lowest security level. As mentioned in previous sections, the proposed scheme provides several security features such as the mutual authentication between all handover entities, anonymity of the user, perfect Backward/Forward secrecy. Furthermore, the proposed scheme is secure against the desynchronization attack, replay attack and redirection attack.

TABLE II. SECURITY PROPERTIES AMONG THE HANDOVER SCHEMES

	Current HO	Lin et al. [15]	Khairy et al. [12]
Mutual Authentication	NO	NO	partially
Anonymity	NO	NO	Hold
Key Backward separation	One-hop	One-hop	One-hop
Key Forward separation	Tow-hop	One-hop	One-hop
Desynchronization attack	NO	Hold	Hold
Replay attack	NO	NO	partially
Redirection attack	NO	NO	partially
	Proposed HO	Mathi and Dharuman [17]	
Mutual Authentication	Hold	partially	
Anonymity	Hold	NO	
Key Backward separation	One-hop	One-hop	
Key Forward separation	One-hop	Tow-hop	
Desynchronization attack	Hold	No	
Replay attack	Hold	NO	
Redirection attack	Hold	NO	

VI. PERFORMANCE ANALYSIS

In this part, the performance analysis is discussed to observe the effect of security level enhancement in the proposed scheme during the X2 handover process.

The numerical results in terms of operations cost of authentication and bandwidth overhead are discussed by comparing the proposed scheme with different handover Key management schemes.

TABLE III. ASSUMPTIONS OF THE LTE NETWORK

Assumptions	Assumptions values
Mean density of UE/USIM ρ .	300/km ²
Total number of UE/USIM.	$2 \times 49 \times 300 = 29400$
Size of MME Area.	49 km ²
Average rate of originating service request.	1/hr/user
Average rate of terminating service request.	1/hr/user
Average speed of UE/USIM \mathcal{V} .	5 km/hr
Number of MME.	2 MMEs.
Number of TA.	128 TAs.
Number of the eNB in each TA	2, 3, 5 eNBs.
Border covered length ℓ .	30 km.

For bandwidth overhead consumption, the handover key management schemes have been simulated in MATLAB running on a 2.10 GHz processor with 4GB memory computing machine. Table 3 illustrates the assumptions of the LTE network.

A. Operations Cost of Authentication

TABLE IV. NOTATIONS OF THE OPERATIONS COST

Notations	Description
Cc	Encryption/ decryption cost
Kc	Key derivation cost
Vc	Verification cost
Gc	Generate new identifier cost
Rc	Refreshment parameter cost

Table 4 illustrates the notations of the operations cost in the authentication process. In this context, assume that the cost of all operations per hop are equal to 1 unit and the operations vector (O_v) in each entity of handover process can be

determined as $[C_c, K_c, V_c, G_c, R_c]$, the O_v represents how many times the operations are executed in the handover entity.

$$V_{sum} = \sum O_v [C_c, K_c, V_c, G_c, R_c] \quad (4)$$

The sum of operations vectors (V_{sum}) for all handover entities represents the operations cost of authentication in handover Key management scheme. Here, the (V_{sum}) is a vector defined as in (4).

TABLE V. OPERATIONS COST IN EACH HANDOVER ENTITIES

	UE	Serving eNB	Target eNB
Mathi and Dharuman [17]	[0, 1, 1, 0, 1]	[0, 1, 0, 0, 1]	[0, 1, 0, 1, 2]
Lin et al. [15]	[0, 1, 0, 0, 1]	[2, 1, 0, 0, 1]	[1, 0, 0, 0, 2]
Khairy et al. [12]	[4, 1, 3, 2, 0]	[2, 0, 0, 0, 0]	[5, 1, 2, 1, 0]
Current	[0, 1, 0, 0, 0]	[0, 1, 0, 0, 1]	[0, 0, 0, 0, 2]
Proposed	[2, 1, 1, 1, 1]	[3, 0, 1, 0, 1]	[6, 1, 1, 0, 2]
	MME	Vsum	Total of Vsum
Mathi and Dharuman [17]	[0, 0, 1, 1, 2]	[0, 3, 2, 2, 6]	13 unit
Lin et al. [15]	[1, 0, 0, 0, 2]	[4, 2, 0, 0, 6]	12 units
Khairy et al. [12]	[8, 0, 0, 3, 0]	[21, 2, 5, 5, 0]	33 units
Current	[0, 0, 0, 0, 2]	[0, 2, 0, 0, 6]	8 units
Proposed	[7, 0, 1, 1, 2]	[18, 2, 4, 2, 5]	31 units

The results in Table 5 show that the operations cost of authentication in the handover key management schemes increase with the increases of the security level. Therefore, if the security level has increased, then the number of operations will be increased, especially the encryption/decryption and verification operations. Compared with other handover key management schemes, the proposed scheme can achieve the highest security level with desirable authentication cost.

B. Bandwidth Overhead

Liang and Wang [14] classify the intra-domain handoff authentication request events. The numerical analysis is taken into account just the event that when the UE starts the request within current MME domain and this request ends before the UE moves to another MME domain.

$$\lambda_1 = \lambda_u P_r (\bar{N}_a - 1) \quad (5)$$

Therefore, the arrival rate of handoff authentication requests (λ_1) can be calculated as in (5) [8], [14]. Where (λ_u) is the service request arrival rate, \bar{N}_a is the average numbers of eNBs passed by the UE in the same MME domain and (P_r) is the probability that X2 handover happens.

$$T_{AP} = \sum_{i=1}^6 \text{Auth}_i \quad (6)$$

Let the authentication parameters size between (UE-Serving eNB), (UE-Target eNB), (UE-MME), (Serving eNB-Target eNB), (MME-Serving eNB), and (Target eNB-MME) be Auen1, Auth2, Auth3, Auen4, Auth5 and Auth6, respectively.

Therefore, the total size of authentication parameters (T_{AP}) that are exchanged between the handover entities is calculated as indicated in (6).

TABLE VI. TOTAL SIZE OF AUTHENTICATION PARAMETERS IN THE HANDOVER KEY MANAGEMENT SCHEMES

Handover schemes	(T_{AP})
Mathi and Dharuman scheme	2432 bits
Lin et al. scheme	2176 bits
Khairy et al. scheme	4224 bits
Current scheme	2176 bits
Proposed scheme	3456 bits

The total size of authentication parameters of the handover key management schemes, as depicted in Table 6, the proposed scheme, Khairy et al. [12] scheme and Mathi and Dharuman [17] scheme, consume during the handover key management (3456), (4224) and (2432) bits, respectively while, Lin et al. [15] scheme and the current scheme consume 2176 bits.

Therefore, the total size of authentication parameters of proposed scheme ranges in the middle. However, the proposed scheme is secure against various attacks and provides more security features than the other schemes.

$$T_{BW} = 128 \times |(\lambda_1 \times T_{AP})| \quad (7)$$

Subsequently, the total bandwidth of the handover process (T_{BW}) for each handover Key management schemes is defined as in (7).

The effect of security level enhancement is shown in Fig. 4 and 5. These figures depict that the relationships between the bandwidth overhead during the handover process with authentication handover key management scheme when the $\bar{N}_a = 1$ and $\bar{N}_a = 2$, respectively.

Therefore, the total bandwidth overhead consumption increases with the increase of the security level. Compared with other handover key management schemes, the proposed scheme provides several security features with desirable bandwidth overhead consumption.

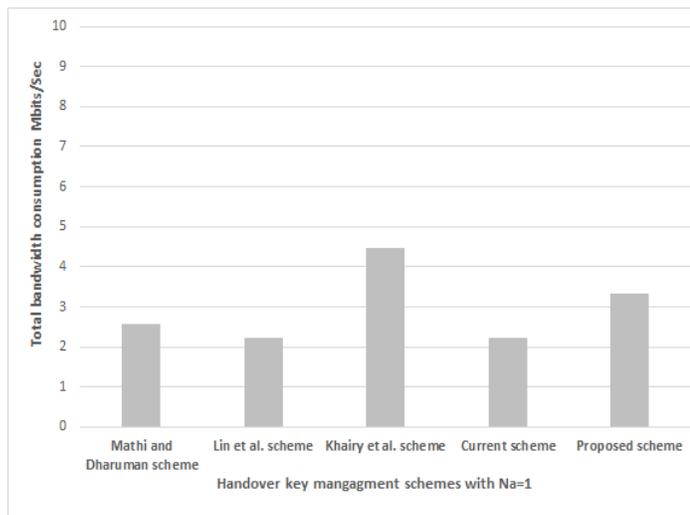


Fig. 4. Total bandwidth when ($\bar{N}_a = 1$).

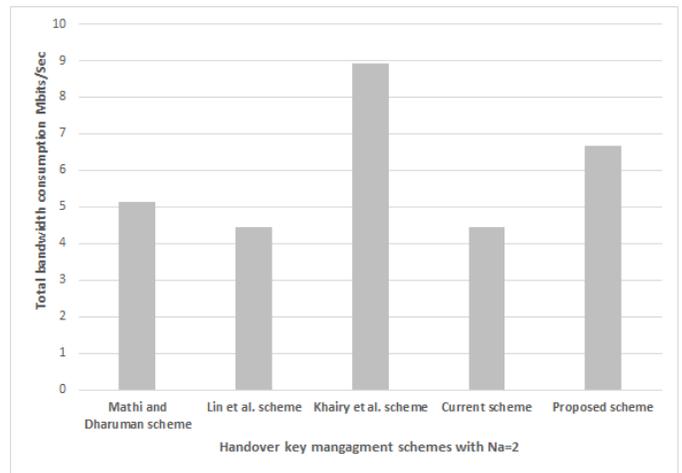


Fig. 5. Total bandwidth when ($\bar{N}_a = 2$).

VII. CONCLUSION

This paper proposes a synchronous handover authentication scheme to prevent the drawbacks of the current handover key management scheme during Inter-eNB handover over LTE networks. Compared with other authentication handover key management schemes, the proposed scheme not only provides strong security features including perfect Backward/Forward secrecy and user anonymity but also the mutual authentication between all handover entities.

The security analysis has shown that the proposed scheme is secure against various attacks such as the desynchronization attack, replay attack and redirection attack. The accurate performance analysis in terms of operations cost of authentication and bandwidth overhead has been discussed, which demonstrates that the authentication cost and bandwidth overhead consumption of the whole handover process are desirable among the other handover Key management schemes.

REFERENCES

- [1] 3gpp-ts 33.401, 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; 3GPP System Architecture Evolution (SAE); Security architecture (2015-12), Release 11 Technical Specification.
- [2] M. Abdeljebba, R. El Kouch, "Fast Authentication during Handover in 4G LTE/SAE Networks", Open Access IERI Procedia, Vol 10, pp. 11-18, 2014.
- [3] P. Agarwal, D. Thomas, and Kumar. A, "Security Analysis of LTE/SAE Networks under De-synchronization Attack for Hyper-Erlang Distributed Residence Time", IEEE Communications Letters, Vol 21, No 5, pp.1055-1058, 2017.
- [4] W. Ahmed, S. Anwar. and M. Arshad, "Security Architecture of 3GPP LTE and LTE-A Network: A Review", INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY SCIENCES AND ENGINEERING, Vol 7, No 1, 2016.
- [5] J. Cao, M. Ma, H. Li, Y. Zhang, and Z. Luo, "A Survey on Security Aspects for LTE and LTE-A Networks", IEEE COMMUNICATIONS SURVEYS & TUTORIALS, Vol 16, No 1, pp.283-302, 2014.
- [6] E. El-Gaml, H. ElAttar, H. El-Badawy, "Evaluation of Intrusion Prevention Technique in LTE Based Network", International Journal of Scientific & Engineering Research, Vol 5, No 12, pp.1395- 1400, 2014.
- [7] D. Forsberg, G. Horn, W. Moeller and Niemi. V, "LTE SECURITY", John Wiley and Sons, United Kingdom, 2013.

- [8] F. Degefa, D. Lee, J. Kim, Y. Choi and D. Won, "Performance and security enhanced authentication and key agreement protocol for SAE/LTE network", *Computer Networks*, Vol 94, pp. 145-163, 2016.
- [9] Z. Haddad, M. Mahmoud, S. Taha, and I. Saroit, "Secure and Efficient Uniform Handover Scheme for LTE-A Networks", In *Wireless Communications and Networking Conference (WCNC)*, pp. 1-6 IEEE 2016.
- [10] C. Han, H. Choi, "Security Analysis of Handover Key Management in 4G LTE/SAE Networks", *IEEE Transactions on Mobile Computing*, vol. 13, no. 2, pp. 457-468, 2014.
- [11] T. Karpagam, S. Sivakumar, "Efficient and Secure Authentication Handover using Network functions virtualization", *International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)*, Vol 17, No 1, pp.36-40, 2015.
- [12] K. Khairy, A. Diao Eldien, A. Abdel-hafez and E. Abd El-Wanis, "Authenticated Key Management Scheme for Intra-Mme Handover Over LTE Networks", *International Journal of Research in Engineering and Science (IJRES)*, Vol, No 10, pp. 19-28, 2016.
- [13] C. Lai, H. Li, R. Lu and X. Shen, "SE-AKA: A secure and efficient group authentication and key agreement protocol for LTE networks", *Computer Networks*, Vol 57, pp.3482-3510, 2013.
- [14] W. Liang, W. Wang, "On performance analysis of challenge/response based authentication in wireless networks", *The International Journal of Computer and Telecommunications Networking*, Vol 48, No 2, pp. 267-288, 2005.
- [15] Y. Lin, W. Longhuang and C. Yang, "Enhanced 4G LTE Authentication and Handover Mechanism", *International Journal of Electrical, Electronics and Data Communication*, Vol 3, No 9, pp.45-47, 2015.
- [16] M. Masud, "Survey of security features in LTE Handover Technology", *Scientific Research Journal (SCRJ)*, Vol, No 8, pp.27-31, 2015.
- [17] S. Mathi, L. Dharuman, "Prevention of Desynchronization Attack in 4G LTE Networks Using Double Authentication Scheme", *Open Access Procedia Computer Science*, Vol 89, pp.170- 179, 2016.
- [18] S. Nashwan, B. Alshammari, "Formal Analysis of MCAP Protocol Against Replay Attack", *British Journal of Mathematics & Computer Science*, Vol 22, No 1, pp. 1-14, 2017.
- [19] S. Nashwan, B. Alshammari, Mutual Chain authentication protocol for SPAN Transactions in Saudi Arabian Banking, *International Journal of computer and communication engineering*, Vol 3, No 5, pp. 326- 333, 2014.
- [20] N. Qachri, O. Markowitch and J. Dricot, "A Formally Verified Protocol for Secure Vertical Handovers in 4G Heterogeneous Networks", *International Journal of Security and Its Applications*, Vol 7, No 6, pp.309-326, 2013.
- [21] B. Sridevi, D. Mohan, "Security analysis of Handover Key Management among 4G LTE entities Using Device Certification", *International Journal of Electrical, Computing Engineering and Communication*, Vol. 1, No 2, pp. 1-7, 2015.
- [22] P. Tayade, P. Vijaykumar, "A Comprehensive Contemplate on Security Aspects of LTE and LTE Advanced in Wireless Communication Network", *International Journal of Control Theory and Applications*, Vol 10, No 31, pp.197-217, 2017.

A New Cryptosystem using Vigenere and Metaheuristics for RGB Pixel Shuffling

Zakaria KADDOURI¹

Laboratory of Computer Science Research, Department of Computers Science,
Mohammed V University Agdal – AbuDhabi, AbuDhabi, United Arab Emirates

Mohamed Amine Hyaya²

Physics Department
Mohammed V University – Faculty of Sciences Rabat,
Rabat, Morocco

Mohamed KADDOURI³

LMPHE Laboratory
Mohammed V University – Faculty of Sciences Rabat,
Rabat, Morocco

Abstract—In this article we present a new approach using Vigenere and metaheuristics to resolve a problem of pixel shuffling to cipher an image. First the image is adapted to match the resolution system by transforming it to a list of intensities and coordinates. The idea is to use Vigenere encryption to maximize the confusion by widening the domain of intensities. Then, metaheuristics play the major role of encryption, generating an appropriate Meta key in order to shuffle the lists. Thus, both Vigenere key and Meta-key are used for encryption and later in decryption by the recipient. Finally, a comparison of different metaheuristics is proposed to find the most suitable one for this cryptosystem.

Keywords—Cryptography; cryptosystem; Vigenere; metaheuristics; image; pixel shuffling

I. INTRODUCTION

In theory, a combinatorial optimization problem can be defined by all its instances. In practice, the problem is reduced to mathematically solving one of these instances, by the algorithmic method [4]. Metaheuristics are a family of optimization algorithms that aim to solve general classes of mathematical problems by combining search procedures to quickly find the best solution.

In 2005, a new encryption system [16], called SEC (Symmetrical Evolutionist-based Ciphering) was introduced, which is strongly linked to evolutionary algorithms and represents the first adaptation of a metaheuristic to the domain of cryptography. Its principle consists in constructing lists containing the different positions of the characters of a plaintext, and it connects the evolutionary processes (evaluation, selection, crossing and mutation operator) applied on the order of these lists to obtain a maximum disorder without modifying their contents. At the end of the algorithm, a key known as “gene key” is generated and used for both encryption and decryption operations [7]-[9], [13]-[15].

In our approach, we used multiple metaheuristics to find a strong key for our encryption. Metaheuristics can be divided into two main groups: 1) Single Solution Algorithms; and 2) Population-based Algorithms.

A. Single Solution Algorithms

Single Search Algorithms, i.e. local and global searches, start with a random solution then tries to optimize it, following a given criteria. Various Algorithms are actually used and improved, such as Hill Climbing (**HC**), which is classified among Local Searches. It optimizes the solution following the highest lean in its neighborhood [10], [18], [19], Simulated Annealing (**SA**), is a global search based on Monte Carlo methods [5], [20]. This algorithm avoids local optimums by choosing a less optimal solution if the aspiration criteria is met [3], and finally the Taboo search (**TS**), is also a global search, that escapes the local optimums by memorizing a list of previous solutions and selecting only unexplored solutions [4], [5], [9].

B. Population-based Methods

Contrarily to the previous methods population based algorithms optimize multiple solutions simultaneously. Among many, Genetic Algorithm (**GA**) uses natural selection. It combines individuals from the initial population to give birth to the next generation of solutions then, only the fittest ones are chosen to reproduce and create the next one and so on [2], [5], [7], [10], while particle swarm optimization (**PSO**) is developed on swarm behavior of birds. The initial population is created then a goal is set, unlike **GA**, **PSO** is not eliminating individuals, but each individual evolves differently so the whole group would reach the goal in an optimal way [3]. Back to **GA** one can say that even a normal individual may have more room for improvement than the fittest, Memetic algorithm (**MA**) solves this problem since it uses a local search to optimize every solution (one individual) before choosing the fittest. **MA** is a hybrid algorithm, a population based method using a local search to optimize intermediate solutions [5], [8], [10], [17].

Section II describes the proposed approach, including the methods used to optimize the cyphering or adapt different components of the cryptosystem.

Section III, shows both qualitative and quantitative analysis conducted on our cryptosystem.

Finally Section IV, discusses briefly the proposed algorithm and potential optimizations.

II. OUR CRYPTOSYSTEM

A. Description

The cryptosystem generates a symmetrical encryption Key [12]. The main idea is to shuffle the colors of the image “M” with dimensions ($W \times H$), using the generated key. First, we apply a preliminary encryption using Vigenere algorithm [18], directly on the RGB image [6], i.e. a random key is generated, as in vigenere encryption the key is repeated until it reaches the length of element to be encrypted. In our case, the ASCII number of each character of the key will be added to the RGB values of a pixel. Then the RGB components of the image are placed vertically, getting a ($W \times 3H$) grayscale image “M’”. The initial solution is created given as a Table “X” of 256 intensities, to each value we assign a list “L” of coordinates of that value in the image V. These lists will help reconstruct the image. The shuffling starts by permuting intensities, i.e.

$$\begin{aligned} X_0 &= \{0,1,2,3, \dots, 253,254,255\} \rightarrow \\ X_f &= \{251,149,50,61, \dots, 13,22,200\} \end{aligned} \quad (1)$$

The best solution is chosen using metaheuristic algorithms and evaluated by the evaluation function “f”:

$$f(i) = \sum_{j=0}^{255} |X_i[j] - X_0[j]| \quad (2)$$

Finally, the encrypted image is reconstructed using the new list of intensities and assigning them to the coordinates stored at beginning, i.e. Let L_i from (1) be the best solution, then all coordinates initially black (intensity=0) will be assigned the value 251, and all coordinates with intensity 1 will be assigned, 149, as for the remaining 254 intensities.

B. Skeleton of our Cryptosystem

Let M be the RGB matrices of the image to be encrypted, with $M(x, y, z)$ a pixel of the image M such as $x \in \{0,1,2, \dots, W\}$, $y \in \{0,1,2, \dots, H\}$ and $z \in \{0,1,2\}$.

We create a list of random values to be our Vigenere Key

$$0 \leq V(w) \leq 255 \text{ with } w \in \{0,1, \dots, V_l\} \text{ and } 30 \leq V_l \leq 50$$

M’ is the encrypted image using Vigenere Key V as follows:

$$\forall x, \forall y, M'(x, y, 0) = [M(x, y, 0) + V((y + x.W) \bmod V_l)] \bmod 256$$

$$\forall x, \forall y, M'(x, y, 1) = [M(x, y, 1) + V((y + x.W) \bmod V_l)] \bmod 256$$

$$\forall x, \forall y, M'(x, y, 2) = [M(x, y, 2) + V((y + x.W) \bmod V_l)] \bmod 256$$

I is a grayscale image made of vertical concatenation of RGB matrices:

$$\begin{aligned} I(x, y) &= M'(x, y, 0) \\ I(x + H, y) &= M'(x, y, 1) \\ I(x + 2H, y) &= M'(x, y, 2) \end{aligned} \quad (3)$$

I is then represented using lists of different Intensities, each list contains the (x, y) coordinates of a given intensity, element of the set $\{0,1,2, \dots, 255\}$. We denote by L_i ($0 \leq i \leq 255$) a list of the different positions of the Intensity and X_{iter} : A list of all intensities in a given iteration.

The goal is to create a maximum disorder between intensities in a manner that the difference transcends a given threshold. Metaheuristics are used to generate a random key while maximizing to a certain degree the disorder in X. We denote X_{final} .

To cipher the image, we reconstruct it using the order of intensities in X_{final} for example (Table 1):

TABLE I. EXAMPLE OF LIST PERMUTATIONS

X_0	Intensities	0	1	2	...	253	$\frac{25}{4}$	255
	Coordinates	L_0	L_1	L_2	...	L_{253}	$L_{\frac{25}{4}}$	L_{255}
X_{final}	Intensities	251	168	59	...	2	$\frac{11}{2}$	15
	Coordinates	L_0	L_1	L_2	...	L_{253}	$L_{\frac{25}{4}}$	L_{255}

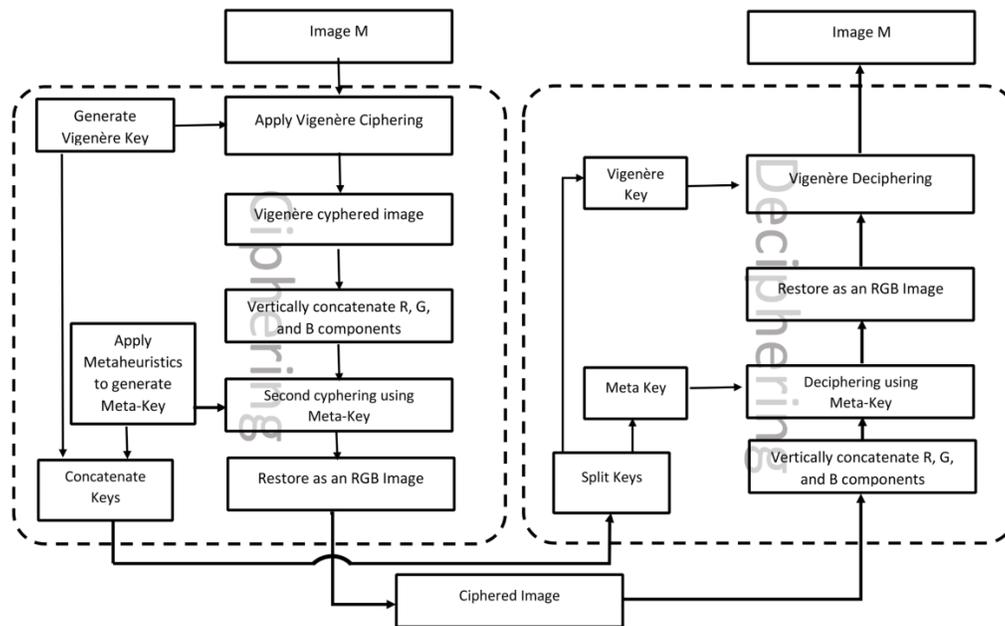


Fig. 1. Diagram of our cryptosystem.

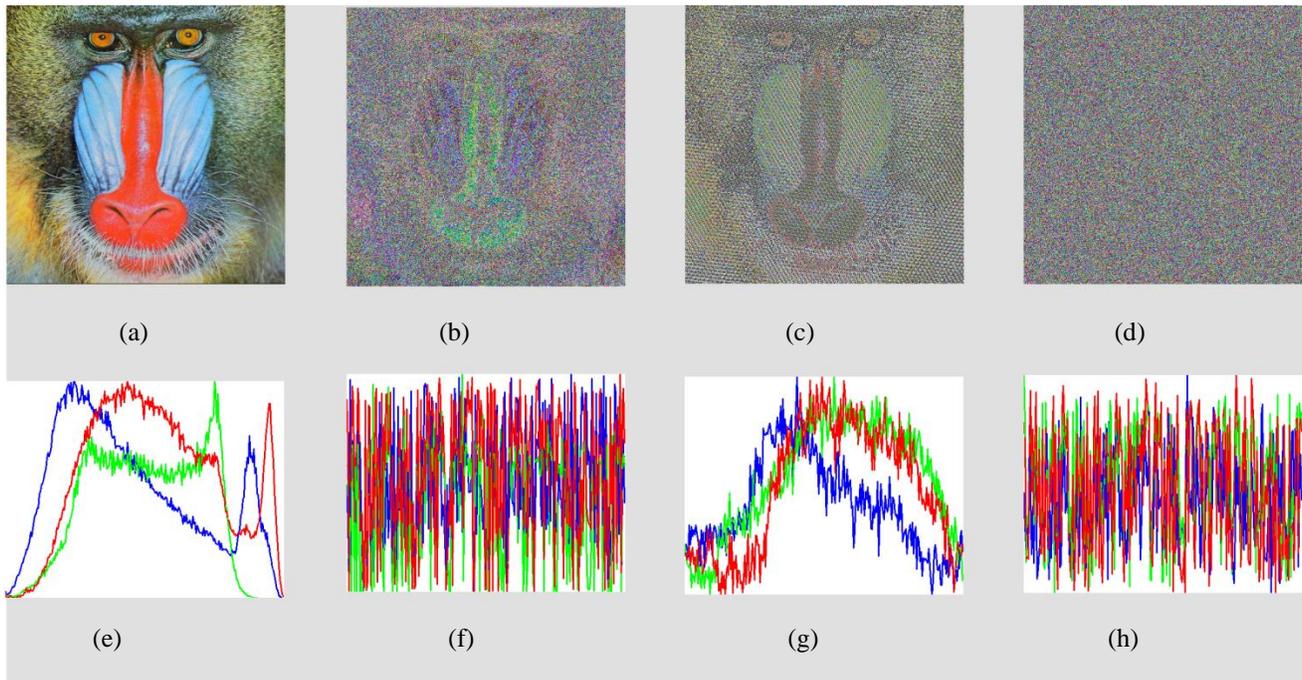


Fig. 2. Baboon Image

(a) original, (b) pixel shuffling only, (c) Vigenere only, (d) proposed cryptosystem, (e) Original histogram, (f) pixel shuffling histogram, (g) Vigenere histogram, (h) cryptosystem histogram.

X0 is the table containing the initial order of intensities and coordinates. The permutation only affects the intensities. As a final result (see X_{final}) all the pixels that initially were pitch black “0” will be assigned the intensity 251 and pixels containing 1 will receive a value of 168 and so on. Fig. 1 summarizes our cryptosystem. During the encryption, the plain image is ciphered, using a randomly generated vigenere key, to enlarge the domain of colors to be shuffled. Then, Red, Green and Blue channels of the resulting image are separated and concatenated vertically, forming a grayscale like

image. At this stage, a list of intensities is derived from the grayscale image. A single solution metaheuristic takes the list to be the initial solution, while population based metaheuristics, derive the initial population using random permutations on that list. At the end of the optimization, the solution returned, is called Meta-key, it allows the permutation of intensities as described previously. Finally, the RGB image is restored by rebuilding a three channels image by dividing the cyphered grayscale image. Decryption, is following the same methods except that both meta-key and vigenere key are

shared. Grayscale image is constructed from the cyphered image. Then intensities get permuted using meta-key. Next, RGB image is restored and finally we use vigenere key to get the Plain Image.

III. EXPERIMENTAL RESULTS

In this section, we use a benchmark image to study the efficiency of our cryptosystem, where we compare multiple metaheuristics including both local searches and population-based Algorithms.

A. Visual Tests

In Fig. 2 and 3, we propose an explanation for combining both Vigenere and metaheuristic keys, as one can observe the images ciphered by Vigenere and meta-key separately still

recognizable by humans. First Vigenere encryption concatenates the key line by line, and changes the colors using the same pattern. If it encounters a big spot containing the same color, the patterns can be easily found (Fig. 2(c) and 3(c)). The same issue occurs for metaheuristic encryption, since the algorithm only permutes the colors. In consequence, we observe an image with similar forms but with different colors as seen in Fig. 2(b) and 3(b). Besides, the image encrypted by the combination of both algorithms is totally unrecognizable (Fig. 2(d) and 3(d)). In fact, Vigenere widens the domain of colors, breaks the contours of the image and adds a strong noise to the spots of similar colors, allowing the metaheuristics to permute intensities and ensure a maximum disorder in the final image. This can be observed by comparing histograms in Fig. 2 and 3 (e-f).

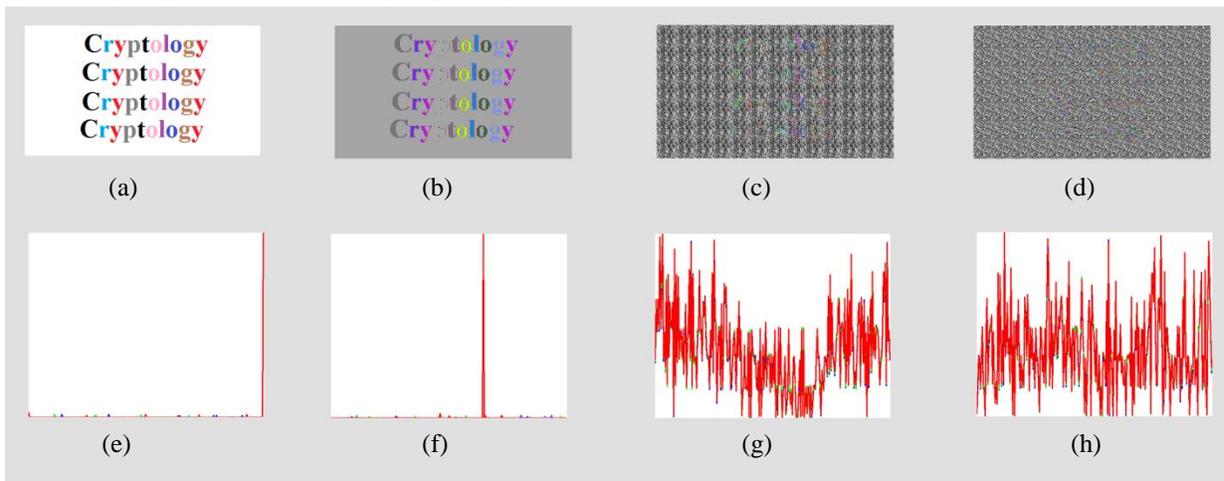


Fig. 3. Cryptology Image
(a) original , (b) pixel shuffling only, (c) Vigenere only, (d) proposed cryptosystem,
(e) Original histogram, (f) pixel shuffling histogram, (g) Vigenere histogram, (h) cryptosystem histogram.

B. Quality Tests

1) NPCR

$$NPCR = \frac{1}{m \cdot n} \sum_{i=1}^m \sum_{j=1}^n \delta_{I_0}^{I_c(i,j)}$$

; with δ : Kronecker delta

The number of pixel change rate (NPCR) is usually used to evaluate the absolute number of pixels change rate [21]. The more pixels change, the closer to 1 we get. In our case, as we can see, Fig. 4 presents the NPCR values between the original and ciphered image, for 10 different runs, the proposed algorithms gave nearly optimal values of NPCR.

However, this maximal value would also involve a binary image and its negative, the last, can be easily recognized. Thus, NPCR proves only that pixels of the original image changed, but it may still be recognizable. This is why we must perform PSNR to evaluate the noise ratio in the ciphered image and SSIM for similarity between the ciphered and the original image.

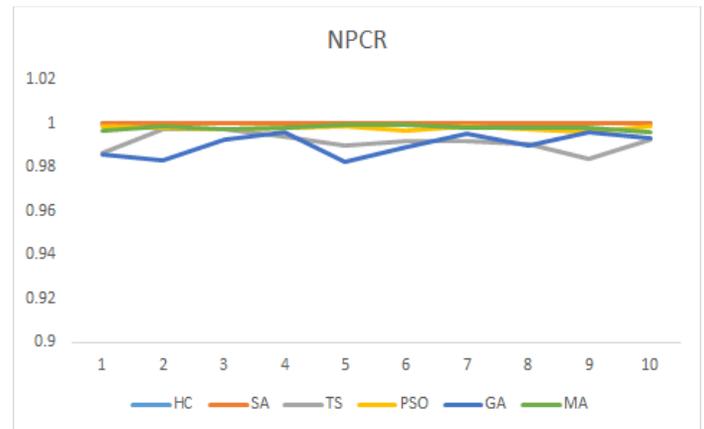


Fig. 4. NPCR values for different metaheuristics.

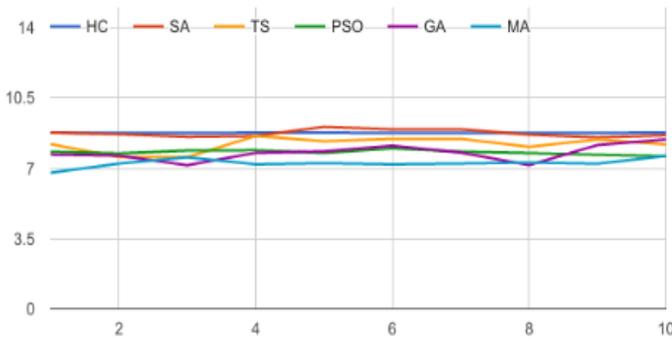


Fig. 5. PSNR values for different metaheuristics.

2) PSNR

$$PSNR = 10 \times \log_{10} \left(\frac{255^2}{MSE} \right)$$

$MSE = \frac{1}{m \times n} \sum_{i=1}^m \sum_{j=1}^n (I_0(i, j) - I_c(i, j))^2$ Peak signal to noise ratio (PSNR) is calculated to measure distortion in a digital image by calculating the amount of noise in the image [1]. The smaller the value of PSNR, the less signal is conserved. Let us consider two identical images and add 1 to one component of one pixel to the second image. The PSNR of those two images is going to be the highest possible after infinity, PSNR of two identical images. If we apply the previous condition to the size of our benchmark image:

$$PSNR_{max} = 102.3162028 \text{ dB}$$

While $PSNR_{min} = 0 \text{ dB}$, considering the log scale, all ciphered images offering a $PSNR < 10 \text{ dB}$ can be considered a good encrypted image. We summarize the values of PSNR given by the experiment, previously described in Fig. 5.

3) SSIM

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2cov_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

The Structural Similarity index map (SSIM) computes a similarity map between two digital images “x” and “y” as confirmed by [11] it allows simulating human perception in comparing two images. The map value $-1 \leq SSIM(i, j) \leq 1$ where one means images are similar around that region. Thus for two identical images, all map values equal one. Meanwhile, negative values attest inverted regions. Finally, zero states totally different regions. Fig. 6 is computed as follows:



Fig. 6. SSIM values for different metaheuristics.

$$\overline{SSIM} = \frac{1}{m \times n} \sum_{i=1}^m \sum_{j=1}^n |SSIM(i, j)| \times 100\%$$

This means that the computed value (in percent) for two completely different Images is 0% while, 100% implies either identical images or an image and its negative.

All values obtained in this experiment are below 25%.

4) Encryption time

This time is actually is for the whole cryptosystem including key generation (Vigenere & Meta key) and pixel shuffling. We observe that the encryption is very fast since the size of the image used is 300KB. For example, in the case of HC metaheuristic, the encryption rate is (17 554 285 bytes/sec). We can notice in Fig. 7 that the execution time for GA and MA is too high compared to the other metaheuristics, but this is due to multipoint crossover that needs to eliminate duplicates every time it generates a child.

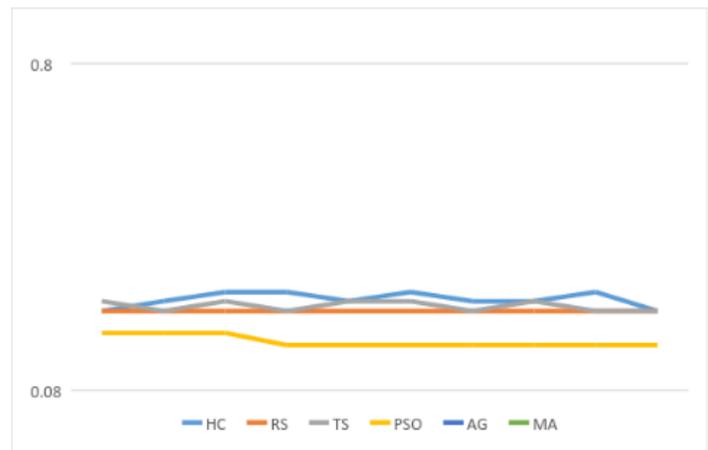


Fig. 7. Encryption time for different metaheuristics.

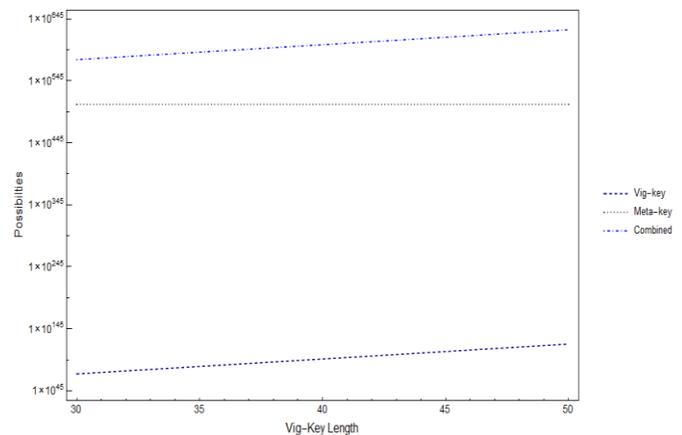


Fig. 8. Number of possible Keys.

5) Key strength

This cryptosystem proposes two complementary and symmetric encryptions. The final key to be shared is a simple concatenation of both keys. The challenge for breaking the key is that the generated key is totally independent from the

image, and the length of the key is not fixed since Vigenere key length is between 30 and 50 Bytes. In addition, the Meta-key is 256 Bytes Long, which give us:

$$286 \leq \text{key length} \leq 306 \text{ Bytes}$$

The total number of combination is given by:

$$N = 256^k \times P_{256}^k; k \text{ Vigenere key length} \\ \text{since: } 30 \leq k \leq 50 \Rightarrow 8.92 \times 10^{575} \leq N \leq 4.13 \times 10^{622}$$

Fig. 8 is a semilog graph showing the number of possible keys for Vigenere key, Meta key separately and the combination of both versus Vigenere key length.

Table 2 gives the means and standard deviation obtained after 10 runs on each metaheuristics. We observe that the numbers are very stable and consistent since the runs were not selected. The experiment is totally independent: no seeds were planted for the pseudorandom generator. The values side by side are quite similar except for SSIM, the values vary from 7 to 15%.

TABLE II. QUALITY TESTS SUMMARY

	NPCR			PSNR			SSIM		
	μ	\pm	σ	μ	\pm	σ	μ	\pm	σ
HC	0.999909	\pm	0.00005	8.773	\pm	0.007	7.007	\pm	0.035
SA	0.999997	\pm	0.00000	8.752	\pm	0.171	7.078	\pm	0.200
TS	0.991533	\pm	0.00404	8.197	\pm	0.350	18.627	\pm	2.561
PSO	0.997794	\pm	0.00092	7.804	\pm	0.108	15.354	\pm	0.487
GA	0.990335	\pm	0.00495	7.780	\pm	0.388	17.146	\pm	1.715
MA	0.997813	\pm	0.00104	7.264	\pm	0.213	17.669	\pm	2.007

IV. DISCUSSION AND CONCLUSIONS

The proposed algorithm reveals very satisfying results. Overall, it is compatible with all the tested metaheuristics. However the parameters have to be set for every metaheuristic to obtain good results, but once set, the results are stable and render the encrypted image unrecognizable. On the other hand, if we compare the proposed metaheuristics to choose the best one(s), NPCR rates Simulated Annealing and Tabu search as the best ones, While PSNR values give a slight preference for population based algorithms, MA, GA and PSO, besides HC and SA, outclass the other metaheuristics according to SSIM. As for the encryption time all the metaheuristics except MA and GA, are very fast. Moreover, the key generated offers a high security level compared to the existing symmetrical cyphers. Despite being very satisfying, the algorithm is very flexible and allows many ameliorations. For instance, improving Vigenere encryption part or choosing different evaluation function.

REFERENCES

- [1] Ahmad, J., & Ahmed, F. (2010). Efficiency analysis and security evaluation of image encryption schemes. *computing*, 23, 25.
- [2] Davis, L. (1991). *Handbook of genetic algorithms*.
- [3] Eberhart, R., & Kennedy, J. (1995, October). A new optimizer using particle swarm theory. In *Micro Machine and Human Science, 1995. MHS'95., Proceedings of the Sixth International Symposium on* (pp. 39-43). IEEE.
- [4] Glover, F. (1989). Tabu search—part I. *ORSA Journal on computing*, 1(3), 190-206.
- [5] Glover, F. W., & Kochenberger, G. A. (Eds.). (2006). *Handbook of metaheuristics* (Vol. 57). Springer Science & Business Media.
- [6] Jain, A. K. (1989). *Fundamentals of digital image processing*. Prentice-Hall, Inc..
- [7] KADDOURI, Z., OMARY, F., & ABOUCHOUAR, A. (2013). "BINARY FUSION PROCESS TO THE CIPHERING SYSTEM" SEC EXTENSION TO BINARY BLOCKS". *Journal of Theoretical & Applied Information Technology*, 48(1).
- [8] Kaddouri, Z., & Omary, F. (2014). *New Symmetrical Ciphering Approach Based on Memetic Algorithm*. algorithms, 1, 5.
- [9] Kaddouri, Z. (2014). *Mise en oeuvre de nouvelles techniques pour la sécurité informatique basées sur les algorithmes évolutionnistes et les fonctions de Hachage*.
- [10] Kumar, R., Tyagi, S., & Sharma, M. (2013). Memetic Algorithm: Hybridization of Hill Climbing with Selection Operator. *International journal of Soft Computing and Engineering*, 3(2), 140-145.
- [11] Li, C., & Bovik, A. C. (2009, January). Three-component weighted structural similarity index. In *IS&T/SPIE Electronic Imaging* (pp. 72420Q-72420Q). International Society for Optics and Photonics.
- [12] Mewada, S., Sharma, P., & Gautam, S. S. (2016). Classification of Efficient Symmetric Key Cryptography Algorithms. *International Journal of Computer Science and Information Security*, 14(2), 105.
- [13] Mouloudi, A., Omary, F., Tragha, A., & Bellaachia, A. (2006, November). An Extension of Evolutionary Ciphering System. In *Hybrid Information Technology, 2006. ICHIT'06. International Conference on* (Vol. 1, pp. 314-321). IEEE.
- [14] Omary, F. (2006). *Application des algorithmes évolutionnistes à la cryptographie*.
- [15] Omary, F., Tragha, A., Bellaachia, A., & Mouloudi, A. (2007). Design and evaluation of two symmetrical evolutionist-based ciphering algorithms. *IJCSNS International Journal of Computer Science and Network Security*, 7(2), 181-190.
- [16] Omary, F., Tragha, A., Lbekkouri, A., Bellaachia, A., & Mouloudi, A. (2005). *An Evolutionist Algorithm to Cryptography*. *Lecture Series and Computational Sciences*, 4, 1749-1752.
- [17] Radcliffe, N. J., & Surry, P. D. (1994, April). Formal memetic algorithms. In *AISB Workshop on Evolutionary Computing* (pp. 1-16). Springer Berlin Heidelberg.
- [18] Stinson, D. R. (2005). *Cryptography: theory and practice*. CRC press.
- [19] Tomita, M. (1982). Dynamic construction of finite-state automata from examples using hill-climbing. In *Proceedings of the fourth annual cognitive science conference* (pp. 105-108).
- [20] Van Laarhoven, P. J., & Aarts, E. H. (1987). *Simulated annealing*. In *Simulated Annealing: Theory and Applications* (pp. 7-15). Springer Netherlands.
- [21] Wu, Y., Noonan, J. P., & Aghaian, S. (2011). NPCR and UACI randomness tests for image encryption. *Cyber journals: multidisciplinary journals in science and technology, Journal of Selected Areas in Telecommunications (JSAT)*, 31-38.

Improved Hybrid Model in Vehicular Clouds based on Data Types (IHVCDT)

Saleh A. Khawatreh

Dept. of Computer Engineering
Faculty of Engineering, Amman University
Amman-Jordan

Enas N. Al-Zubi

Dept. of Computer Engineering
Faculty of Engineering, Amman University
Amman, Jordan

Abstract—In Vehicular Cloud (VC), vehicles collect data from the surrounding environment and exchange this data among the vehicles and the cloud centers. To do that in an efficient way first we need to organize the vehicles into clusters, each one works as a VC, and every cluster is managed by the cluster head (broker). The vehicles are grouped in clusters with adaptive size based on their mobility and capabilities. This model is an approach that forms the clusters based on the vehicles capabilities and handles with different types of data according to its importance to select the best route. A hybrid model is proposed to deal with these differences; Long-Term Evolution (LTE) is used with IEEE 802.11P which forms the traditional wireless access for Vehicular Ad hoc Networks (VANETs). This merge gives the high data delivery, wide-range transmission, and low latency. However, using only LTE based VANET is not practical due to its high cost and the large number of occurrences in the base stations. In this paper, a new Vehicular Cloud (VC) model is proposed which provides data as a service based on Vehicular Cloud Computing (VCC). A new method is proposed for high data dissemination based on the data types. The model is classified into three modes: the urgent mode, the bulk mode, and the normal mode. In the urgent mode, Long-Term Evolution (LTE) is used to achieve a high delivery with minimum delay. In the bulk mode, the vehicle uses IEEE 802.11p and chooses two clusters to divide this huge data. In the normal mode, the model works as D-hops cluster based algorithm.

Keywords—Vehicular Cloud (VC); Vehicular Cloud Computing (VCC); Vehicular Ad hoc Networks (VANETs); cloud algorithms; hybrid transmissions; IEEE 802.11p; Long-Term Evolution (LTE); transmission cost

I. INTRODUCTION

In recent years, Vehicular Cloud Computing (VCC) has attracted the concern of researchers to deploy the Intelligent Transportation System (ITS). VCC is the concept of merging Mobile Cloud Computing (MCC) and Vehicular Ad-hoc Networks (VANETs). MCC is the study of the characteristics of mobile agents (people, vehicles, robots). The mobile agents interact and collaborate among each other to sense the surrounding environment, process the data, propagate and aggregate it. The result of the interaction is to be shared among the network where this cannot be done using conventional Internet Cloud. On the other hand, VANET is a type of networks which consists of networks of vehicles to achieve a specific purpose.

VANETs have been established as an efficient network in which vehicles communicate among each other on highways and urban environments.

Nowadays, most vehicles are equipped with embedded systems, integrated computers, processing units and sensors. All these improvements provide a good platform to deliver Data as a Service (Daas), therefore, providing an efficient and timely data diffusion about such events as traffic jams, accidents, and road conditions. In VANETs, the mobility of vehicles, different network density and the frequent changes in topologies are the most challenges that must be considered. In addition to these challenges, real time applications are strict to delay and data delivery of these safety messages. Up to now, most VANETs have applied IEEE 802.11p as a communication method, which forms the conventional way for wireless access. IEEE 802.11 p supplies data rate range from 6 to 27 Mb/s at a short transmission distance, 300 m. Diffusion safety messages over a huge area needs an intelligent multi-hop broadcast techniques dealing with broadcasting overhead and frequent disconnections.

Recently, cellular technology has been used as an alternative to IEEE 802.11p. Due to the standardization of advanced broadcast/multicast in the Third-Generation (3G) which is called a Universal Mobile Communication System (UMCS), provides efficient and dynamic data dissemination over the network. As the rapid improvements on the communications, the Fourth Generation (4G) which is called Long-Term Evolution (LTE) is presented to support high data rates up to 300 Mb/s for downlinks, and up to 75 Mb/s for uplinks, with low delay of less than 5 ms, for up to 100 Km transmission range. Although these high benefits, LTE is not used alone in VANETs due to the high cost of communications between the vehicles among each other and between them and the base stations, take in mind the mobility of the vehicles and the high overload on the base station.

In recent years, hybrid solutions have been presented to achieve the benefits of the two merging techniques; IEEE 802.11 p and LTE. Therefore, we obtain the low cost, the high data rate, high transmission range, and low delay [16].

In this paper, a new improved model is presented compared with the DHCV [15] which is based on the use of the vehicles capabilities and the hybrid transmission for data dissemination. This model is the first one that takes vehicles capabilities into consideration when clustering the vehicles. In this model, the

clustering technique is used to organize the exchange of data between vehicles and between the vehicles and the base stations. Every cluster is formed based on the QoS of the data collected from the sensors, and selects the CH based on its mobility and capabilities. After the clusters are performed, the Cluster Member (CM) which has an emergency data uses the LTE to transmit this data; therefore, we insure the delivery of this important data with low latency.

Section II presents most of the related works. Section III describes the system model. Section IV provides comparison scenarios and evaluation of the simulation results. Finally, Section V presents some concluding remarks and provides future work.

II. RELATED WORKS

Several studies have been proposed in VANET in general, and in data allocation routing specifically. In this section, a review of some existing methods is presented. To have a better understanding of these methods two categories of cluster-based algorithms have been studied so far; the first one is based on location; where speed, location, and direction of movement are used for cluster formation. The second one is based on computable collective parameters, for example, network density and connectivity. The most existing cluster-based algorithms concerning with these categories form one-hop cluster. Such as, Stability Based Clustering Algorithm (SBCA) [1] where this algorithm is one-hop cluster and constructing clusters based on the relative mobility among the vehicles and the accessibility of the Cluster Head (CH). To obtain this stability a Secondary Cluster Head (SCH) is selected along with the cluster head, this algorithm chooses Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) for transmissions, therefore a frequent contention occurs.

CSMA/CA is the basic for the MAC protocol which operates in IEEE 802.11p. MAC protocol is not the efficient way to use due to the features of VANETs such as vehicles mobility interference and hidden nodes. In situations of high vehicle density, the vehicles wait for a long time for sharing medium access, therefore, the network will suffer from low data throughput and long delay. Many protocols were presented to address these problems, such as; Space Division Multiple Access (SDMA) protocol [2] which allocates different frequency ranges to the space units in the cluster. Ad-Hoc MAC protocol [3] makes the transmission scheduling according to assign time slots to the vehicles which aim to access the medium. So the performance of these protocols decreases in high dense networks due to the hidden node and the congestion. Many protocols presented to address these problems, one of them is Clustering-based multichannel MAC protocol [4], this protocol is one-hop clusters approach that makes some improvements on the MAC layer by clustering the vehicles into clusters and let the cluster head control the cluster member's transmission in the shared medium, and support real time data transmission and not-real time data within quality of service. Therefore, the congestion is decreased compared with the previous protocols.

Hidden node problem is solved by Cluster-based MAC protocol (D-CBM) [5] that clusters the vehicles according to their mobility; it uses CSMA/CA or TDMA to schedule the

transmission. In [6], a transmission based on QoS-TDMA is proposed; this protocol assigns pre-reserved slots of time that satisfy the priority. In [7] TDMA cluster-based MAC protocol assigns different slots to the cluster members in its one hop, therefore a fairness is achieved. These MAC protocols address the intra-cluster transmissions without collision.

All the above algorithms are one-hop clusters. These clusters have small coverage range; therefore, the movement of vehicles will reconstruct the clusters. On the other hand, multi-hop clusters achieve better performance due to its stability and the decrement of reconstructions.

Hierarchical Clustering model (HC) [8] is a randomized clustering algorithm. It clusters the vehicles based on the connectivity data among the neighbors without using a GPS to locate their locations. The size of the clusters is limited to two hops only. HC algorithm constructs the clusters in the first stage and does the adjustments in the maintenance phase. HC algorithm does not consider the mobility of the vehicles and that was the major drawback of it. In [9] a Modified Distributed Mobility-Adaptive Cluster (Modified DMAC) is proposed. It improves DMAC [10] by clustering the vehicles which only have the same direction of movement. The major drawback here is it does not consider the mobility of the vehicles in selecting the cluster heads. In [11], a Distributed Multi-Hop Clustering Scheme for VANETs based on a neighborhood Follow (DMCNF). In this algorithm the vehicles follow their one-hop neighbors according to these factors: the historical cluster membership, relative mobility, and the number of followers. The clusters in DMCNF tend to be large and this decreases the network throughput and increases the delay due to the large number of cluster members and that makes a bottleneck in the cluster heads. In [12] a multi-hop clustering model is proposed. It clusters the vehicles based on the relative mobility between them in multi-hop range. Each node (vehicle) selects its cluster head in at most D hops. Beacon messages are exchanged among the vehicles within the D hops, and each node calculates the beacon delay. The node with the least delay among D hops broadcasts itself as a cluster head. This algorithm improves the stability of the clusters by avoid reconstructions of the clusters when two cluster heads are located within D hops. The high overhead is the major drawback of this algorithm. In [13], a Vehicular Multi-Hop Algorithm for Stable Clustering (VMSC) is proposed. In this algorithm, the least mobility vehicle is selected as a cluster head to provide the clustering stability. To do that vehicles calculate the average speed of the vehicles within D hops. The drawback here is the high overhead as in the previous one. In [14] a Vehicular Deterministic medium Access controls (VDA). It schedules transmissions up to two hops, therefore, decreases the transmission delay and the collisions. In [15], a D -Hops Clustering Vehicles (DHCV) is proposed. In this algorithm the vehicles are grouped into D hops clusters according to location and speed differences between the vehicles within its D range. Each vehicle has a GPS to obtain its speed and location and to broadcast these data by WAVE standard to its neighbors within D hops. After the cloud construction, the cluster head manages the transmission scheduling inside the cluster by mathematical optimization model. DHCV provides better performance than the previous

algorithms, due to, its stability and the usage of both physical layer and Medium Access Control (MAC) layer to schedule transmissions.

The main drawbacks of DHCV, it cannot work with real time applications and urgent data, due to its cluster-based model, and as known, one of the characteristic of VANET is the frequent re-clustering due to the mobility of vehicles. Another case, when the vehicle has a huge data, the model does not achieve a high throughput with low delay. There are other drawbacks that DHCV did not consider such as, the availability of the CH and the vehicle capabilities.

Finally, in this paper some improvements are proposed to DHCV, first include the vehicle capability into consideration when electing the cluster head. Second, I propose a merging technique of IEEE802.11p and Long-Term Evolution (LTE) to transmit real time data. Third, let the vehicles which have a lot of data be CM in two clusters to ensure its delivery and to collaborate among the nodes.

III. SYSTEM MODEL

An Improved Hybrid model in Vehicular Clouds based on Data Types (IHVCDT) is the unique model which checks the data type and according to this data it will decide which mode to operate.

A. System Modes

As shown in Fig. 1, we have three different modes which classified according to the type of data.

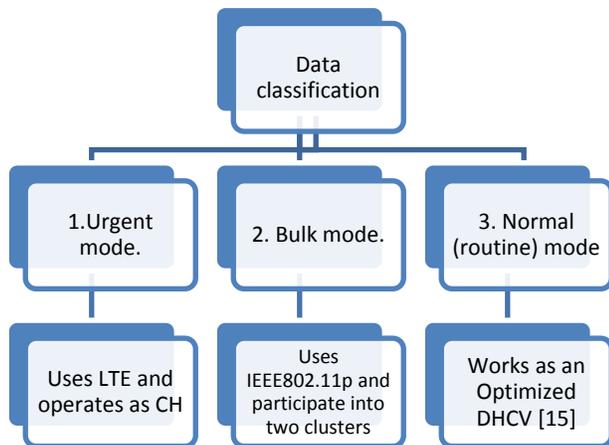


Fig. 1. The system modes.

1) The Urgent Mode

This mode operates if the data is urgent such as, huge accident, earthquake, and road crashes. The most important metric to be concerned here is the delay so we must deliver the data with the minimum delay. To obtain this goal, the vehicle that has an urgent data will assume itself as a CH or a separate vehicle and transmit this data directly to the nearest base station using LTE.

2) The Bulk Mode

If the data is huge such as long video, and is not urgent to be sent, the vehicle chooses a CH that has an efficient capabilities including:

- **Availability:** It means whether this CH has the ability to accept this CM or not.
- **Bandwidth:** It is the data rate which is based on the following factors: interfaces of the devices, transmission medium, the weather, and the service provided by the Internet Service Provider (ISP).
- **Vehicle capabilities:** It is the vehicle can handle all of requests, the size of the buffer it has, the cost of transmissions.

In this case, the most optimization here is the cost. The vehicles choose two CHs according to the previous factors. After cloud creation, the vehicle will be a CM in two clusters, and divide the data between them. The transmission of data to the CH is done by IEEE 802.11p. The equation to select the CH according to transmission cost is as follows:

Let C_{xy} , denotes the cost difference between X and Y.

$$C_{xy} = |C_x - C_y| \quad \text{where } y \in N(x) \quad (1)$$

Where C_x and C_y present the transmission cost of nodes X and Y.

3) The Normal (Routine) Mode

This mode operates when the data is small and not urgent such as advertisements, fuel stations locations, hotels, etc.

This mode operates as an optimized DHCV [15], where the clusters are created based on the distance differences between the vehicles and after the cloud creation a mathematical optimization managed by the CH controls the transmissions from the CM to CH and from CH to base stations.

IV. PERFORMANCE EVALUATIONS

In this section, the expected performance evolutions are presented. The expectation results obtained from this research are compared with these simulation results that obtained in DHCV [15]. The simulator was NS2 to evaluate the proposed DHCV. The transcendence choice of the vehicles is considered as the velocity limit, the distance between the vehicle in front of it, and acceleration of the vehicles.

The authors of DHCV [15] compared the optimization scheduling technique between CSMA/CA and VDA [14]. Based on throughput and delay, VDA is a deterministic model to access the medium that schedules the transmission in contention free durations up to two hops. Two VANETs are used, one is low density which has 2 vehicles per km, the other is high density with 12 vehicles per km. Here, all the vehicles have the same transmitting power, which is 5 mW. The other parameters are listed below in Table 1.

TABLE I. SIMULATION PARAMETERS

Parameters	Value
Propagation model	Nakagami
System bandwidth	10 MHz
Message payload size	500 byte
MAC and PHY	802.11p
Noise power density	-131 dbm
Raw bitrate	1 to 6 Mbps
Modulation	BPSK 1/2
Simulation time	10 sec
Vehicle speed	40 km/h

To evaluate the performance, the following metrics are considered:

- 1) The average data throughput: is the average data that is received from the CMs to the CHs.
- 2) The average delay: is the average time needed from sending the message from CMs to CHs.

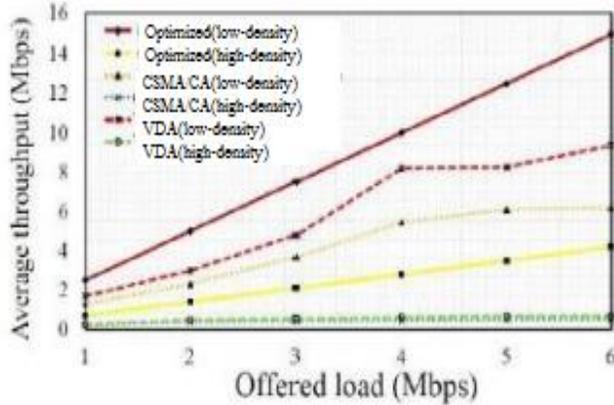


Fig. 2. Average throughput.

A. Throughput

Fig. 2 presents the average throughput with the different loads. As shown in Fig. 2, as the load increases the throughput increases. Due to the fact that as the load increases the data delivery will also increase. According to the results, the optimized DHCV has the best delivery in all scenarios due to many facts: the transmission links are optimally scheduled to decrease the contentions, the CH is responsible about the transmissions inside the cloud, and the hidden node is addressed in the optimized DHCD, therefore, it obtains the best result among the compared models [15].

B. End to End Delay

Fig. 3 presents the average end-to-end delay with different loads. As shown in Fig. 3, as the load increases the delay decreases. According to these results, the optimized DHCV model has the lowest delay in all scenarios due to the fact that the optimized DHCV solves the hidden nodes and interference problems, by CH which considers the physical condition of the medium [15].

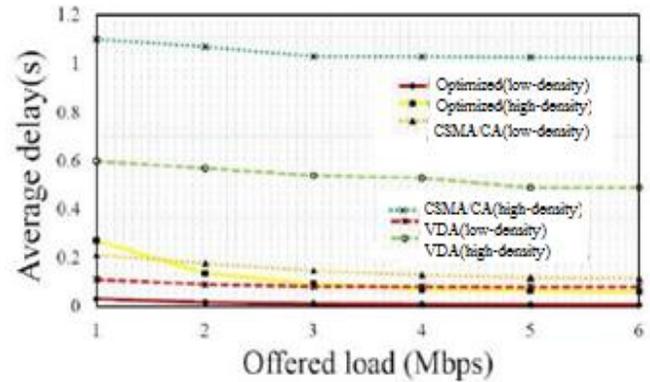


Fig. 3. Average end to end delay.

If the simulation results [15] are compared with IHVCDT, best performance is obtained in the urgent case and in the bulk case, where in the normal case the same results of the optimized DHCV will be obtained. These perspectives are considered from the advantage of using LTE instead of IEEE 802.11 p in the first case, and of the participation of the vehicle into two clusters as in the second case. While, in the third case, the vehicles operate as in the optimized DHCV [15], and according to their results, it has the best throughput. And the minimum delay ever. As a result, if the normal case has the best results then our improved model will be the best of all.

V. CONCLUSION

The model presented in this paper is an Improved Hybrid model in Vehicular Clouds based on Data Types (IHVCDT). This model expected to provide the best delivery ratio within the minimum latency. The model is based on the data classification. The model is classified into three cases; the first one, is the urgent mode where data is urgent, and therefore, the concern here is to obtain a high data delivery with low latency. The vehicle which has an urgent data assumes itself as a cluster head or a separate vehicle and sends the data directly to the cloud station (base station) by using LTE. The second case, is the bulk mode, where the data is huge where the vehicle chooses the most efficient CH which has an efficient capability such as; the processor units in its interface, bandwidth, availability, and its transmission cost. All these factors should be considered to select the CH. In this case, the vehicle can be participated in two clusters to divide this huge data between the clusters to insure its delivery. The transmission here is based on IEEE 802.11p. The third case is the normal (routine) mode, where the data is small and not urgent. Here, the vehicle chooses the CH based on the distance as in DHCV [15]. This mixture produces a new hybrid model that provides the advantages of LTE and IEEE 802.11p.

The main drawback of this model is that each vehicle should have two interfaces for transmission, one for LTE, and one for IEEE 802.11 p.

As future work, as the mobility in VANETs can be predictable, we propose to make the CH selection is based on the vehicles mobility in addition to the proposed factors, this provides another QoS model, where vehicles can send a prerequisite to a specific CH to become its CH.

REFERENCES

- [1] Ahizoune, Ahmed, and Abdelhakim Hafid. "A new stability based clustering algorithm (SBCA) for VANETs." *Local Computer Networks Workshops (LCN Workshops), 2012 IEEE 37th Conference on*. IEEE, 2012.
- [2] Bana, Soheila V., and Pravin Varaiya. "Space division multiple access (SDMA) for robust ad hoc vehicle communication networks." *Intelligent Transportation Systems, 2001. Proceedings. 2001 IEEE*. IEEE, 2001.
- [3] Borgonovo, Flaminio, et al. "ADHOC MAC: new MAC architecture for ad hoc networks providing efficient and reliable point-to-point and broadcast services." *Wireless Networks* 10.4 (2004): 359-366.
- [4] Su, Hang, and Xi Zhang. "Clustering-based multichannel MAC protocols for QoS provisionings over vehicular ad hoc networks." *IEEE Transactions on Vehicular Technology* 56.6 (2007): 3309-3323.
- [5] Mammu, Aboobeker Sidhik Koyamparambil, Unai Hernandez-Jayo, and Nekane Sainz. "Cluster-based MAC in VANETs for safety applications." *Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on*. IEEE, 2013.
- [6] Mu'azu, Abubakar Aminu, et al. "A QoS approach for cluster-based routing in VANETS using TDMA scheme." *2013 international conference on ICT convergence (ICTC)*. IEEE, 2013.
- [7] Almalag, Mohammad S., Stephan Olariu, and Michele C. Weigle. "Tdma cluster-based mac for vanets (tc-mac)." *World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2012 IEEE International Symposium on a*. IEEE, 2012.
- [8] Dror, Efi, Chen Avin, and Zvi Lotker. "Fast randomized algorithm for hierarchical clustering in vehicular ad-hoc networks." *Ad Hoc Networking Workshop (Med-Hoc-Net), 2011 The 10th IFIP Annual Mediterranean*. IEEE, 2011.
- [9] Wolny, Grzegorz. "Modified DMAC clustering algorithm for VANETs." *2008 Third International Conference on Systems and Networks Communications*. IEEE, 2008.
- [10] Basagni, Stefano. "Distributed clustering for ad hoc networks." *Parallel Architectures, Algorithms, and Networks, 1999.(I-SPAN'99) Proceedings*. Fourth International Symposium on. IEEE, 1999.
- [11] Chen, Yuzhong, et al. "Distributed multi-hop clustering algorithm for VANETs based on neighborhood follow." *EURASIP Journal on Wireless Communications and Networking* 2015.1 (2015): 1-12.
- [12] Zhang, Zhenxia, Azzedine Boukerche, and Richard Pazzi. "A novel multi-hop clustering scheme for vehicular ad-hoc networks." *Proceedings of the 9th ACM international symposium on Mobility management and wireless access*. ACM, 2011.
- [13] Ucar, Seyhan, Sinem Coleri Ergen, and Oznur Ozkasap. "VMaSC: Vehicular multi-hop algorithm for stable clustering in vehicular ad hoc networks." *2013 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2013.
- [14] Rezgui, Jihene, Soumaya Cherkaoui, and Omar Chakroun. "Deterministic access for dsrc/802.11 p vehicular safety communication." *2011 7th International Wireless Communications and Mobile Computing Conference*. IEEE, 2011.
- [15] Azizian, Meysam, Soumaya Cherkaoui, and Abdelhakim Hafid. "An Optimized Flow Allocation in Vehicular Cloud." *IEEE Access* 4 (2016): 6766-6779.
- [16] Ucar, Seyhan, Sinem Coleri Ergen, and Oznur Ozkasap. "Multihop-Cluster-Based IEEE 802.11 p and LTE Hybrid Architecture for VANET Safety Message Dissemination." *IEEE Transactions on Vehicular Technology* 65.4 (2016): 2621-2636.

FPGA Implementation of SVM for Nonlinear Systems Regression

Intissar SAYEHI

University of Tunis Elmanar, Faculty of Mathematical, Physical and Natural Sciences of Tunis
Laboratory of Electronics and Microelectronics, (E. μ . E. L),
FSM, Monastir, Tunisia

Mohsen MACHHOUT

University of Monastir, Faculty of Sciences of Monastir
Laboratory of Electronics and Microelectronics (E. μ . E. L)

Rached TOURKI

University of Monastir, Faculty of Sciences of Monastir
Laboratory of Electronics and Microelectronics (E. μ . E. L)

Abstract—This work resumes the previous implementations of Support Vector Machine for Classification and Regression and explicates the different methods and approaches adopted. Ever since the rarity of works in the field of nonlinear systems regression, an implementation of testing phase of SVM was proposed exploiting the parallelism and reconfigurability of Field-Programmable Gate Arrays (FPGA) platform. The nonlinear system chosen for application was a real challenging model: a fluid level control system existing in our laboratory. The implemented design with fixed point precision demonstrates good enough results comparing with the software performances based on the Normalized Mean Squared Error. Whereas, in term of computation time, a speed-up factor of 60 orders of time comparing to MATLAB results was achieved. Due to the flexibility of Xilinx System Generator, the design is capable to be reused for any other system with different data sets sizes and various kernel functions.

Keywords—Machine learning; nonlinear system; SVM regression; Reproducing Kernel Hilbert Space (RKHS); MATLAB; Field-Programmable Gate Arrays (FPGA); Xilinx System Generator

I. INTRODUCTION

The support vector machine is a machine learning created by Vapnik at the 60's. It was created first for classification tasks then extended to regression. The most important advantage in this method that is applicable to different fields are like medicine biology, signal processing, sensor networks, computer sciences, etc.

The difficult challenge in the use of the SVM method is to compromise between the model performances and the data sets size. There from the need to hardware platforms that accelerate the computation time and provide a flexible support for classifying or regressing new systems.

This paper treated the previous hardware implementations of support vector machine on Field-Programmable Gate Arrays (FPGA) for classification and regression and explains the different approaches adopted for developing the SVM

architecture. The FPGAs devices offer many advantages like concrete development tools, simple reprogram ability and quick development time. Furthermore, parallelism can be attained, that is a benefit above other devices, like microcontrollers and DSPs.

The majority of implementations of SVM were targeted to classification task for simple and specific problems. However the SVM regression task still abandoned and neglected.

In this present, we propose a hardware design on FPGA for nonlinear systems regression.

The arrangement of the article is as pursues. In the second section the theoretical basis of nonlinear systems identification in the Reproducing Kernel Hilbert Space (RKHS) space was described. In Section 3, the Methods of SVM implementations on FPGA were presented with the related works. In Section 4, the FPGA designing tools were explained. In Section 5 our SVM Implementation approach and the Parameters Selection were presented with results and plots for the regression of fluid level control system. Finally we compare our work to similar ones and conclude with Conclusion.

II. NON LINEAR SYSTEMS IDENTIFICATION IN REPRODUCING KERNEL HILBERT SPACE

The identification of linear systems is accomplished via mathematical representations on the bases of vibration measurements. Thus, the relation between the input and output of a system, called transfer function, stays stable at all excitation levels. Accordingly, the mathematic model acquired at one operating point can be generalized for predicting the system behavior at another operating point. Whereas, it is not the same case for nonlinear systems because it is hard and complicated to find a general mathematical model describing the system by relying on the system identification only at a particular excitation level.

The difference between linear and non-linear systems can be explicated by Fig. 1, that for non-linear systems the transfer function is not independent of the input.

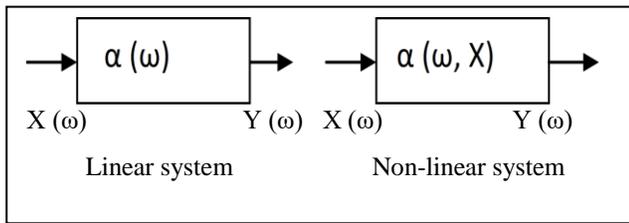


Fig. 1. Distinction between linear and nonlinear system.

It exist many types of nonlinear models like state-space models, Input/output-models, block-oriented models, etc. it didn't exist a nonlinear universal model suitable for all appliances but it depends on each one. Consequently, diverse approaches for identifying and modeling nonlinear systems were conceived. Essentially two categories can be distinguished: parametric models and nonparametric models. The next figure clarifies the different constitutions of these categories.

Another part of researchers were also interested by the system identification field by creating diverse techniques in machine learning allowing nonlinear systems identification like k-Nearest Neighbors and Regularization networks [1]-[3].

In next paragraphs, we introduce the mathematical foundation of learning machine and explain the functionality of support vector machine method.

A. Statistical Learning Theory (SLT)

The goal of the Statistical Learning Theory [4] is to obtain a function f modeling a given system since a set of observed data $O = \{(x_i, y_i)\}_{i=1}^N$ composed of inputs x_i and outputs y_i . This function has to repeat the process behavior by diminishing the functional risk presented by this expression:

$$R(f) = \int_{X,Y} V(y, f(x))P(x, y)dx dy \tag{1}$$

The expression $V(y, f(x))$ is named cost function. It computes the deviation between system output y_i and the estimated output $f(x)$. The couple (X, Y) is composed of random vectors and of the independent samples (x_i, y_i) . The risk $R(f)$ can't be approximated caused by unknowing $P(x, y)$. To resolve this problem we have to ease the following expression:

$$R_{emp}(f) = \frac{1}{N} \sum_{i=1}^N V(y_i, f(x_i)) \tag{2}$$

However, minimizing frankly $R_{emp}(f)$ in the functions space H don't give the best approximation of $R(f)$ minimization and could guide to over fitting. As a resolution, Vapnik presented the theory of structural risk minimization (SRM). It penalizes empirical risk via a function that approximates the complexity of the retained model.

This guides to minimizing the constraint defined by the following expression:

$$\min_{f \in H} D(f) = \min \left(\frac{1}{N} \sum_{i=1}^N V(y^{(i)}, f(x^{(i)})) + \lambda \|f\|_H^2 \right) \tag{3}$$

The first word measures how the function f fits a given data and the second word is the squared norm of f in the RKHS space H that controls the complexity (smoothness) of the solution. The parameter λ is the regularization parameter that equilibrium the tradeoff among the two terms.

The regularity of the solution is most important and not the value of λ . Whereas it is not evident to minimize the constraint (3) on whichever arbitrary function space H , whatever is it with finite or infinite dimension. Therefore, to overcome this difficulty, we will consider the space H as a RKHS.

B. Reproducing Kernel Hilbert Space (RKHS)

We suppose that X the random variable is evaluated in the space $E \subset \mathbb{R}^d$ and we suppose that exists a function K called kernel function: $K : E^2 \rightarrow \mathbb{R}$. It is symmetric and positive definite. In this case, there is a function $\phi : E \rightarrow H$ that:

$$K(x, x') = \langle \phi(x), \phi(x') \rangle_H \tag{4}$$

H is the Reproducing Kernel Hilbert Space (RKHS) [5] of kernel K . This space has some rigorous properties:

- ✓ $\forall x \in E$ et $f \in H$ $\langle K(x, \cdot), f \rangle_H = f(x)$ (5)
- ✓ Due to represented theorem the resolution of the optimization problem presented by (3) in this space is given by :

$$f_{opt} = \sum_{i=1}^N a_i K(x_i, \cdot) \tag{6}$$

TABLE I. KERNEL FUNCTIONS

Kernel function	Mathematical expression	Parameters
Linear kernel	$K(x, x') = x \times x'$	-
Polynomial kernel	$K(x, x') = (1 + \langle x, x' \rangle)^n$	$n \in \mathbb{N}^*$ and $\langle x, x' \rangle$ is an Euclidian scalar product.
Sigmoid kernel	$k(x, y) = \tanh(\alpha \cdot x^T \cdot y + c)$	α and c are adjustable
Radial Basis Function (RBF) kernel	$K(x, x') = e^{-\frac{\ x-x'\ ^2}{2\sigma^2}}$	σ is a real positive parameter
Laplacian kernel (ERBF, Extended RBF)	$K(x, x') = e^{-\gamma \ x-x'\ }$	γ is a real positive parameter

It exist a variety of kernel functions that could be considered in Table 1.

C. Support Vector Machine for regression

SVM is a supervised learning model founded on the Vapnik and Chervonenkis learning theory. It was first developed for classification problems then extended to regression tasks.

For SVM regression, the goal is to find a model for the data set $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ that matches the input x_i to the real output y_i (with $x_i \in \mathbb{R}^l$ and $y_i \in \mathbb{R}$).

By resolving the following quadratic programming problem with linear restrictions and an ϵ -insensitive loss function:

$$\min_{\alpha_i} \text{imise } \frac{1}{2} \sum_{i,j=1}^n (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) K(x_i, x_j) + \epsilon \sum_{i=1}^n (\alpha_i^* + \alpha_i) - \sum_{i=1}^n (\alpha_i^* - \alpha_i) y_i \quad (7)$$

$$\text{St : } 0 \leq \alpha_i^*, \alpha_i \leq C \text{ for } i=1, \dots, n \quad \sum_{i=1}^n (\alpha_i^* - \alpha_i) = 0$$

where $K(x_i, x_j)$ is a kernel function, C is the regularization term and ϵ is a positive constant presenting an insensitive region in the interior of which the training errors are unseen. C and ϵ are predefined constants. The feed-forward evaluation function of a new, unlearned vector x is:

$$y(x) = \sum_{i=1}^{Nsv} (\alpha_i^* \alpha_i) K(x_i, x) + b \quad (8)$$

The parameters α_i , α_i^* and b are calculated in learning phase. As in the classification model, just the resulting support vectors are used in the feed-forward phase.

III. METHODS OF SVM IMPLEMENTATION ON FPGA

The related hardware implementations of SVM model on FPGA was accurately reviewed. In this paper we are focused in certain class of SVM implementations that implemented just the testing phase on FPGA. Unfortunately those for SVM regression were rare and exceptional. Most of designers needed to implement classifier for different applications whereas we found only three works [6]-[8] that implement a design for both classification and regression. The SVM for regression has the same importance as SVM for classification but is infrequently employed due to the complexity of the feed forward function. Next paragraphs give a consistent recap of different techniques and architectures employed for SVM implementation on FPGA.

A. Parallel Systolic Array Architecture

In different fields of sciences, the operations involving important linear system of equations like matrix algebra are indispensable. Consequently, the need of fast and speedy computers equipped with efficient software programs is crucial and increasing. Whereas, the main disadvantage of a general-purpose computer is the limited memory space for big matrices computing. To avoid this problem, novel methods and

approaches have to be invented to benefit simultaneously of highly parallel computational machines.

The solution was the association of a big number of same processing elements (PEs) that rhythmically compute and pass data to neighboring connected PEs. The produced set of well ordered PEs connections corresponds to the Systolic architecture that can be arranged in a linear or two-dimensional array with rectangular or hexagonal geometry.

The systolic array could be employed as a coprocessor combined with a host computer that pass data through the PEs and the final result is came again to the host computer. As in Fig. 2, this operation is similar to the flow of blood throughout the heart called "systolic".

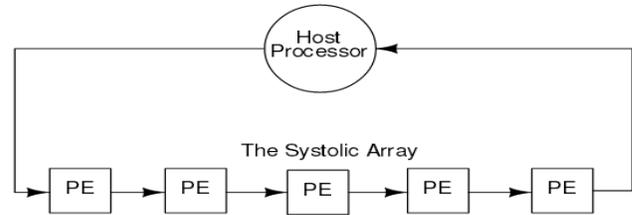


Fig. 2. Systolic system mechanism.

This arrangement is very suitable for VLSI technology that offers an exceptionally high operating with low cost array of speedy computational processors. It was broadly implemented on FPGA to attain high levels of parallelism, that was exploited by various SVM implementations.

R. Patil et al. [9] employed this architecture for implementing a SVM multiclass classifier. The hardware was Xilinx Virtex-6 FPGA for the recognition of facial expression. The training phase computed by MATLAB. Thanks to a power optimization of the FPGA design based on the difference-based partial reconfiguration technique, the power decrease up to 3 to 5% was attained by using Xilinx EDA tool.

C. Kyrkou et al. developed an SVM classifier for object detection based on systolic array architecture [10]. The design can be expanded and adjusted to convene multiclass classification and various applications. Many tasks of object detection like face, walker, and car were done. Simulation consequences proved a high performance of 40, 46, 122 fps for three applications, with no precision loss in comparison with the precision of software detection of the SVM model executed in MATLAB (77, 76, 78%).

B. Multiplier-less Approach

The Multiplier-less techniques were needed to diminish the implementation cost because the multipliers are the mainly costly blocks in term of surface occupation. Therefore, many researchers have hardly worked to make the multiplication simpler and faster by applying the fact that multiplication by a power-of-two could be achieved by simple shift and add operations. The number of these operations depends on the design restrictions. There are a number of conventional representations to speed up multiplication. One is by reducing the number of operands to be added; the other is by adding the operands faster (accelerating accumulation) [11]. Most designers combine and profit from the two methods to reduce

significantly the number of operands employed in the hardware. In the work [12], the authors benefit from this technique and aiming a diminution in hardware complexity and power consumption by implementing simplified multiplier-less kernel using shifts and add operations as an alternative to traditional vector product kernel for classification. The implementation of SVM classification was performed on the modern Xilinx Virtex-7 FPGA.

They presented a different approach of applying the CSD and CSE representation methods for vectors data to decrease the number of needed adders to reduce the hardware complexity. Three classifiers were implemented to compare it against other three implemented classifiers using the conventional vector product kernel. The power reductions of 1, 2.7, and 3.5% were achieved by the proposed CSD-based multiplier-less kernel against the vector product kernel relating to resources utilization.

C. Dynamic Partially Reconfiguration

There are two methods to modify the Hardware functionality on FPGA. First, the Static reconfiguration which consists to shutting down the application then downloading the new configuration and restarting the implementation. Second, the Dynamic Reconfiguration permits varying hardware functionality on FPGA without taking the purpose offline. This offers a flexibility to adjust the hardware online and to gain a lot of time. The modification can be either total or partial according to the need of designer. In total reconfiguration the configuration bitstream, affords the information concerning the chip and it arranges whole FPGA. In partial reconfiguration, just a part of the platform is reconstituted, whereas the rest maintain operating securely with respect to the development procedure. The Dynamic Partially Reconfigured approach (DPR) grants the modification on a selected section of the FPGA while the other sections stay working without necessitating to turning off. This great improvement is excellent for real time embedded systems when the shutting down of the system is expensive and detriment during system runtime. Moreover, DPR diminished significantly power consumption and decreasing reconfiguration time.

This practice was utilized by H. Hussain et al. [13] for implementing SVM classifier for bioinformatics applications. It was implemented on an old FPGA panel; Xilinx ML403, where the kernel calculation was implemented using two pipelined stages. An acceleration up to 85x was accomplished above a corresponding GPP software execution using MATLAB bioinformatics toolbox DPR was applied to change the diverse parameters of SVM, it was 8x more rapidly than reconfiguring the entire of FPGA.

D. Common Pipelining Technique

The pipelining technique is a technique implementing a form of parallelism with a single processor. It accelerates the central processing unit throughput at a certain clock rate by performing multiple operations at the same time. The fundamental training cycle is broken in a series named a pipeline. Pipelining searches to let the processor works on as many instructions as there are dependent steps. It augments instruction throughput however does not diminish the required time to end one instruction.

The researchers in works [14], [15] wanted to compare the performances of FPGA and GPU implementations of a human skin SVM classifier against the software performances. The critical hardware composes of FPGA were designed using HDL in a completely pipelined organization, even as the other elements like FIFO and interfaces were implemented in HLL. The implementation results confirmed the excellence of the implemented fully pipelined FPGA architecture on GPU and CPU for a small number of image pixels, while the GPU implementation was the fastest for a big number of pixels. The advantage of FPGA implementation is that consumes less power than the GPU implementation [15].

Y. Ago et al. [16] employed a new fully pipelined DSP architecture on FPGA for accelerating SVM classification. The proposed design was executed on Xilinx Virtex-6 FPGA with different types of kernel functions; sigmoid, polynomial, and RBF kernels. Consequently, an important throughput of 2.89x10⁶ times per second for classifying 128-dimension feature space running at 370.096 MHz was obtained. Other implementations intended to develop the common pipelined fashion for accomplishing powerful designs.

Whereas, a combined circuit was designed in a parallel architecture with two-stage pipeline for linear and non-linear SVM classification [17] in order to diminish the circuit size by sharing multipliers and adders necessary for inner product computation. The proposed circuit was synthesizing with 65nm standard cell library, representing 661,261 gates with 152 MHz maximum operating frequency. Moreover, high performance was attained from handing out up to 33.8 640x480 image frames per second.

E. CORDIC Algorithm

The CORDIC algorithm is a fundamental iterative algorithm using a fixed vector rotation technique to calculate sequentially the trigonometric functions. The entire conception orbits around employing just a simple shifter and adder to simplify the implementation of the CORDIC algorithm for computing complex functions. The CORDIC algorithm was originally presented by J.E. Volder [18] for implementing fundamental mathematical functions like the multiplication, division and trigonometric functions. It was helpful for diverse domains like neural networks, video and image processing, etc. For majority of applications the CORDIC algorithm offers a speed-up of time and reduction of power consumption. SVM method for both classification and regression were implemented by M. Ruiz-Llata et al. [6] on FPGA using the CORDIC iterative algorithm. The implemented system consumed 3/4 of the FPGA logic (Cyclone II) and an exterior memory was used for storing support vectors leading to 2ms limitation in the classification speed, with an error rate of 4%.

Another FPGA implementation of fast SVM presented by J. Sarcia et al. [19] based on CORDIC algorithm for kernel calculations. The implemented system achieved speed improvement over their previous CORDIC circuit implemented in [20] with a factor of 6, with limited hardware resources utilization.

IV. FPGA DESIGNING TOOLS

Field-Programmable Gate Arrays (FPGAs) is composed of configurable logic blocks (CLBs) that can be reprogrammed to realize different functions in few seconds. The flexibility offered by the FPGA goes with the increasing programming complexity. Consequently there is a critical necessity for high level fast prototyping systems that can help designers and eases the mapping from algorithm to hardware. The algorithms are classically written and tested via MATLAB code or Simulink model based environment, and there are a number of tools that convert such algorithms to a hardware description language such as AccelDSP Synthesis Tool from Xilinx, Simulink HDL Coder from Mathworks, C-based High Level Design Tools and System Generator for DSP. Briefly, these tools are explained at the next paragraph.

AccelDSP Synthesis Tool is a high level tool particularly designed for Digital Signal Processing (DSP) applications. The principle of AccelDSP is to translate a MATLAB floating-point design into a hardware implementation that objects FPGAs. AccelDSP automatically creates bit-true and cycle-accurate HDL codes which are complete to synthesize, implement and map onto FPGA hardware. AccelDSP generated designs result in inefficient architectures in terms of area and timing compared to hand-coded results.

The Simulink HDL Coder is a high level design tool which generates HDL code from Simulink models and State flow finite state machines. It can provide also interfaces to combine manually-written HDL codes, HDL Co-simulation blocks and RAM blocks in its environment. Whereas, not the whole Simulink included blocks are supported. Embedded MATLAB Function Block has its own limitations and do not support all the operations such as nested functions and use of multiple values in the left side of an expression.

The design C-based high level design tools [21] are used for automatic hardware generation offering a quicker path to hardware with a low cost comparing to traditional methods. It expresses parallelism through variations in C (pseudo-C) or compiler or both. Ideally, it is best to use pure ANSI-C without any variation in C and exploit parallelism through compiler that ports C code into hardware; therefore a user does not need to learn a new language.

System Generator is a high level design tool designed by Xilinx to be used in model-based design environment and implemented in FPGAs. Simulink provides a powerful component based computing model including several different blocks to be connected together by the user for designing and

simulating functional systems. System Generator provides similar blocks which are used and connected the same way Simulink blocks does but target FPGA architectures to design discrete time systems which can be synchronous to a single or more clocks. The simulation results of the designed systems are bit and cycle accurate which means simulation and hardware results are exactly match together. System Generator is the best tool provided for MATLAB code environment because it's a "push button" transition from specification to implementation.

V. IMPLEMENTATION OF SVM REGRESSION

A. Approach and the Parameters Selection

A variety of practices for hardware implementations have been developed for improving the online SVM testing phase on different FPGA devices. The idea is to train SVM model first offline on software (MATLAB), and then the trained data are extracted to be used for online regression on hardware. Only the resulting support vectors are used in the feed-forward phase. The feed-forward estimation function of a new, non-learned vector x is:

$$y(x) = \sum_{i=1}^{N_{sv}} (\alpha_i^* \alpha_i) K(x_i, x) + b \quad (8)$$

Where parameters α_i , α_i^* and b are given in the learning phase. $K(x_i, x_j)$ is the kernel function that can be polynomial functions or Gaussian functions. Our system uses the polynomial function because this kernel significantly simplifies the SVM feed-forward phase computation in constrained hardware while conserves good classification performance with respect to the system nonlinearities.

The parameters to be fixed prior the training step are the parameter of the kernel which is the degree of polynomial kernel, the regularization parameter C and the ϵ parameter of the ϵ -insensitive loss function. All these parameters are chosen to be used in the fixed-point arithmetic. To locate C , and the ϵ parameter in regression, we use an iterative training strategy with the goal of minimizing errors while keeping a reduced number of support vectors.

The basic hardware architecture to perform (8) is represented in Fig. 3.

The inputs parameters are: the testing vector and the support vectors. The calculation of the prediction function is realized through a kernel function processing block that acquires the input parameters and then calculates the Gramian matrix to be multiplied by the Lagrange Multipliers.

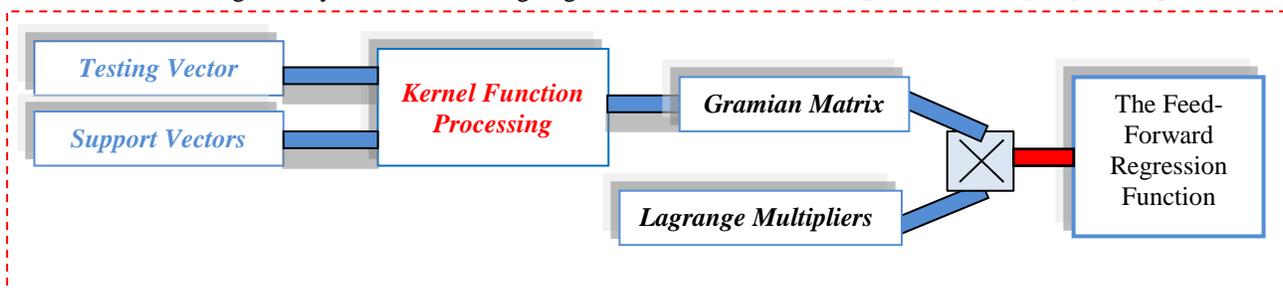


Fig. 3. Hardware architecture of SVM testing phase.

The Gramian matrix $G \in \mathbb{R}^{N \times N}$ is like that:

$$G_{ij} = (K(x_i, x_j)), i, j = 1, \dots, N \quad (9)$$

N is the number of observations and K is the kernel function. It can be selected as sigmoid or polynomial kernel. The computation of Gramian matrix is done in the training and testing phase. As the input vector X can be 1-Dimensional or 2-Dimensional Array, we suggested a streaming Approach to calculate the Gramian matrix.

This approach exploits the FPGA parallelism to automatic compilation of software programs into hardware. Effectively, three fundamental approaches are distinguished to automatic compilation of software into hardware.

First approach is to find an accessible parallel programming model. Then programmers map a program written in it onto hardware [22]. This approach permits to establish parallelism, however many problems like synchronization, deadlocks and starvation have to be arranged.

A different approach, the behavioral synthesis compilers those investigate programs written in a high-level sequential language, for example C, and challenge to extort instruction-level parallelism by analyzing dependencies in the middle of instructions, and mapping in reliant instructions to parallel hardware components. Various compilers have been accomplished, like C2H from Altera that is entirely incorporated inside their SOPC design flow. The major difficulty with such approach is that the overall of instruction-level parallelism in a classic software program is limited. Therefore, constantly have to reorganize their code and without a doubt control hardware resources, like mapping of data to memory units.

The third approach is to exploit a sophisticated language that permits to the programmers to express parallelism without be troubled about synchronization and associated matters. This type of languages is based on the streaming paradigm articulated on data that are collected into streams [23] similar to arrays, but with a mutually independency between the elements. In our work this approach is entirely used to execute all the testing phase on the FPGA.

In next paragraph the experimental process is described with explanation of the implementation steps.

B. Description of the Fluid Process and Results

The process subject to regression is a fluid level control system consisting of two cascaded tanks with free outlets fed by a pump. The water is transported by the pump to the upper of the two tanks. The process is depicted in Fig. 4.

The input signal to the process is the voltage applied to the pump and the two output signals consist of measurements of the water level of the tanks. Since the outlets are open, the result is a dynamics that varies nonlinearly with the level of water. The process is controlled from a PC equipped with MATLAB interfaces to the A/D and D/A converters. All data was collected in open loop experiments using zero-order hold (ZoH) sampling. The data was recorded from the cascaded tanks and collected in a data file. The sampling period of 4.0 s

provides 7500 samples of input-output data for both the upper and lower tank.

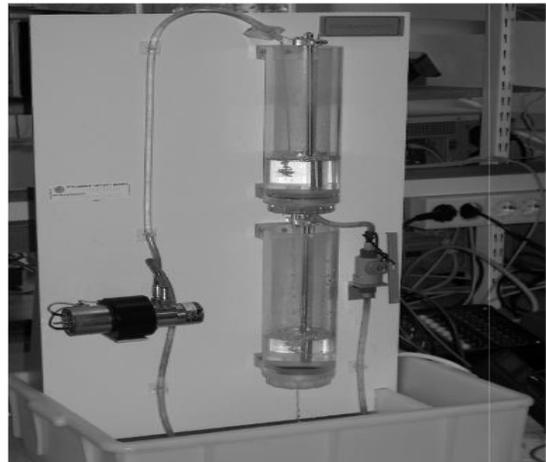


Fig. 4. Fluid process composed of two cascaded tanks.

To construct the SVM regression model for the fluid level control system, 2000 observations taken for the training phase and the validation phase was performed on 1000 new observations.

Firstly, the optimal parameters for training phase like the regularization parameter C and the ϵ -insensitive loss function are obtained using MATLAB with an iterative training approach to reduce computing errors while maintaining a reduced number of support vectors. After this preparation, it is possible to prove the efficiency of the modeling parameters by testing it with novel data. The type of kernel used: polynomial and sigmoid. The optimal parameters λ and σ of the machine learning are mentioned in Table 2:

TABLE II. PARAMETRERS SELECTION

Parameters Method	Kernel function	Optimal parameter	Trade-off parameter	ϵ -insensitive loss function
SVM Regression	Polynomial	$\sigma=3$	$C=100$	$\epsilon = 0.01$
SVM Regression	Sigmoid	$\alpha =10, c=1$	$C=1000$	$\epsilon = 0.01$

Secondly and according to this table, the resulting support vectors were six (for polynomial kernel). The basic hardware architecture to perform this equation:

$$y(x) = \sum_{i=1}^{Nsv} (\alpha_i^* \alpha_i) K(x_i, x) + b$$
 is represented in Fig. 5. It is

composed of input vector X , Nsv parallel support vectors SV blocks, a kernel processing block, a Gramian matrix block, Lagrange multipliers block, FIFO buffers and memory buffers.

The testing vector X is passed through the FIFO to be streamed and then calculated by the kernel functions and support vectors. To benefit from data parallelism, the number of streams is identical to the number Nsv of suport vectors because stream elements are independent. The number of streams is often limited by the accessible hardware resources.

The objective of our research is to benefit from the powerful parallelizing of FPGA and the flexible software programming. The design user has only to choose the system to be predicted and the kernel type suitable to the application, and then in offline calculate the support vectors that will be used for feed forward function on hardware.

In Fig. 5, the system generator project for implementing the testing phase is presented.

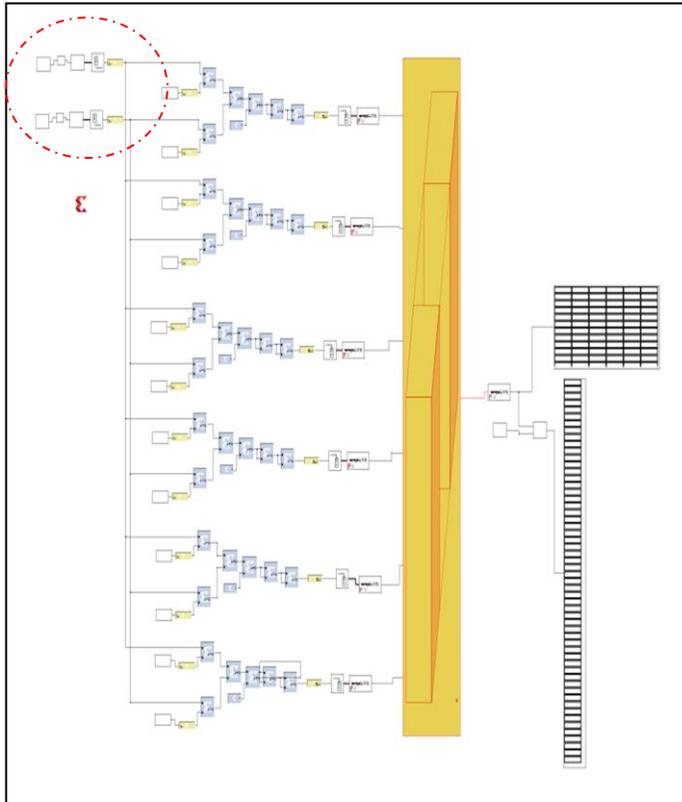


Fig. 5. System generator project for computing the testing phase of SVM.

As mentioned in Fig. 6 the regression vector is 2d-vector: $x(i) = (y(i-1), y(i-2))$. The computation of this type of vectors is generally difficult in hardware. Consequently, the streaming approach is used for this vector. The two columns of

the regression vector are streamed in parallel and simultaneously computed by the kernel processing function and support vectors. The kernel function is polynomial with third order. Then the result passed through the yellow block responsible of the concatenation of Gramian matrix elements.

Finally, the equation $y(x) = \sum_{i=1}^{N_{sv}} (\alpha_i^* \alpha_i) K(x_i, x) + b$ is ready and displayed.

As mentioned in Fig. 6 the regression vector is 2d-vector: $x(i) = (y(i-1), y(i-2))$. The computation of this type of vectors is generally difficult in hardware. Consequently, the streaming approach is used for this vector. The two columns of the regression vector are streamed in parallel and simultaneously computed by the kernel processing function and support vectors. The kernel function is polynomial with third order. Then the result passed through the yellow block responsible of the concatenation of Gramian matrix elements.

Finally, the equation $y(x) = \sum_{i=1}^{N_{sv}} (\alpha_i^* \alpha_i) K(x_i, x) + b$ is ready and displayed.

After verifying the of the system functionality on the Simulink environment the generation of hardware components are executed. While building the hardware system, ISE flow generates a bit-stream that will be later used to configure the FPGA. When the compilation is finished, a new one block is created including all the purposes necessary for the executing system on FPGA. The produced hardware capsule the SVM testing phase is allied to a bit-stream file. After downloading this file in the FPGA via the Digilent USB JTAG Cable, then System Generator reads the output back from JTAG and sends it to Simulink. When execution is accomplished, the displayed results are compared to the results expected by the simulation.

The more important criteria in this comparison is the computation time because the software (MATLAB) is incapable to realize matrix multiplication for big dimension (more than 1000). Therefore, the use of FPGA gains a lot of time and big data to be computed in one time. In Table 3, we illustrate the results for computing the testing phase on hardware (FPGA) called SVMsoft and software (MATLAB) called SVMhard for polynomial and sigmoid kernel.

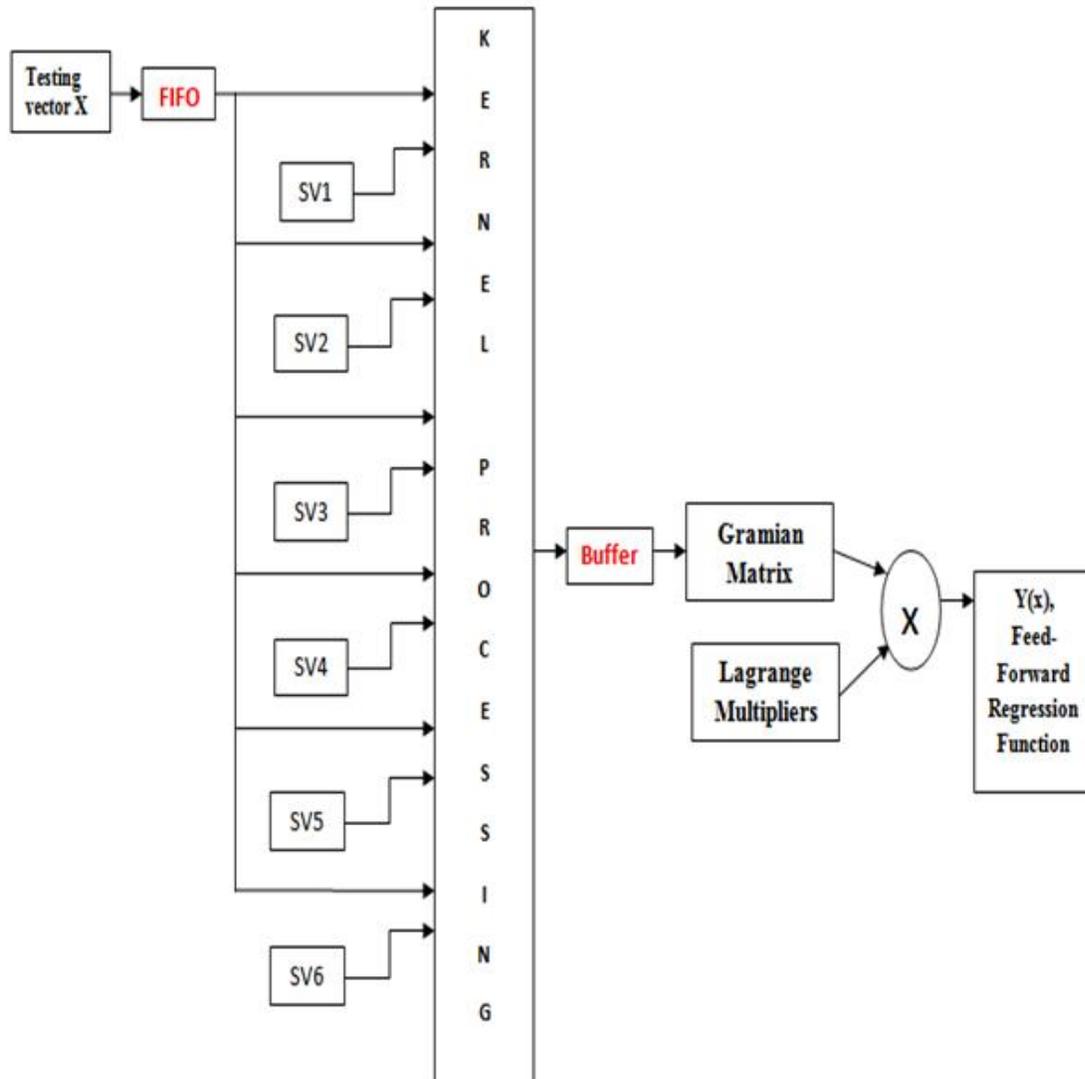


Fig. 6. Architecture for the feed-forward estimation function.

TABLE III. SVMSOFT AND SVMHARD PERFORMANCES

	Kernel type	NMSE Testing	CT(s)
SVMsoft	Polynomial	$5.0766 \cdot 10^{-09}$	740.553
SVMhard		$9.0463 \cdot 10^{-03}$	12.032044
SVMsoft	Sigmoid	$4.8928 \cdot 10^{-08}$	664.931
SVMhard		$2.1974 \cdot 10^{-02}$	10.056215

The exploited FPGA platform was VIRTEX 5 with the clock time period is 10 ns. The difference of computation time

between SVMsoft and SVMhard was very big in order of 60 times. This acceleration was reached thanks to the FPGA computation power. The error rate of SVMhard calculated by the Normalized Mean Squared Error (NMSE) is higher than error rate of SVMsoft. The little difference in accuracy was in order 10^{-5} caused by the fixed point arithmetic used in hardware implementation.

We draw the different plots of SVMsoft, SVMhard and the real output of process in the same Fig. 7. As seen, there is a good concordance between the three plots that demonstrate the efficiency of the adopted approach.

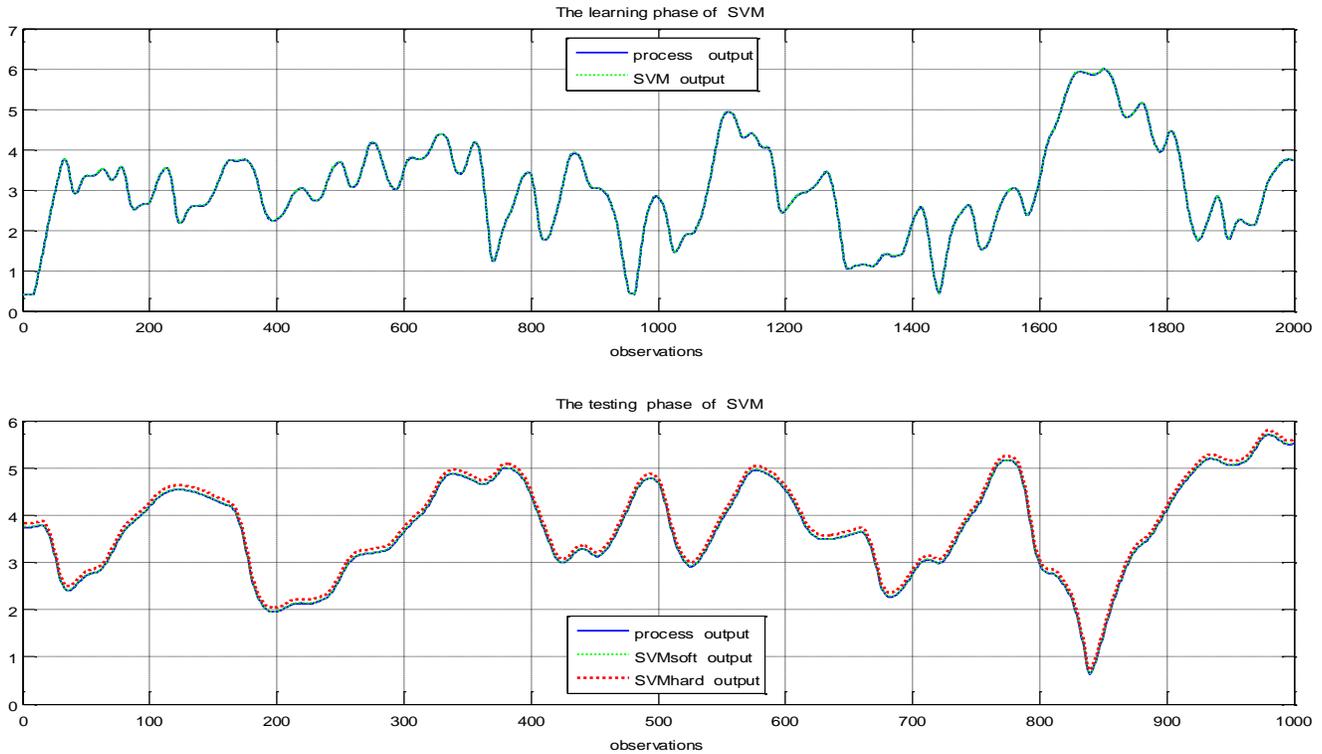


Fig. 7. Learning and testing phase of SVM (hard & soft).

VI. COMPARISON WITH SIMILAR WORK

In this section we propose to compare the performances of our design to the design of Marta Ruiz-Llata [6] that calculates the testing phase of SVM for classification and regression by the same design. As we interested to the regression task, we implemented the same system used in Marta work. It was the sinus cardinal function “sinc” corresponded to this expression:

$$y = \frac{\sin \sqrt{x_1^2 + x_2^2}}{\sqrt{x_1^2 + x_2^2}} \quad (10)$$

To obtain convincing model for testing the prediction function quality the selected datasets were arbitrarily engendered by 400 input (x_1, x_2) values appertaining to $x_i \in (10, 10)$ and conniving its corresponding output value y .

The training phase and parameters selection were also achieved by MATLAB. After finding the optimal values like the regularization parameter C , the insensitive zone ϵ and the parameter of kernel function γ , the efficiency of the model have to be tested with new unknown data. Then the estimated

values were compared with the true values for sinus cardinal function. The testing dataset was composed of 100 arbitrarily values chose in the same way as testing dataset.

The author of this work employed the hardware friendly kernel function described by this expression:

$$2^{-\gamma \|x_i - x_j\|_1}$$

Where $\gamma = 2^{-2}$, $C=1$ and $\epsilon=0.01$. There were 283 support vectors. The middling error between the predicted output and the real one was 0.02. The selected device for implementation was Altera EP2C20 Cyclone II using 8 bits resolution with fixed point arithmetic for representing data. The clock rate of the system was restricted to 30 MHz.

With the same method SVM for regression, the same approach (implementing just the testing phase) and the same platform (FPGA), we exploited our hardware architecture to implement the same sinus cardinal function for the same data sets. The platform exploited by Marta was very old. Therefore, there is no benefit to implement our design on Altera cyclone 2.

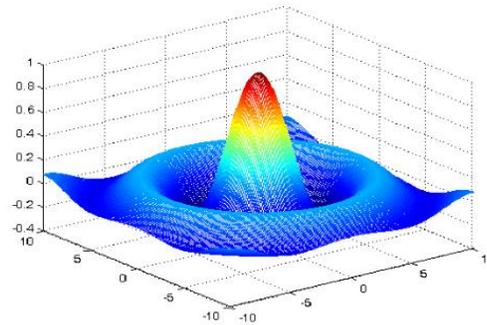


Fig. 8. A three dimensions plot of the sinus cardinal function.

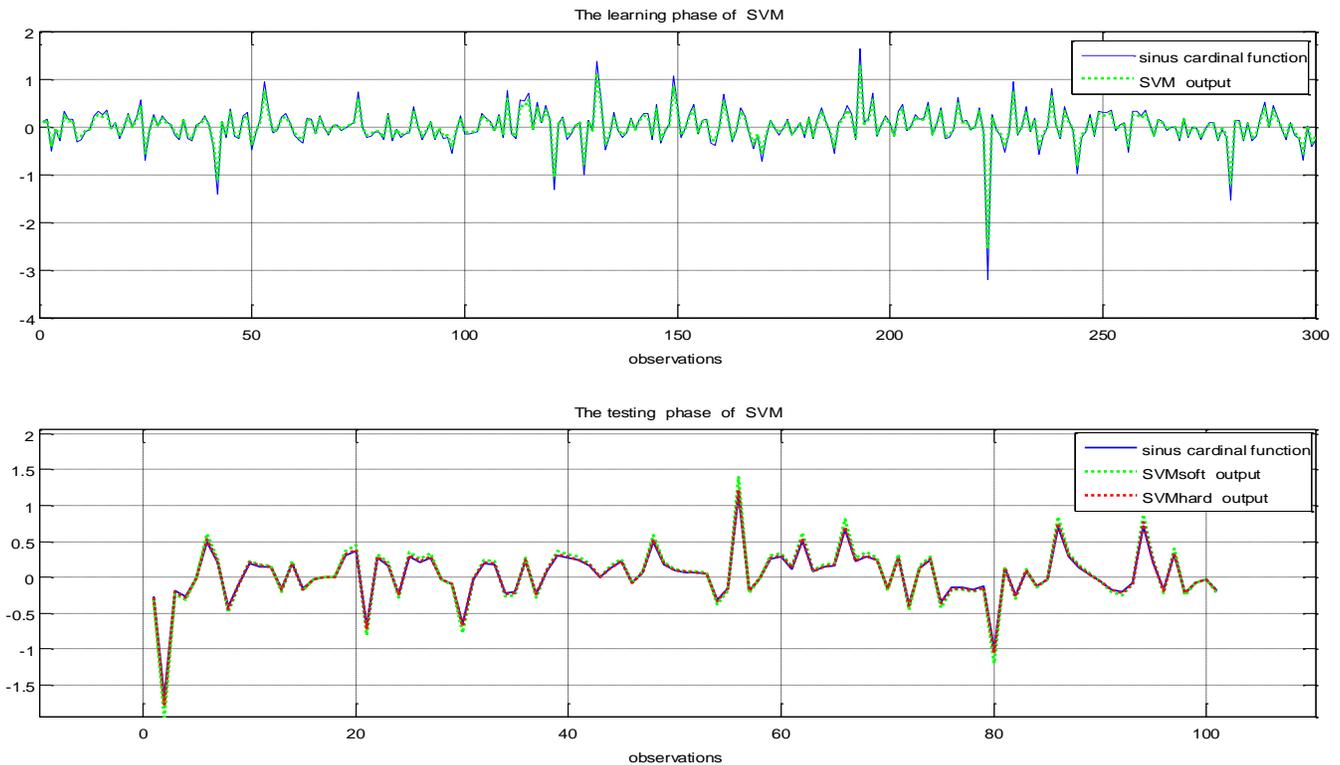


Fig. 9. Learning and testing phase of SVM for sinus cardinal function.

The used kernel function was the sigmoid function with the parameters: $\alpha=1$ and $c=10$. The optimal parameters of SVM were $C=100$ and $\epsilon=0.001$. There were just 10 support vectors. The reached performances were better in term of computation time and error rate. The NMSE was equal to $5.4437 \cdot 10^{-04}$ and the total time was 08.1075 seconds. In Fig. 8, the different outputs are drawn. In the learning phase, the sinus cardinal function and SVM output are drawn. Then in the testing phase, also the sinus cardinal function is plotted with SVM output (SVMsoft) and the SVMhard (the implementation result).

In Fig. 9, it seems clear the strong resemblance between the sinus cardinal function output and the expected output in the training and learning phase. The SVMsoft and SVMhard were approximately confused thanks to the effectiveness of the SVM method.

Therefore, it is easy to predict the output of any process either linear or nonlinear in very short time. The number of support vectors is based on the value of the margin ϵ .

VII. CONCLUSION

In this paper, an efficient method of implementing the testing phase of SVM method was advised. The basic contribution of this approach is to accelerate the computation of the RKHS model by use of powerful FPGA. The experiments prove the excellent speedup attained that is more than 60 times compared with the software computation time.

The advantage of the designing tool Xilinx System Generator is the possibility to implement the software and the hardware in the same environment. Furthermore, Simulink offers a pleasant graphics interface for supple modellization

and simulation. The design was well organized into streaming approach along the testing phase.

This practice permitted to construct a robust model for nonlinear system using novel data in the testing phase.

Future works will incorporate the use of the Xilinx System Generator for the development of other kernel functions to increase the simulation precision. Also, better performances can be reached with newer and more powerful FPGA type.

REFERENCES

- [1] Lennart Ljung. Some aspects on nonlinear system identification. In Proc 14th IFAC Symposium on System Identification, Newcastle, Australia, March 2006.
- [2] Lennart Ljung. Perspectives on system identification. Annual Reviews in Control, 34(1), March 2010.
- [3] Alaa F. Sheta, "A Comparison between Regression, Artificial Neural Networks and Support Vector Machines for Predicting Stock Market Index", communication on (IJARAI) International Journal of Advanced Research in Artificial Intelligence, Vol. 4, No.7, 2015
- [4] Bernhard Scholkopf and Alexander J. Smola, "Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond", MIT Press Cambridge, MA, USA 2001
- [5] Cristian Preda, "Regression models for functional data by reproducing kernel Hilbert spaces methods", Journal of Statistical Planning and Inference Volume 137, Issue 3, 1 March 2007, Pages 829–840.
- [6] M.Ruiz-Llata, G. Guarnizo, and M. Yébenes-Calvino, "FPGA Implementation of a Support Vector Machine for Classification and Regression," in The 2010 International Joint Conference on Neural Networks (IJCNN), 2010, pp. 1-5.
- [7] D. Anguita, S. Pischiutta, S. Ridella, and D. Sterpi, "Feed-Forward Support Vector Machine without Multipliers," IEEE Transactions on Neural Networks, vol. 17, pp. 1328-1331, 2006.
- [8] X. Pan, H. Yang, L. Li, Z. Liu, and L. Hou, "FPGA Implementation of SVM Decision Function Based on Hardware-friendly Kernel," in International Conference on Computational and Information Sciences, ICCIS 2013 Proceedings, 2013, pp. 133-136.
- [9] R. Patil, G. Gupta, V. Sahula, and A. Mandal, "Power Aware Hardware Prototyping of Multiclass SVM Classifier Through Reconfiguration," in 2012 25th International Conference on VLSI Design (VLSID), 2012, pp. 62-67
- [10] C. Kyrkou and T. Theocharides, "A Parallel Hardware Architecture for Real-Time Object Detection with Support Vector Machines," IEEE Transactions on Computers, vol. 61, pp. 831-842, 2012.
- [11] C.-W. Hsu and C.-J. Lin, "A Comparison of Methods for Multiclass Support Vector Machines," IEEE Transactions on Neural Networks, vol. 13, pp. 415-425, 2002
- [12] B. Mandal, M. P. Sarma, and K. K. Sarma, "Implementation of Systolic Array Based SVM Classifier Using Multiplierless Kernel," in International Conference on Signal Processing and Integrated Networks (SPIN), 2014, pp. 35-39.
- [13] H. Hussin, K. Benkrid, and H. Seker, "Reconfiguration-Based Implementation of SVM Classifier on FPGA for Classifying Microarray Data," in 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2013, pp. 3058-306
- [14] M. Wielgosz, E. Jamro, D. Zurek, and K. Wiatr, "FPGA Implementation The Selected Parts of Fast Image Segmentation," in Studies in Computational Intelligence vol. 390, ed, 2012, pp. 203-21
- [15] M. Pietron, M. Wielgosz, D. Zurek, E. Jamro, and K. Wiatr, "Comparison of GPU And FPGA Implementation of SVM Algorithm for Fast Image Segmentation," in Architecture of Computing Systems-ARCS 2013, ed: Springer, 2013, pp. 292-302.
- [16] Y. Ago, K. Nakano, and Y. Ito, "A Classification Processor for Support Vector Machine with Embedded DSP Slice and Block RAM in the FPGA," in IEEE 7th International Symposium on Embedded Multicore Socs (MCSoc), 2013, pp. 91-96
- [17] S. Kim, S. Lee, and K. Cho, "Design of High-Performance Unified Circuit for Linear and Non-Linear SVM Classifications," Journal of Semiconductor Technology and Science, vol. 12, pp. 162-167, 2012.
- [18] J. Nayak, B. Naik, and H. Behera, "A Comprehensive Survey on Support Vector Machine in Data Mining Task: Applications & Challenges," International Journal of Database Theory and Application, vol. 8, pp. 169-186, 2015.
- [19] J. Gimeno Sarciada, H. Lamel Rivera, and M. Jiménez, "CORDIC Algorithms for SVM FPGA Implementation," in Proceedings of SPIE - The International Society for Optical Engineering, 2010.
- [20] H. Lamela, J. Gimeno, M. Jimenez, and M. Rruiz, "Performance Evaluation of FPGA Implementation of Digital Rotation Support Vector Machine," in SPIE Defense and Security Symposium, 2008, pp. 697908-697908-8
- [21] T. Bollaert, "Catapult Synthesis: Practical Introduction to Interactive C Synthesis," SpringerLink Book Chapter, Pages 29-52, ISBN 978-1-4020-8587-1.
- [22] Y. L.C.N.W. Wong. Generating hardware from OpenMP programs. Proc.IEEE Int.Conf. on Field Programmable Technology, pages73ñ80,Dec.2006.
- [23] J. Gummaraju and M. Rosenblum. Stream Programming on General-Purpose Processors. In Proc.38th Int. Symp.on Micro architecture, pages343ñ354, Washington,DC,2005.

A Synthesis on SWOT Analysis of Public Sector Healthcare Knowledge Management Information Systems in Pakistan

Arfan Arshad

Department of Information System
KICT, International Islamic University
Malaysia

Mohamad Fauzan Noordin

Department of Information System
KICT, International Islamic University
Malaysia

Roslina Bint Othman

Department of Library & Information
System, KICT, International Islamic
University, Malaysia

Abstract—Healthcare is a community service sector and has been delivering its services for the betterment of civic health since its establishment at communal level. For working efficiently and effectively, this sector profoundly relies on correct and complete health information of people and a proficient integrated healthcare knowledge management information system (HKMIS) to manage this information. The performance of Healthcare organizations has significantly augmented by inception of Information and Communications Technology (ICT) in HKMIS in developed countries, but is yet to exhibit its full potential in developing countries specifically those with huge populations like Pakistan. An exploratory qualitative research methodology was adopted to conduct this study. The purpose and objective of this study was to determine and investigate the internal and external factors that influence the performance of HKMIS by performing SWOT analysis on two of the largest public-sector healthcare organizations of Pakistan. The findings of this study will certainly help authorities to devise methods of improvement in Pakistani HKMIS eventually paving ways towards a better and improved healthcare in the future.

Keywords—Healthcare; knowledge management; healthcare knowledge management information system; information and communications technology; SWOT analysis; internal and external factors; healthcare organizations

I. INTRODUCTION: AN OVERVIEW

Developing countries have always been under extreme pressure and facing challenges in providing community services to the public sector as healthcare being one of them to be mentioned. Specifically, nations like Pakistan having huge population (where more than 60% population is colonized in rural areas) face a great deal and variety of challenges related to healthcare systems and their effective implementation mainly due to limited resources and working capability [1].

Healthcare organizations now-a-days rely on healthy healthcare knowledge management information systems (HKMIS) to provide reliable data that report on health system performance [1]. Hence availability of relevant, timely and accurate information related to healthcare system performance is the key to successful strengthening of the healthcare systems [2]. Further, Information and Communications Technology (ICT) has been an important tool being used to expand the healthcare systems working capabilities beyond

physical limits. Some of the examples in this respect are Tele-medicine, E-Health, and Tele-pharmacy, etc. [3].

Now-a-days, healthcare management information systems are used to support the overall management of healthcare organization related to data and information about patient care, diseases, up-to performing resource management and knowledge management [6]. An advanced innovated and integrated form of these systems is known as Knowledge Management System (KMS). HKMIS are affected by a great variety of internal and external influential factors [26]. How well an HKMIS performs, depends on these influential factors and issues related to them [7]. A typical HKMIS architecture for clinic to large healthcare organization access application anytime from anywhere is shown in Fig. 1.

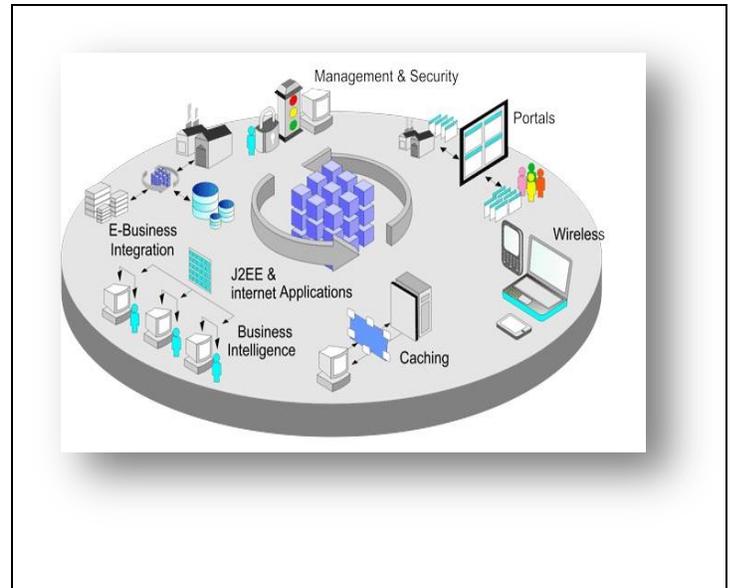


Fig. 1. A Typical HKMIS Framework Architecture Adapted from <http://www.hospycare.com/technology.html>.

A perfect HKMIS is basically triangulation of three professional disciplines named as Medicine, ICT and Management [4]. At times, the HKMIS have been evolved and taken different names as Hospital Information Systems, Clinical information systems, Hospital information

management systems and Healthcare information systems etc, all exhibiting and serving the same purpose and objectives [5].

This study used a professional and scientific approach named as SWOT analysis to identify and determine these internal and external influential factors for evaluating the existing condition of HKMIS in public sector healthcare organizations of Pakistan. The research questions comprised of:

- 1) What is the condition of existing HKMIS implemented in public sector organizations of Pakistan?
- 2) What are the Strengths of the existing HKMIS under study?
- 3) What are the weaknesses and their possible influential factors that impact existing HKMIS?
- 4) What are the opportunities to improve the working capability and performance of public sector HKMIS?
- 5) What are the possible threats that may lead to failure of HKMIS?

II. BACKGROUND AND LITERATURE REVIEW

A. History of Healthcare Computing

Use of computers in healthcare is not a contemporary phenomenon. It's been more than six decades, since some of the major healthcare organizations started using computers for the processing of batch files of health related records [8]. During early 1950s, most of the major hospitals of G7 countries started using mainframe computers for routine data processing related to healthcare. Although it was very slow as compared to latest computing systems [9], [10].

The era starting from early 1960s through the mid 1970s experienced new emerging essence in healthcare computing. The performance of automated computing medical systems also improved considerably and it paved the path towards increased use and integration of computing systems into healthcare systems [9], [10]. The era of early 1980s was dedicated to computer miniaturization and drastic increase in processing power along with storage capacity. This helped many healthcare organizations round the globe to implement new healthcare management information systems (HMIS). This helped a lot to process massive and huge health data processing being fast and accurate [9], [10].

Later on, Innovation in networking and telecommunication systems during 1990s took healthcare computing systems i.e. HMIS further step ahead, enhancing them into integrated and distributed HMIS. This helped to reduce distance by introducing the innovative concepts of E-healthcare such as Tele-medicine, Tele-pharmacy, etc. The E-healthcare was further advanced to E-nursing systems, E-homecare systems, E-Clinical Decision Support Systems, Knowledge Based Systems and expert systems [9], [10].

B. Healthcare Computing / HKMIS in Pakistan

Due to its massive data creation and handling sensitive confidential information, healthcare is considered as an important industry now-a-days [29]. So, keeping in view the importance of healthcare data and its effective utilization, the Government of Pakistan considered revolutionizing the

healthcare industry and changing from manual data processing to electronic data processing methods in early 1990s. In response to this need the Ministry of Health, Government of Pakistan, in collaboration with the provincial healthcare departments and international agencies developed a National Health Management Information Systems (HMIS) during 1990-93 [17]-[19].

An effort was instigated by the Basic Health Services Cell (now National HMIS Cell of Ministry of Health, Government of Pakistan) to establish a countrywide HMIS facility in 1990s keeping in view the importance of management information systems as an essential tool required for improved performance and quality of working of healthcare organizations [30]. This effort was supported by significant Provincial Health Departments also. The initiative was financially and technically supported by international agencies like USAID, UNICEF and WHO also [31]. The objective was to provide support to tactical and operative managers in decision making process [17]-[19].

Further, the HMIS was institutionalized in all the provincial healthcare headquarters of the country by the help of Family Health Projects initiated by World Bank support in mid 1990s [32]. As per reports produced by Ministry of Healthcare, Government of Pakistan, more than 90% of the primary healthcare facilities of public sector account under this system which was implemented in a phased manner [17]-[19].

Subsequently, in 1994, the Government of Pakistan also developed a parallel community based information system (CBIS) under the National Program for Family Planning and Primary Healthcare (NPF & PHC). Additionally, some other several types of information systems related to Malaria, AIDS, and TB programs etc. are also running at the district levels but not fully integrated with National HMIS [17]-[19].

Despite all these efforts and initiatives, still a lot more is to be done in public sector HMIS to improve the performance and provision of quality healthcare facilities. This includes processing of collective information and data at centralized data processing facility, improvement in data quality, data collection methods, knowledge management, knowledge innovation and sharing, availability of summarized and scrutinized information to establish effective plans and assessment of healthcare services etc. Currently, reports generated by the facility based HMIS receives low priority, monitoring is poor and facility staff looks upon HMIS as an additional workload [20], [21].

The scope of the current HMIS is however, limited to the Primary Level healthcare facilities only and no data from inpatient/hospital, private care facilities or from the health facilities other than Provincial Health Departments are captured [22], [25].

C. An Introduction to SWOT Analysis

Every organization requires continuous improvements in its day to day processes for better performance and excellent quality of working [27]. This can be done if all the factors that influence the working of the organization are properly identified and defined [11]. Such factors are divided into

internal factors (Strengths & Weaknesses) and external factors (Opportunities & Threats) [28]. The assessment and evaluation of these factors is done by using an exploratory scientific technique named as SWOT analysis. So, SWOT analysis can be defined as:

“**SWOT Analysis** is an examination and evaluation of an organization’s internal strengths and weaknesses, its opportunities for growth and improvements and the threats the external environment poses to its survival and working [11].”

The definition can be summarized and exhibited in a matrix as shown in Fig. 2.

The internal factors are considered relatively controllable and can be manipulated by organization itself. On the contrary, the external factors are somewhat out of the control of the organization (may be controllable up-to some extent) and imposed by the environment in which the organization operates [12]. The SWOT analysis matrix illustrated in Fig. 2 clearly explains that (if it is performed realistically and unbiased) an organization must strive hard to improve the strengths and opportunities that are helpful towards attaining its objectives and goals. While on the other side, an organization must strive hard to eradicate its weaknesses and threats that hinder its performance and quality of working. The key to success while performing SWOT analysis is to be honest, neutral, impartial and realistic [11].

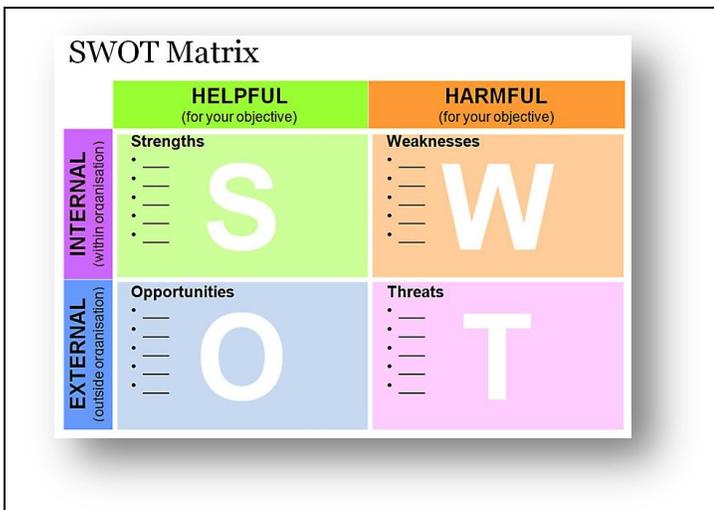


Fig. 2. SWOT Analysis Matrix Diagram – Adapted from “www.Business-Docs.co.uk”.

D. SWOT Analysis: As a Process

SWOT analysis is a preliminary exploratory tool that helps the organizations to investigate and evaluate their internal and external environment and the related influential factors. This tool helps the organization in their decision making process for taking steps that improve and increase the organizational performance and capability of quality working [13].

SWOT analysis can be divided into following four steps:

Step 1: The first step is used to identify, collect and analyze key data about the organization that may be related to working environment of the organization, community health

status, funding resources, existing medical and computer technology along with demographic information of the population [13], [14].

Step 2: In second step the collected data is sorted and categorized according to strengths, weaknesses, opportunities and threats that influence the organizational working capability [13], [14].

Step 3: The third step is used to develop a SWOT matrix for each of the business alternative that is being considered to be likely a potential substitute or a better option [13], [14].

Step 4: The fourth step hence assists in decision making process. It helps the organization to integrate the findings and results of SWOT analysis with the decision making process for establishing and determining the best business alternative that best fits for the organizational strategic plans [13], [14].

III. SWOT ANALYSIS IN HEALTHCARE

SWOT analysis and evaluation technique was initially intended to present an overall systematic analysis of businesses that were related to organizations other than healthcare, but it has proven its usefulness and advantages in healthcare organizations as well. Recent years have experienced its increased use and great impact in healthcare industry [15]. It is pertinent that healthcare organizations also require improvement and innovation for increased performance, quality and optimum functionality. For this purpose, progressive adjustments and up-gradations are a necessity to acquire. To identify areas for improvements, SWOT analysis has been a helpful and successful tool that was ignored by healthcare organizations for many years [16].

The sorting and categorization of organizational data in Healthcare systems can be illustrated as under:

A. Strengths

Strengths are those internal factors that support and illustrate extraordinary performance of a healthcare organization e.g. extra ordinary IT infrastructure, highly qualified and experienced healthcare professionals, state of the art equipments, excellent services, etc. [13], [14].

B. Weaknesses

On the contrary, weaknesses are those internal factors that hinder the working capability and negatively affect the performance of a healthcare organization. They can be mismanagement of resources, lack of financial resources, incompetent healthcare professionals, outdated equipments, etc. [13], [14].

C. Opportunities

Opportunities are those factors that are external to healthcare organizations. They provide initiatives for improvements. Examples include collaborations with other organizations for better services, plans for better organization and management, new funding programs for better IT infrastructure, effective training and informative programs for community development, etc. [13], [14].

D. Threats

Threats are those external factors which are considered to be potential risks or dangers that could cause harm to the quality of working and performance of healthcare organizations. Economic instability, rapidly changing technology, budget deficits, un-necessary political intervention, and political insecurity are some of the related examples [13], [14].

IV. OBJECTIVES OF THE STUDY

1) To perform a comprehensive and in-depth overall analysis of existing healthcare management information systems implemented in public sector organizations of Pakistan.

2) To identify and define Strengths of the existing HKMIS under study.

3) To evaluate weaknesses and their possible influential factors.

4) To discover better opportunities to improve the working capability and performance of public sector HKMIS.

5) To identify and evaluate the possible threats that may lead to failure of HKMIS.

V. METHODOLOGY AND APPROACH ADOPTED

This cross sectional study was conducted in two famous and most visited healthcare public sector organizations present in Pakistan. An exploratory qualitative research methodology was adopted to conduct this study and a detailed analysis was performed on the data collected through detailed interviews, structured walkthroughs, observation, available forms and individual surveys at the end user level.

The study was also complemented and supported by a comprehensive methodological review of available literature and related data. Published and unpublished documents including government reports, peer review journals and other literature such as local journals were also a source of information for this study. The analysis was helped by discussion during interviews with the experts in the relevant field including government, stakeholder agencies and public sector healthcare specialists.

VI. THE CASE STUDY: PUBLIC SECTOR HEALTHCARE ORGANIZATIONS

A. Introduction

This section provides an excerpt of information from a case study conducted on two government hospitals (A & B—pseudonym used to ensure confidentiality) in one of the major provincial capital city in Pakistan. These were one of the largest and oldest public healthcare organizations governed and controlled by the local government and comprising of more than 2000 beds collectively. Hospital “A” employs 1025 healthcare professionals with a total count of 1800 employees and staff in its 36 departments and is one of the oldest public sector healthcare organizations in Southeast Asia. Hospital B employs 1225 healthcare professionals with a total count of 2100 employees and staff in its 32 departments with its roots being older than 70 years. The hospitals seem to be

overcrowded with a significant number of in-patients (A>30000, B>25000 per year) and out-patients (A>1000000, B>900000 per year).

B. Conduction of Case Study

The case study was conducted to obtain an account of information on practical grounds related to existence and implementation of HKMIS in the top and famous public sector Pakistani healthcare organizations. The researchers collected data and information for their study by help of generalized interviews, structured walkthroughs, existing document reviews and self observations. A substantial number of people including healthcare professionals, IT staff, support staff and some educated patients were met to dig out the information. The interviews and data collection objectives of the authors were generally based upon the following major domains and points:

- Know how about the terminologies “Knowledge” & “Knowledge Management”.
- Know how about the terminology “HKMIS” (Healthcare Knowledge Management Information System / Healthcare Management Information System)
- Know how about the importance of ICT in healthcare.
- Role of ICT in healthcare.
- Knowledge about working of HMIS / HKMIS.
- Benefits of a healthy HKMIS in healthcare.
- Financial and top management support for successful HKMIS implementation.
- Implementation of HKMIS processes in the organization.
- Role of HKMIS in strengthening the healthcare organization.
- Weaknesses of existing HKMIS.
- Opportunities and strategies to improve existing HKMIS
- Salient threats faced by HKMIS.
- Role of HKMIS in decision making.
- Best practices about HKMIS for patient care.
- Organizational and Technology infrastructure, etc.

C. Findings and Results

Compiling the findings and results was a hectic and tedious task. The data collected by the help of detailed interviews, discussions and individual surveys ended up with a great deal of variation and gaps among the answers produced by the tactical management, operational management, healthcare professionals and other individuals of the organizations. Nonetheless, the results of the discussions and observations revealed that most of the healthcare professionals, management personals, and other stakeholders involved in the research population from both the healthcare organizations do have some sort of understanding of the term

HKMIS, its benefits and its role in improving the healthcare services.

Most of the time, the findings and results were supported by the help of data collected through self observations, structured walkthroughs and study of available forms and reports. Anyway, the available data was carefully compiled and was distributed and sorted in the four major categories according to the SWOT analysis tool.

Furthermore, it was established that both the healthcare organizations does not have a competent HKMIS implementation to support daily based healthcare activities with significant weaknesses, but it was also recognized that there are certain specific areas of opportunities available for improvements of-course keeping in mind the posed threats.

Tables 1 and 2 enlist the internal factors identified while conducting SWOT analysis on public sector healthcare organization's HKMIS.

TABLE I. SWOT ANALYSIS OF PUBLIC SECTOR HEALTHCARE – INTERNAL FACTORS (STRENGTHS)

Internal Factors – Strengths identified:	
Strengths	1) Provision of improved healthcare services
	2) Decrease in data and medical errors
	3) Better and improved data storage
	4) Better safeguarding and improved confidentiality of sensitive information
	5) Better and fast communication between stakeholders
	6) Increased and better cost savings
	7) Improved access to accurate and relative information
	8) Increased productivity of end users
	9) Availability of timely data
	10) Better and improved reporting
	11) Reduced possibility of data loss
	12) Improved quality and originality of documentation
	13) Enhanced paperless environment
	14) Solid and sound IT infrastructure

TABLE II. SWOT ANALYSIS OF PUBLIC SECTOR HEALTHCARE – INTERNAL FACTORS (WEAKNESSES)

Internal Factors – Weaknesses identified:	
Weaknesses	1) Shortage of competent staff
	2) Lack of specific and professional training programs
	3) Lack of top management commitment and seriousness

Weaknesses	4) Lack of effective system integration
	5) Unprofessional reporting structure
	6) Fragmentation and disintegration of health information
	7) Lack of E-health services
	8) Limited access to internet and collaborative tools / applications
	9) Lack of interest and professional ethics in learning new systems
	10) Significant errors in recording healthcare data
	11) Lack of accountability and transparency
	12) Lack of sufficient hardware and software maintenance staff
	13) Lack of motivation or reward criteria
	14) Increased and high costs of IT adoption
	15) Limited or no interoperability between service providers

Tables 3 and 4 enlist the external factors identified while conducting SWOT analysis on public sector healthcare organization's HKMIS.

TABLE III. SWOT ANALYSIS OF PUBLIC SECTOR HEALTHCARE – EXTERNAL FACTORS (OPPORTUNITIES)

External Factors – Opportunities Identified:	
Opportunities	1) Improvement in reporting and data presentation capabilities
	2) Improvement in quality of healthcare services
	3) Effective and efficient resources utilization procedures
	4) Improvement in patients trust and satisfaction
	5) Encouragement in proactive healthcare practices
	6) Public awareness and community support programs
	7) Training programs and facilities
	8) Unification and integration of Public and Private sector health records
	9) Improved support for knowledge management and decision making
	10) Productive, efficient and effective healthcare management
	11) Better human resource management
	12) Costing and budget analysis for enhanced funds utilization
	13) Sufficient allocation of resources for supporting IT infrastructure
	14) Internet availability and enhanced bandwidth

TABLE IV. SWOT ANALYSIS OF PUBLIC SECTOR HEALTHCARE –
EXTERNAL FACTORS (THREATS)

External Factors – Threats Identified:	
Threats	1) Lack of top management commitment and seriousness
	2) Ineffective and inefficient governance
	3) High staff turnover rate due to political interventions
	4) Shortage of human resources and untrained staff
	5) Undue transfers and postings of professional staff
	6) Patients perceptions on privacy and confidentiality of health data
	7) Load shedding and electrical surges
	8) Rapid changes in technology and IT systems
	9) Unreliable and unrealistic system and reporting requirements
	10) Data under security and hacking threats
	11) End users resistance to systems change and implementation

VII. DISCUSSION AND CONCLUSION

There are no two opinions on the fact that ICT and healthy HKMIS have been a key factor for providing timely, accurate and reliable information for supporting decision making process in improving healthcare services in public as well as private sector healthcare organizations [23]. The working capability and performance of such systems can improve by dealing professionally with internal and external factors that influence them. This has a positive impact on quality of service, growth and sustainability of the healthcare programs and strategic plans induced by healthcare organizations [24].

On the other side, it is also important that HKMIS must be easy to use, flexible and easy to understand as well as must incorporate with user requirements. Public sector healthcare organizations face a drastic increase in patient visits every year due to constant increase in population of developing countries and deteriorating conditions of healthcare and non availability of healthy environment. So, HKMIS at public sector healthcare organization is under extreme constant pressure to deal with loads of healthcare data and information along with its processing and provision for decision making. These objectives can be attained in a better way by taking care of weaknesses and threats posed to the healthcare organizations. Further, taking benefit of prevailing opportunities also helps improvise the working capability and performance of the HKMIS.

This study performed a detailed SWOT analysis on public sector healthcare organizations of Pakistan and presented the facts and findings in detailed as well as summarized manner. An influential and healthy HKMIS is considered to be useful when it provides information that helps decision making effectively and efficiently.

For further research, each and every point identified and mentioned as strengths, weaknesses, opportunities and threats in public sector healthcare organizations of Pakistan, could be

studied in depth and alternate solutions could be identified and implemented to cater them.

REFERENCES

- [1] Hameed, Shafqat, Jawad Karamat, and Kashif Mehmood, "Effectual dynamics and prolific usage of knowledge management & engineering in health care industry," *Life Science Journal*, vol. 9, issue 2, 2012.
- [2] C. AbouZahr, and T. Boerma, "Health information systems: The foundations of public health," *Bulletin of World Health Organization*, vol. 83/8, pp. 578-583, August 2005.
- [3] M. C. Azubuike, J. E. Ehiri, "Health information systems in developing countries: Benefits, problems, and prospects," *Journal of the Royal Society for Promotion of Health*, vol. 119, pp. 180-184, 1999.
- [4] Hussain, Fehmida, and S. Ali Raza, "A knowledge management framework to operationalize experiential knowledge: mapping tacit medical knowledge with explicit practice guidelines," *National Conference on Emerging Technologies*, 2004.
- [5] N. Mahmood, A. Burney, Z. Abbas, and K. Rizwan, "Data and knowledge management in designing healthcare information systems," *International Journal of Computer Applications*, vol. 50, no. 2, July 2012.
- [6] A. Razzaque, and A. Jalal-Karim, "Conceptual healthcare knowledge management model for adaptability and interoperability of EHR," *European, Mediterranean & Middle Eastern Conference on Information Systems*, 2010 (EMCIS2010), April 12-13 2010, Abu Dhabi, UAE
- [7] S.C. Chong, and Y.S. Choi, "Critical Factors in the Successful Implementation of Knowledge management," *Journal of Knowledge Management Practice*, June 2005, [Electronic], available: <http://www.tlinc.com/articl90.htm>
- [8] T. Lippeveld, R. Sauerborn, and C. Bodart, "Design and implementation of health management information systems," Geneva: World Health Organization, 2000.
- [9] World Health Organization. 2008. "Health information systems, toolkit on monitoring health system strengthening," Geneva: WHO; 2008.
- [10] World Health Organization. World Health Report. 2003. "Shaping the Future," Geneva: WHO; 2003.
- [11] Z. Terzic, Z. Vukasinovic, V. Bjegovic-Mikanovi, V. Jovanovic, and R. Janicic, "SWOT analysis: The analytical method in the process of planning and its application in the development of orthopedic hospital department," *Srpski Arhiv Za Celokupno Lekarstvo – Journals*, vol. 138, pp. 473-479, 2010.
- [12] M. Smith, S. Madon, A. Anifalaje, M. Lazarro-Malecela, and E. Michael, "Integrated Health Information Systems in Tanzania: experience and challenges," *The Electronic Journal of Information Systems in Developing Countries*, vol. 33, pp. 1-21, 2008.
- [13] M.H. Marilyn, J. Nixon, "Exploring SWOT analysis – where are we now?: A review of academic research from the last decade", *Journal of Strategy and Management*, Vol. 3 Issue: 3, pp.215-251, doi: 10.1108/17554251011064837, 2010.
- [14] J.D.H Wijngaarden, G.R.M.Scholten, and K.P.V Wijk, "Strategic analysis for health care organizations: the suitability of the SWOT-analysis," *International Journal of Health Planning and Management*, vol. 27, pp. 34-49, doi:10.1002/hpm.1032, 2012.
- [15] M. Ali, and Y. Horikoshi, "Situation analysis of health management information system in Pakistan," *Pakistan Journal of Medical Research*, vol. 41, pp. 64-69, 2002.
- [16] S. Nishtar, "The Gateway Paper: Health System in Pakistan – A way forward," Islamabad, Pakistan: Pakistan's Health Policy Forum and Heart-file, 2006, Available from URL: <http://www.heartfile.org/pdf/phpfGWP.Pdf>
- [17] Ministry of Health, Government of Pakistan. 1995. "Health Management Information System: National Feed Back Report September 1995," Primary Health Care Cell. Islamabad, 1995.
- [18] Ministry of Health, Government of Pakistan, "Health Management Information System: National Feed Back Report 1996," Primary Health Care Cell. Islamabad, 1996.

- [19] Ministry of Health, Government of Pakistan, "Health management information system: National feedback report," September 1997-98: National MIS Cell, 1999.
- [20] M.Q. Suleman, and M. Ali, "Pakistan's Health management information system: Health manager's perspectives," Journal of Pakistan Medical Association, vol. 59, pp.10-14, 2009.
- [21] Hafeez, Z. Khan, K.M. Bile, R. Jooma, and M. Sheikh, "Pakistan human resources for health assessment," Eastern Mediterranean Health Journal, (16-Suppl), pp. 145-51, 2010.
- [22] M.S. Qazi, and M. Ali, "Health Management Information System utilization in Pakistan: challenges, pitfalls and the way forward," Bioscience Trends, vol. 5, pp. 245-254, 2011.
- [23] A. Rizwan, Bhutto and A. Maqsood, "Evaluating knowledge & information-based healthcare system: A case study of DOW university of health sciences & civil hospital Karachi," MS/MS152006ISR, 53 pages, January 2007.
- [24] H. Lee, and B. Choi, "Knowledge management enablers, processes, and organizational performance: An integration and empirical examination," Journal of Management Information Systems, vol. 20(1), pp. 179-228, 2000.
- [25] M. Akram and F.J. Khan, "Health care services and government spending in Pakistan," PIDE Working Papers, 2007.
- [26] V.S. Anantamula, and S.Kanungo, "Modeling enablers for successful KM implementation," Proceedings of the 40th Hawaii International Conference on System Sciences, 2007.
- [27] M.F. Noordin, R. Othman, and N.A. Zakaria, "Peopleware and heartware – The philosophy of knowledge management, research and innovation in information systems (ICRIIS)," International Conference on 23-24 Nov. 2011.
- [28] S.Y.S. Danesh, N.S. Rad, S.N. Mobasher, and M.M.S. Danesh, "The Investigation of Mutual Relations of Success Factors of Knowledge Management in Project-Centered Organizations," Journal of Basic and Applied Scientific Research, vol. 2, no. 4, pp. 3888-3896, 2012.
- [29] WHO Report of the Health System Review Mission – Pakistan, Available from URL:<http://gis.emro.who.int/HealthSystemObservatory/PDF/HealthSystemReviewMission>
- [30] Ministry of Health, Government of Pakistan, "National Health Management Information System (HMIS): An Overview," http://www.pakistan.gov.pk/divisions/ContentInfo.jsp?DivID=25&cPath=254_260&ContentID=1635
- [31] Government of Pakistan, "Pakistan Bureau of Statistics. Statistics Division," National Health Accounts Pakistan, 2009-10. Available from URL: www.pbs.gov.pk/content/nationalhealthaccountspakistan200910.
- [32] Government of Pakistan, "Survey of private medical, dental and other health services. Islamabad, Pakistan," Federal Bureau of Statistics, 2001.

Multi-Agent based Functional Testing in the Distributed Environment

Muhammad Fraz Malik

Shaheed Zulfikar Ali Bhutto Institute of Science and
Technology, Islamabad, Pakistan

M. N. A. Khan

Shaheed Zulfikar Ali Bhutto Institute of Science and
Technology, Islamabad, Pakistan

Uzma Bibi

Shaheed Zulfikar Ali Bhutto Institute of Science and
Technology, Islamabad, Pakistan

Muhammad Ayaz Malik

Chalmers University of Technology,
Gothenburg, Sweden

Abstract—Verification and testing are two formal techniques of defect reduction applied on designing and development phases of SDLC to rationalize quality assurance activities. The process of testing applications in the distributed environment becomes too complex. This study discusses a distributed testing framework that consists of many parallel tester components. The idea is based on utilizing client server environment to conduct software testing efficiently and in a short span of time. It is pertinent to mention that this study is restricted to testing of functional aspects of the software while testing of performance and other quality-of-service aspects are outside the scope of the study. An important factor influencing the use of agent technology in software testing is the dynamic nature of events. Since agents are characterized by intelligence and autonomy, their ability to interact with the environment offers added functionality to make decisions based on the needs of the scenarios that are dynamic in nature. This study shows that the use of agents to build a dynamic model for software testing in the distributed environment results in a more robust and efficient design. The proposed framework is based on distribution of test cases among multiple agents deployed across a distributed system which collaborate with each other to perform testing in an efficient manner. The proposed framework also provides an in-depth visibility into the software quality by providing the defect statistics on-the-fly. The experiments have been conducted using Selenium test automation tool. The test cases along with their test scripts and the test run results are described herein.

Keywords—Software quality assurance; software testing; distributed environment; input variation testing; test vectors; multi-agents

I. INTRODUCTION

Software testing is a process which is used to ensure quality of the product by assessing software behavior according to the specifications. The abnormal software behavior is generally termed as a bug or defect which could be a fault, error or failure of software that causes it to produce unexpected results or exhibit unwanted behavior. It is important to mention that faults lead to errors and errors lead to failures. The term failure is generally used from user's perspective which means that certain functionality is either missing or is not producing the desired results. The software quality is mostly viewed from customers' perspective mainly the customer satisfaction and is

generally termed as fitness for purpose. IEEE and ISO define software quality as meeting the "user needs or expectations" and ability to "satisfy specified or implied needs" respectively.

Testing is a critical phase in designing and development of software and computer systems. In case of distributed software applications, the testing process particularly becomes too complex as the distributed applications inherently are multifaceted and more convoluted than the applications developed in collocated environment. Software testing of distributed systems even becomes a daunting task if manual testing is employed. This paper addresses the issue of automated testing of distributed applications by looking into the common challenges in the distributed systems testing followed by proposing a framework that automates the testing process.

Software testing is one of the most difficult tasks to assure the quality of software and plays a vital role in the SDLC process to ensure that it adheres to the client or customer needs. When it comes to the distributed applications environment, software testing is considered as backbone for applications. Testing the distributed software application is much more difficult as compared to testing standalone software applications, because the distributed system behaviors are dynamically changed with respect to time and platforms. There are several key challenges linked to testing the distributed applications; e.g., the same test run executed frequently on the same scenario with same input, may generate different outputs. This happens due to the non-linear behavior of the distributed systems i.e., event timing can also affect the end results. The functional and non-functional requirements play key role in web applications testing.

Generally, the term Software under Test (SUT) refers to the application that needs to be tested to determine correctness of its operations/functionality. For this purpose, a correctness-centered approach needs to be employed to ensure that software meets the software quality assurance requisites. For this purpose, the "design for testability" rules are applied on the specifications to prepare a suitable test run. Test harness is an umbrella term used to refer to collection of software and input data variation datasets to be used for software testing. Test harness is usually applicable at the level of unit testing

and is based on the concept of executing software under varying conditions of input followed by observing the correctness of the produced output. Test harness consists of two key components known as test script repository and test execution engine. Multi-agent is a predestined technology which synergizes the power of autonomous computational components that have control over their behavior and mutually work to achieve their specific individual objectives.

Within this context, multi-agent systems impart more functionality and intelligence to computer systems as agents operate in a flexible and rationale manner through interaction with other agents or even humans. With agents having the capacity to operate in parallel, multi-agent systems not only speed up efficiency but also enhance reliability and robustness. Also, since multi-agent systems follow a more modular structure with each agent performing independently of the other, it addresses the problem of scalability as programmers can simply add new agents to enhance functionality of a program.

Current research shows that multi-agent systems are less costly than most of the centralized systems as they are composed of subsystems of low unit cost. These agents can be reused in different scenarios without the hassle of coding new programs from the scratch for newly emerging scenarios. Agents act autonomously in order to solve complex problems in real time that are beyond the scope of human capability. Multi-agent systems make use of the expertise of individual agents coupled with their ability to collaborate, cooperate and also interact with one another to formulate powerful systems beyond the scope of individual agents alone. It is this very feature of agent technology that forms the basis of multi-agent systems [14]-[16].

II. RELATED WORK

A software agent is a computer code or program that operates in the dynamic environment on behalf of an additional entity either human or computational. Software agents are designed to be autonomous, proactive, collaborative and operate asynchronously as well as in parallel fashion. These agents communicate through the message passing instead of method invocation. Since agents run autonomously and cooperatively among the other agents, so they may run properly by themselves and may perform inaccurately while working together. For the specific nature of the software agents, it is tricky to apply the normal software testing and debugging technique to software agents. Software agents require special techniques dedicated for testing.

Collins et al. [1] combine the advantages of Distributed Software Development and agile software development methods and state that the old works did not cover the scenario where the tasks related to testing in distributed environments with individual work groups that are located on different spatial locations. DiLucca et al. [2] performed analysis of different testing method for web applications with respect to functional and non-functional requirements. The research highlights that functionality testing of a web application relies on the following basic aspects: testing models, testing levels, testing strategies, test cases and test processes.

Clune et al. [3] state that systematic testing is crucial for the complex scientific software systems. The objective of this study was to analyze testing techniques for scientific software so that they can be maintained and evolve in a systematic way. The complex scientific software evolves due to the growing requirements and the developers introduce new line of codes in the software to meet the additional changes. Every new line in the system carries risk of introducing bugs and may result in performance degradation. Identifying and fixing such bugs require additional cost, time and effort. Early detection of bugs could considerably reduce the efforts needed to implement a correction. The authors suggest that scientific software should be covered with systematic testing.

Chu et al. [4] state that it might become more informative to perform the testing inside live situations. The vivo testing focuses on easing the burden by simply sharing load across a number of multiple instances of the software application. This approach elevates the scope in vivo testing from a single instance to a couple of instances. A central server coordinates this effort by monitor the size of the community and collecting the test results. This approach extends to presented in vivo testing framework which is called Invite. Applying this distributed method to in vivo testing technique help amortizing the workload above many instances cause higher performance impact lacking of sacrificing the quantity of tests being conduct. In addition, in vivo testing supports testing as many permutation of states as possible, in the hope that it would encounter the ones that are not correctly handled by the code. Testing software applications that use nontrivial databases are increasingly being outsourced to test centers for reducing cost and achieving higher quality [5]. However, for security reasons the sensitive information is not shared with the test centers to perform testing with the real data. The author introduced a novel approach called PISTIS for minimizing database for software testing tasks. PISTIS used on a weight-based information clustering formula that divide the test data by making using of pertinent information obtained by way of program evaluation. This way a large database is reduced to a few centric objects. This main benefits of this testing is that tests are not dependent able the designer and the tester.

Agent oriented software engineering methodologies provide us a platform to develop agents based systems. These methodologies mainly focus on development rather than the testing. It is not possible to map all agent properties e.g. autonomy, reactivity etc. to object oriented constructs. Therefore, a proper testing technique for agent-based software solutions is needed. Sivakumar et al. [6] propose an effective and specialized testing technique for agent-based systems. The proposed technique focuses the main attribute of an agent which is role. It follows a v-based model which starts from requirements and ends at role-based acceptance testing. The proposed approach provides better solution for industrial, commercial, medical, networking and educational applications related problems. The purpose of software product lines is to create efficient products in a systematic manner, and Uzuncaova et al. [7] build on one such systematic technique referred to as "scope-bounded testing" in order to develop a novel specification based methodology far efficiently creating tests regarding products within a software product line.

Lv et al. [8] propose hybrid approach that uses Adaptive Testing and Random Partition Testing is an alternating manner. The motivation for this approach is that both strategies are employed such that the underlying computational complexity of Adaptive Testing is reduced by introducing Random Partition Testing into the testing process without affecting the defect detection effectiveness. A case study with seven real-life subject programs is presented in the study. The Adaptive Random Testing is to enhance the failure-detection ability of Random Testing. Chen et al. [9] consider and compare towards the performance regarding adaptive randomly testing from code coverage perspective.

Eassa et al. [10] introduce a dynamic testing tool that use a temporal logic assertion language for detecting run time errors in agents and agent-based systems. The proposed technique is based on the syntax and semantic of the temporal logic assertion language. A dynamic testing tool has been built and tested for ascertaining its effectiveness against its use as a dynamic testing tool.

Serrano et al. [11] propose a framework to record and order interactions among agents in a MAS. To capture the interactions in the distributed MAS, the proposed solution uses a generic registration layer based on aspect oriented programming. In MAS, the distributed events are ordered using vector clocks which are combined with graph theory to produce abstract graphs. The framework is based on debugging errors in different testing environments and removes the matching errors.

Hao et al. [12] introduce the technique of performance assessment and offered a platform of agent-based functionality testing about web based services. The study includes some specific features Test Flow Generator, Scenario Creator, Test Manager and Load Generator Agent. Communication and the coordination between distributed testing components are more complex features. The typical reactions of such systems are the generation of errors such as time outs, locks, observability, controllability and synchronization problem. Azzouzi et al. [13] show how to cope with these problems by using a distributed testing method including timing constraints. Afterwards, a multi-agent architecture is proposed to describe behavior of testing a distributed chat group application on high level of abstraction. The study focuses on the temporal properties that specify the time required for exchanging messages between the various components of the distributed test applications.

Based on the literature review, we observed that there is a need for speedy execution of the testing activities to curtail the testing costs and save time. In this regards, testing of application using the distributed environment can be a viable and speedy solution. Multi-agent based framework is proposed to address the issue of performing software testing in a robust manner. In view of this, a suitable proposition could be to formulate a network of multi-agents that possess learning capabilities and are intelligent as well as collaborative. The cooperative nature of the multi-agents in this study is expected to enhance the multi-agent based software testing frameworks.

III. FUNCTIONAL TESTING FRAMEWORK

In view of the problem statement mentioned in the last paragraph of the previous section, we propose a multi-agent based functional testing framework within the distributed environment. The proposed framework only encompasses the functional testing and does not cater for non-functional or quality attribute testing. The functional testing is primarily based on input-output relationship. While testing a software, the obtained results are matched against the expected/desired results and based on this comparison the decision whether the test has passed or failed is taken. Our proposed framework is shown in Fig. 1. The detail description of the various components/artifacts that constitute our framework is described below.

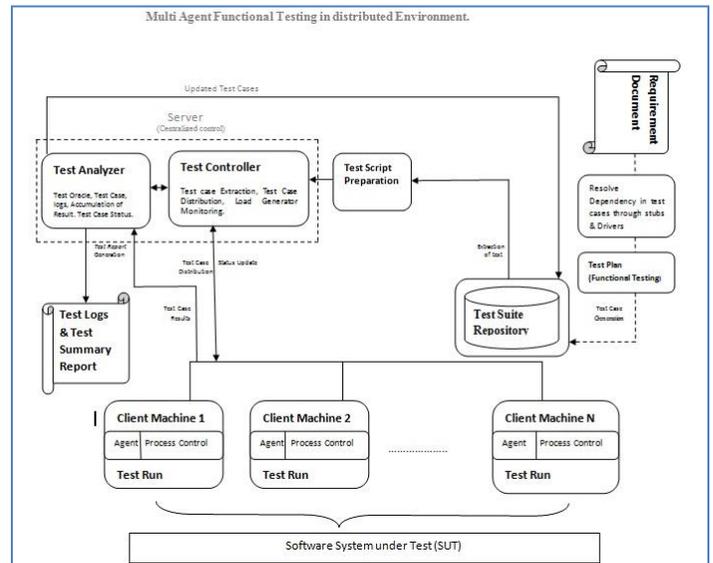


Fig. 1. Framework for multi-agent based functional testing in distributed environment.

A. Centralized Client-Server Environment

Network server is a key component to support the distributed environment (i.e. Client Server Environment) by providing a centralized control. The purpose of a server is to share data or hardware and software resources among the clients. This architecture is called the client-server model. In our framework, we imply that the client server environment is already in place and we do not deal with any of its hardware configuration or network protocol improvement. We are using distributed environment as a test bed. Our proposed framework supports parallel testing. It is used for speedily performing the testing activities by different distributed bunches of test cases across the multiple client machines. Parallel testing means testing multiple applications or subcomponents of an application concurrently to reduce the time required to test the entire system. Parallel tests consist of two or more parts (projects or project suites) that check different parts or functional characteristics of an application.

B. Test Suite Repository

Test Suite Repository maintains all the test cases corresponding to different software functionalities. It is assumed that test cases have already been prepared/generated by the test engineers either automatically or manually from the software specifications. In this study, we assume that the necessary test cases to be run for an application have already been made available in the test repository. And the dependency among different test cases has already been resolved using appropriate measures. The test plan used in this study merely corresponds to the functional testing.

C. Test Oracle

In software testing paradigm, the Test Oracle that determines whether a test has passed or failed based on certain criteria bears three capabilities: a generator, a comparator and an evaluator. The generator furnishes the expected result for a test case which is examined by the comparator against the obtained result. Finally, the evaluator then determines whether the comparison was successful or not. The key limitation of the oracles is that they can be applied only on a small subset of all the possible inputs and outputs pairing. Thus, it makes them suitable for small scale testing activities.

D. Test Script Generator

The task of test script generator in our framework is to prepare an executable test scripts (e.g. in HTML Format) for each and every test case. We can also assign/add priority with the test cases in order to decide upon their order of execution. A test script in software testing is a set of instructions that will be performed on the system under test (SUT) to verify that the system functions as expected. The test scripts can be either generated automatically or prepared manually. There are several testing tools such as Selenium and QTP (Quality Testing Professional) that can generate test scripts.

E. Multi-Agents

Software agents are programs or code snippets that are placed across the network and have their own control and goals. There are several types of agents and among them Interface agent, Information agents, Heterogeneous agent, Mobile agent, Reactive agents, Collaborative agents are commonly used. Agents are generally categorized based on their properties. Agent should be message passing, collaborative, proactive and autonomous. In this study, agents are supposed to receive, execute and send results of the test cases. Intelligent agents deployed on the client machines will decide to load the relevant software component/artifact based on the received test case and the procedure or plan to execute them.

F. Test Controller

Test controller is an important module of our framework. The main task of test controller is to extract the test cases from the test suite repository and prepare the test scripts accordingly followed by distributing test cases on different client machines. Test controller continuously receives the status of test cases

from different client machines and updates their status accordingly. Test controller is also connected with another important artifact known as test analyzer. Test Controller has a two way communication channel with the test analyzer and client machines, which enables it to send and receive data across both the modules. Therefore, we can safely assume that Test Controller is the core module of our testing framework which initiates, executes and monitors the whole testing process.

G. Test Analyzer

After the Test Controller, Test Analyzer is the next important part of our framework. It keeps record of the number of test cases passed and failed. Such results are used to determine the level and quality of a software product. It is connected with test controller and different client machines. It analyzes the output of client machines and accumulates results and then it sends failed and deferred test cases back to the test suite repository so that test controller can fetch them again and try their re-run. It also helps test controller to update the test cases in test suite repository so that the successfully executed test cases should not be executed again. Test Analyzer also maintains test case logs which are needed by the Test summary module to generate different summary and analytic reports.

IV. IMPLEMENTATION AND EXPERIMENTATION

To validate our proposed framework, we prepared a test bed consisted of one server machine and one another machine which hosts multiple client machines in the form of virtual machines. For this purpose, we created four virtual machines on the computer using VMware Workstation 10.0. All these virtual machine were connected to the server and we installed soft bots (i.e., agents) on each of the client machine as well as on the server side. These agents are in fact a piece of code to communicate and coordinate with other agents deployed on other client machine as well as server machine.

A. Case Study

As a case study to perform the testing activities, we used a web-based application for "Employee Management". This web-based application was prepared for a company "Cafedunord". Employee management is a shift management platform to prepare a shift roaster for different employees of an organization. The manager can create different shifts and can allocate them to different employees. Employees receive emails about their whole weekly or monthly working schedule. Manager can also generate shifts related work report for employees. We prepared 50 functional test cases which correspond to different functionalities provided by the software such as login (authentication), adding employee, and assigning tasks to the employee, preparing shift schedule, making duty rostrum etc. Some of the selected test cases are appended as Annex-1 to this report. The prepared test cases were then converted into test scripts using the Selenium testing tool. The test cases were run on the Selenium using the "record" option, for which Selenium prepared the test scripts accordingly. We saved these test scripts for future use. The test scripts, which were in executable form, were then passed on to the test controller for distribution to the client machines for their execution.

TABLE I. DESCRIPTION OF TEST CASES BASED ON SYSTEM FUNCTIONALITY

Sr #	Functionality	Module/ Webpage	Test cases	Description of the selected test cases provided in Annex-1
1	Employee Authentication	Login.aspx	6	a. Check for valid Username and Password. b. Test with invalid Username and Password.
2	Registering New Employee	Add Employee.aspx	8	Valid user name and particulars for New employee.
3	Shift Management	Shift Template.aspx	13	Different shifts allocated different staffs members.
4	Scheduling	Employee Scheduling.aspx	17	Scheduling/rotation of employees by admin
5	Generate Reports	Reports.aspx	6	Generate the reports when required.

These sample test scripts generated by the Selenium are also attached as Annexure-II to this report.

B. Test Case Execution

We selected a web-based application named “Cafedunord employee management system” to test its functionalities for validation of our framework. It is actually a shift management application for the employees. We can create different shifts e.g. day-shift or night-shift and then we can allocate them to different employees. Employee receives their whole weekly or monthly schedule by email. We have created 50 test cases in the beginning from which we selected specific test cases to perform testing which are described in Table 1. Our framework is a distributing functional testing with multi-agent in which we have a server and a small bunch of client machines. In our test bed, the client machines are not physical machines but are virtual machines created using VM-ware Workstation. For the testing purpose, we use Selenium automation testing tool. Selenium is a suite of tools to automate web browser across many platforms. This testing tool is free and open source software.

Running tests cases in parallel calls for two things: an infrastructure to spread the tests and a framework which will run these kinds of tests with parallel in the given infrastructure. So, we can first make a distributed infrastructure and then create several tests cases, which will be executed in this distributed test environment. Selenium is powerful tool which can work with distributing environment and we can also record a test script for a particular test case. When we have to verify a test case, we will run its correspondent test script. Selenium automation testing tool and multi-agents are be deployed on server and client machines. Software agent is in fact a piece of code snippet that monitors and controls all the work related to communication and collaboration among the network nodes. To begin with the testing, all of our test cases are placed in the

Test Suite Repository. Test Controller fetches the test cases from the repository. In Test Controller milieu, we use Selenium to create test scripts for those cases and agents will distribute those test scripts among the client machines depending upon the load on each machine. Every client machine will verify the test script using testing tool and will provide the result. Agents can share those results with each other through message passing which can create a speedy execution of test cases as well as reliable and robust testing environment. Every test result is sent to the Test Analyzer. Test Analyzer updates the repository with the test case status whether the test has been passed or failed. Test Controller can run the failed test cases again at a later stage. The execution summary of passed and failed test cases is provided in Table 2 whereas, statistics about the percentage of passed and failed test cases in shown in Table 3.

TABLE II. TEST CASES EXECUTION SUMMARY

Functionality	Test cases	Passed	Failed	Remarks
Employee Authentication	6	5	1	The logoff functionality was not working properly.
Registering New Employee	8	6	2	Change password option on the first login attempt was not active. Also employee's roles modification was not being carried out.
Shift Management	13	13	0	All the test cases passed in this module.
Scheduling	17	16	1	Erroneous behavior observed while assigning off days. Same off day could be assigned to all the employees.
Generate Reports	6	4	2	Two of the summary reports in the menu list did not generate anything.

TABLE III. PASSED AND FAILED TEST CASE EXECUTION SUMMARIES (PERCENTAGE)

Test Cases	Quantity	Percentage
Passed	42	84 %
Failed	6	12 %
Deferred (Error in test script)	2	4 %
Total	50	100 %

Graphical representation of Table 3 is shown in Fig. 2.

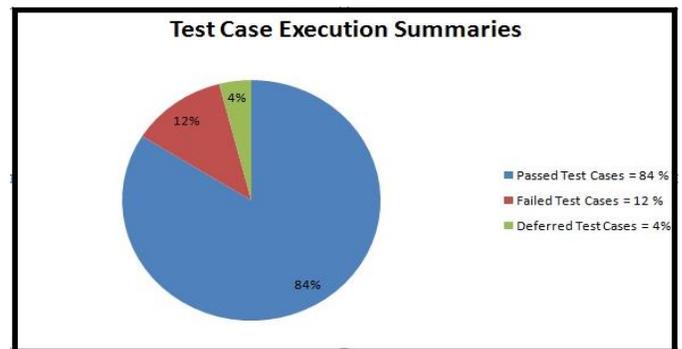


Fig. 2. Test execution summary.

A time based comparison of the test execution on monolithic and multi-agent based system in the distributed environment is shown in Table 4.

TABLE IV. EXECUTION TIME ANALYSIS USING MONOLITHIC AND MULTI-AGENT SYSTEM

Functionality (Module)	Input fields	Test Cases	Test Run (Execution) Time on Monolithic Environment (in seconds)	Time of Execution using multi-agents (i.e., JADE agents in Distributed Environment) (in seconds)		
				2 Client machines	3 Client machines	4 Client machines
Employee Authentication	2	6	12.86	8.21	6.91	6.72
Registering New Employee	10	8	150.41	84.7	70.13	48.12
Shift Management	4	13	34.12	23.78	16.37	15.93
Scheduling	5	17	180.24	104.89	75.12	60.13
Generate Reports	3	6	41.23	26.36	19.52	18.12

A graph showing the time of execution using multi-agents on multiple machines and on a standalone system which had no agents deployed on it is illustrated in Fig. 3.

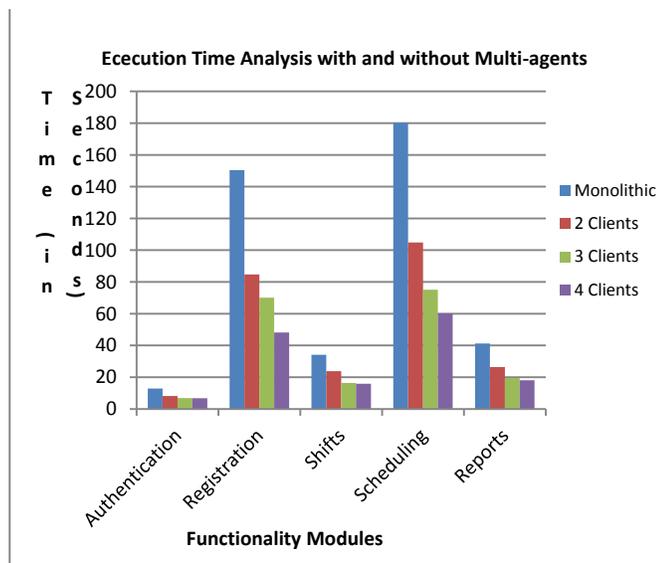


Fig. 3. Execution Time Analysis using monolithic and multi-agent system.

V. DISCUSSION

The framework proposed in this study will facilitate the regression testing of the applications which undergo several releases/builds by automating the testing process. The core reason for using collaborative multi-agent in this research was to make the testing activities faster and economical. The multi-agent approach has been widely used in the domain of computational intelligence as it has been proved to be an

adequate approach where cooperative traits of specialized agents (or bots) are required. Since multi-agent are themselves distributed in natural and operate autonomously in different environment therefore, they support better utilization of computational resources.

In this study, we have employed a three-layered multi-agent architecture. We used JADE (Java Agent Development framework) to deploy multi-agents. A comparative analysis of the similar studies shows that our framework supports robustness and enhances test execution speed besides supporting controllability. Further, our model is scalable as well. Hence, our model supports the key performance measures as reported by other researchers. The distinctive part of our study is that it focuses on reducing Test execution time by utilizing more and more resources. A comparative analysis of our framework with other studies is provided in Table 5.

TABLE V. COMPARATIVE STATEMENT OF CONTEMPORARY STUDIES

Ref	Purpose	Evaluation Parameters	Benefits/Strengths
[6]	Agent oriented software testing approach is presented to enhance efficiency and quality of software products.	Efficiency	Agent based approach helped achieve better management of the software testing process.
[11]	Different methods for debugging the multi-agent system for software testing.	Debugging helps enhance the quality of software.	This paper proposes the methodology to test a relational database server as a central storage mechanism.
[12]	This approach enhances performance testing on distributed agent based web services.	Reliability, Accuracy, Dynamicity	It also combines the features of performance testing as well as functional testing and improves the system with respect to reliability and accuracy.
[13]	The proposed approach is used to improve the correctness of testing in distributed systems.	Coordination, Communication, Controllability, Observability	Several problems influencing fault detection during the conformance testing process arise. So this approach reduces the problems of coordination and improves the controllability of system and enhances the fault detection.
[10]	Build a temporal logic assertion language to help detect the identified errors as well as build a dynamic analyzer based on temporal assertion language for testing agents.	Reliability, Communication, Fault Detection	The main advantage of using agent based testing is that it can generate test cases automatically and it can run continuously. This framework is more scalable in dealing with the distributed environment.
Ours	Our approach use collaborative multi-agents the core	Controllability, Efficiency (Speed)	Our framework supports robustness and enhances test

reason for using collaborative multi-agent in this research was to make the testing activities faster and economical. Our study is focuses on reducing Test execution time by utilizing more and more resources.	Scalability, Observability	execution speed besides supporting controllability. Further, our model is scalable as well. Hence, our model supports the key performance measures as reported by other researchers.
--	----------------------------	--

VI. CONCLUSION

In this research we presented a multi-agent based framework to perform the functional testing in the distributed environment. The core reason for performing testing activities in the distributed environment was to reduce the cost, time and efforts ordinarily required to perform functional testing. We choose to deploy multi-agents on the client machine and server side to better coordinate the testing activities. To validate our proposed framework, we created 50 test cases for a web application called “Cafedunord”. All the test cases were passed through Selenium automation testing tool to generate their test scripts which were also run through Selenium testing tool using the agents deployed on different client machines. The experimental results show that time to execute test cases was reduced by a proportional factor depending on the number of client machines.

REFERENCES

- [1] E. Collins, G. Macedo, N. Maia., and A. Dias-Neto, “A Industrial Experience on the Application of Distributed Testing in an Agile Software Development Environment”, In *Global Software Engineering (ICGSE) on IEEE Seventh International Conference*, pp-190-194,2012.
- [2] G.A. Di Lucca, and A, R Fasolino, “Testing Web-based applications: The state of the art and future trends”on*Information and Software Technology*, vol 48, pp-1172-1186, 2006..
- [3] T. Clune, M. Rilee, and D. Rouson, “Testing as an essential process for developing and maintaining scientific software”.on In *The 2nd Workshop on Sustainable Software for Science: Practices and Experiences*, 2014.
- [4] M. Chu, C. Murphy, and G. Kaiser, “Distributed in vivo testing of software applications. In *Software Testing Verification, and Validation*”, on *1st International Conference on*, pp. 509-512, 2008
- [5] B. Li, M. Grechanik, and D. Poshyvanyk, “Sanitizing and minimizing databases for software application test outsourcing. In *Software Testing, Verification and Validation (ICST)*” on *IEEE Seventh International Conference on*, pp. 233-242,2014
- [6] N. Sivakumar, and k. Vivekanandan, “Agent Oriented Software Testing–Role Oriented approach” on *International Journal of Advanced Computer Science and Applications*, vol 3, 2012
- [7] E. Uzuncoava, S.khurshid,, and D. Batory, “Incremental test generation for software product lines” on *Software Engineering, IEEE Transactions*, vol 36, pp. 309-322,2010.
- [8] J. Lv, H. Hu, K. Y. Cai, and T. Y Chen,” Adaptive and Random Partition Software Testing”,2014.
- [9] T. Chen, F., Kuo, H., Liu and E. Wong, “Code coverage of adaptive random testing”, on *IEEE Transactions on Reliability*, vol 62, pp. 226-237.2013.
- [10] F.E, Eassa., L.J Osterweil, M. A, Fadel, S. Sandokji and A Ezz,” DTTAS: A Dynamic Testing Tool for Agent-based Systems” on *Pensee Journal*, vol 76. 2014.
- [11] E. Serrano., A. Munoz. And J. Botia. “An approach to debug interactions in multi-agent system software tests.” On *Information Sciences*, vol 205, pp. 38-57,2012.
- [12] D. Hao, Y., Chen., F. Tang, and F. Qi. “Distributed agent-based performance testing framework on Web Services” In *Software Engineering and Service Sciences (ICSESS) on International Conference on*, pp. 90-94, 2010.
- [13] S., Azzounzi., M., Benattou, and M.E.H Charaf, “A temporal agent based approach for testing open distributed systems.” on *Computer Standards & Interfaces*, vol 40, pp. 23-33.2015.
- [14] G.,Weiss, “Multi-Agent Systems.” MITPress, Cambridge, MA.1999.
- [15] M. F. Malik & M.N.A. Khan, “An Analysis of Performance Testing in Distributed Software Applications.” *International Journal of Modern Education and Computer Science*, 8(7), 53, 2016.
- [16] M. Wooldridge, “An Introduction to Multi Agent Systems.” Wiley Second Edition.ISBN 978-0-470-51946-2.2009.

Modeling and Implementing Ontology for Managing Learners' Profiles

Korchi Adil

Laboratory of Signals, Systems and
Components
Faculty of Science and Technology
University Sidi Mohamed Ben
Abdellah,
Fez, Morocco

El Amrani El Idrissi Najiba

Laboratory of Signals, Systems and
Components
Faculty of Science and Technology
University Sidi Mohamed Ben
Abdellah,
Fez, Morocco

Oughdir Lahcen

Department of Mathematics, Physics
and Informatics
Polydisciplinary Faculty -Taza
University of Sidi Mohamed Ben
Abdellah,
Taza, Morocco

Abstract—This paper presents an issue that is important to consider when developing a learning environment whose field is constantly evolving mainly in terms of the use of training platforms. Research in this field has enabled the successful use of information technologies for the benefit of human learning, while placing the learner at the heart of pedagogic situations. It is also an environment that integrates human agents (tutors, learners) and artificial (computers) and allows them to interact locally or through computer networks, as well as conditions for accessing local or distributed training resources. Moreover, several computing environments for human learning (CEHL) platforms are available on the web for free access. These platforms are environments that offer a learner a multitude of courses in various formats in order to satisfy the learner's desire to learn. Several CEHL platforms are available on the web for free access. But learning itself is not enough and that is why a new generation of advanced learning systems that integrate new pedagogical approaches giving the learner an active role to learn and acquire knowledge has emerged by offering more Interactivity and incorporating a more learner-centered vision. These new generations of advanced learning systems adapt to learners and their profiles by taking into account their cognitive, intellectual and motivational characteristics. An adaptation that cannot be achieved without the complicity of ontological engineering, which plays a very important role in the sharing of knowledge between humans and computers, and between computers and sharing, and reuse of concepts through computational semantics. By the same way, this paper aims at creating a process of modeling and managing profiles of learners based on ontology whatever the learning situation may be. This management process is implemented in computer's environment based on the learner's ontology that supports the learner by detecting the gaps in several factors in order to improve them and adapt the pedagogical content to the learner's profile.

Keywords—*Ontology; computing environments for human learning (CEHL)-Learner – Learner's Profile – XML/RDF – JENA API – OWL – PERFECT-LEARN – inference; Learner modeling – SPARQL - semantic links - concepts – sub-concepts*

I. INTRODUCTION

Several CEHL platforms are available on the open access web. They constitute environments that offer the learner a multitude of courses in various formats in order to satisfy the learner's desire to learn [1]. But CEHL must adapt to learners

and their profiles, and take into account their cognitive, intellectual and motivational characteristics of the learner.

In order to adapt the learning profile to the learning environment, we need to ask ourselves some questions such as:

- 1) What are the factors that characterize the learner's profiles ?
- 2) Among these factors, which ones are positive and which ones are negative ?
- 3) How can these factors be automatically detected and evaluated ?
- 4) What are the functional aspects of the learning process that depend on these factors ?
- 5) How can the functional aspects of the learning process be adapted to adapt to the learning profile and, on the other hand, improve the factors that characterize it ?

We will attempt to answer these questions in order to adapt the learning to the ontology-based learner profile.

II. LEARNING PROFILE AND ONTOLOGY

Just learning is not enough because the learner eventually gets tired of the heap of information he receives. The current CEHL allow the adaptation of the pedagogical content to the learning profile to a certain extent where the parameters that distinguish it are detected in the process through which it tries to learn. These parameters include behavior, preferences, cognitive level and interaction [2]. Each learner has his own way of learning which constitutes what is called his profile. And to encourage him to develop his knowledge, he must be placed at the heart of the pedagogical situation and take into account only the elements that really influence his learning. Such an operation can only be realized with the complicity of the ontologies that participate fully in the modeling of the profiles and modeling knowledge [3].

The ontology development process refers to what activities you need to carry out when building your ontologies. However, the ontology development process does not imply an order of execution of such activities. Its goal is to identify the list of activities to be completed. Usually verbs are used to refer to such activities [4].

III. PROFIL'S MODELING AND IMPLEMENTATION

Taking into account the characteristics of the learner requires a modeling of his profile. Such an operation helps to adapt the pedagogical content to the needs of each learner in order to evaluate his skills, behavior and interaction for generating new personalized learning situations. To summarize, the modeling process is a diagnosis on the traces of the learner and the learner's profile is the result of this diagnosis .

We have been thinking about modeling that will identify the learner's limitations, abilities and gaps in order to initiate learning situations appropriate to the learner's cognitive level. This modeling process is shown in Fig. 1 which represents the proposed ontology and its various semantic links that link concepts to sub-concepts [5].

Fig. 1 gives an overview of the Learner ontology. In order to implement it in the designed learning system, we have represented it in PROTEGE (an open-source ontology editor) to recover the XML (Extensible Markup Language) file and the graph associated with it. Fig. 2 schematizes this graph.

This ontology proposes a modeling way comprising several concepts linked semantically to each other namely:

A. Learner

Admits personal data to be defined in an XML file (Name, First name, Age, Sex, Email ...). The XML parser that we are going to create is none other than an XML analyzer that will:

- Find a data item in the XML file and read it to build a computer object.

Extract the data in xhtml format and display it.

- Check that the new extracted document is well-formed.

During the registration or authentication step of a learner, the system will need this parser to load learner's information. Once the file containing the various updated information about the learner is read, the system classifies it in one of the predefined levels according to the defined criteria.

B. Profile:

This concept is linked to four other concepts characterizing the learner, which are:

- The knowledge.
- The behavior.
- Interaction.
- Skills.

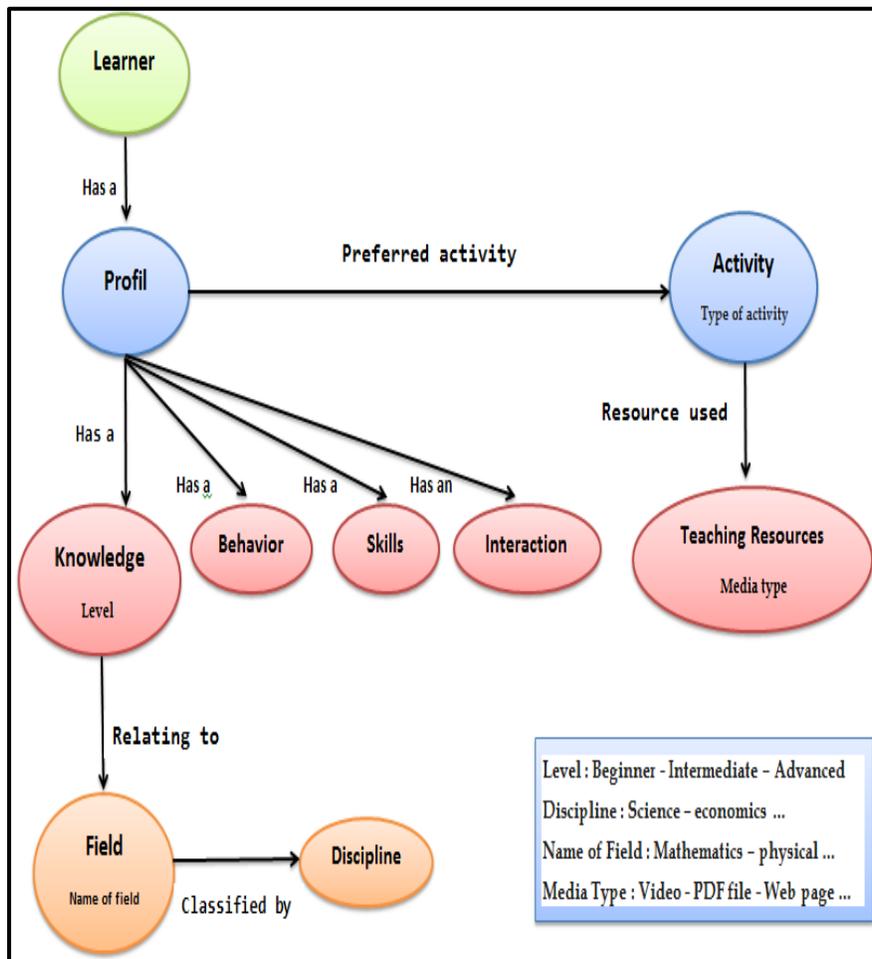


Fig. 1. Learner ontology elaborate.

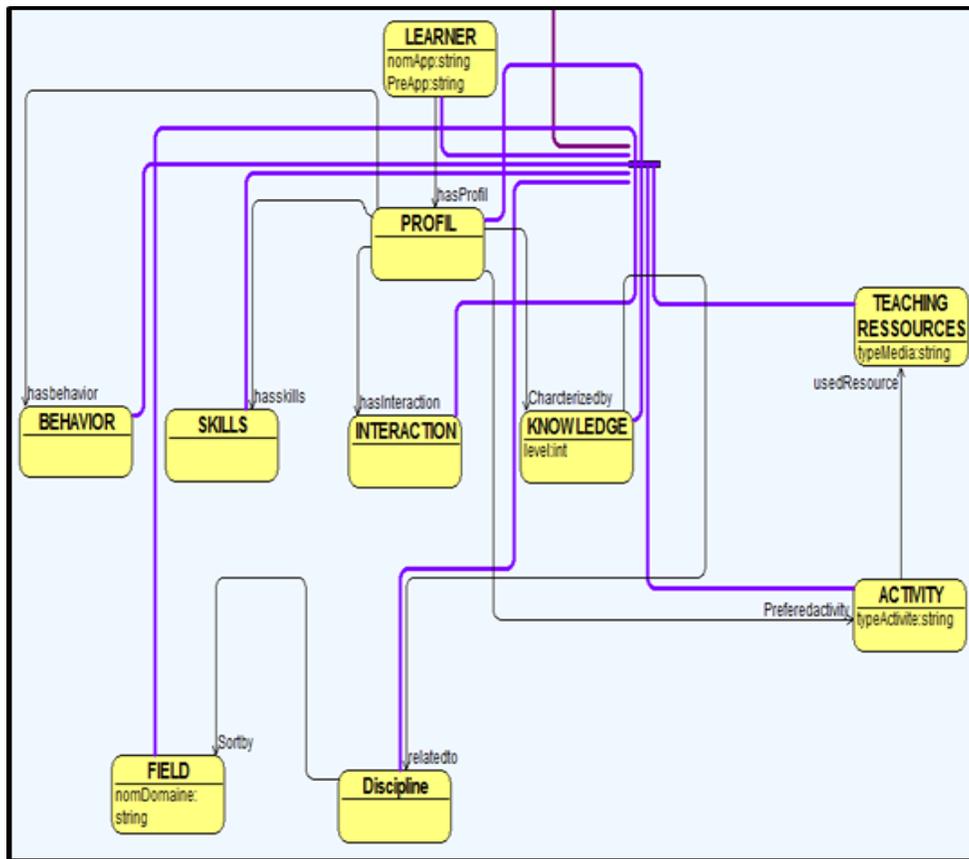


Fig. 2. Learner ontology graph elaborated in Protégé editor.

C. Learning Activity and Ressources

During a learning activity associated to a learner, an educational resource is assigned according to the learner’s preferences in accordance with his profile. This resource can be a media file, PDF file or others. Ontology participates in the representation and organization of those educational resources during a learning session.

D. Domain and Discipline

Our ontology classifies the different disciplines by fields of knowledge, if for example, the domain name is the Sciences then the discipline can be Mathematics, Physics or Natural Sciences.

At the end of each learning session, the system has a set of information about the learner that must be organized and saved using the ontology to generate the new profile that will be taken into account at the next session. The system will, then, interact intelligently with the learner by dynamically adapting the subjects to be presented to him according to the acquired results and the mode of learning that suits him best.

But long before that, we will discuss the process of the registration or authentication step of a learner. This process is shown in Fig. 3.

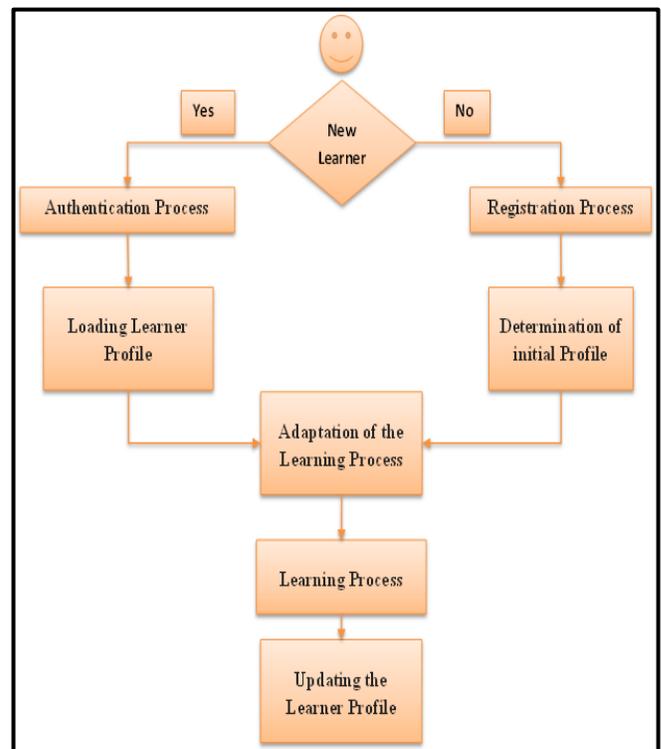


Fig. 3. Sequence diagram of registration or authentication step of the learner and profile update.

E. Registration or Authentication Step of the Learner

During the registration or authentication step, our system will need a parser to browse the XML file concerning the learner's personal information, to load it.

F. Registration Process

The case illustrated in the sequence diagram of the registration process (Fig. 3) is triggered by any candidate wishing to learn in our system. When the learner clicks on a link to sign up, the system displays a form and prompts the learner to provide certain personal information to register. Following this operation, the model is initialized and the learner becomes recognized into the system. Afterwards, it is necessary that the learner authenticate itself in order to be able to access in reserved space in the system and begin to learn.

G. Authentication Process

As soon as the learner connects to the learning platform, his identity is stored, which will later allow him to locate his workspace. Through this identity, the learner profile is recovered and is associated with the pedagogical activity in order to adapt it to the right profile. All this is done using XML parser and ontology.

H. Process Access to the Course

The designed environment structures and adapts resources according to the profile. These resources are displayed in a personalized way for each learner who solicits them. This adaptation is made possible through the use of the different languages gravitating around the XML technology as well as

the ontologies. This technology also supports multimedia content.

Each learner level listed in our system admits specific courses. Our ontology controls access to courses according to the level of the learner in such a way that he can only access courses of his level which is defined beforehand by the system.

When the learner clicks on the "access to the course" link, the system loads its profile and looks for the fragments related to the concept to be presented and according to the different characteristics indicated in the learning profile. It executes adaptation rules which are already predefined, applies the theme to the resulting XML file. Then it presents the content to the learner while observing his behavior and interaction according to the course presented. Fig. 4 gives a brief overview of this operation.

I. Learner Profile Update Process

The updating of the learner model consists of modifying the values representing the level of knowledge of the learner and this for a certain number of resources of a given concept.

The learner profile is updated before or after a learning session. Fig. 3 and 4 show exactly where this process is located. Several techniques are used to update the profile, namely:

- Level tests.
- Determination of learner interactions.
- Behavior determination.
- Type of preferred course material.

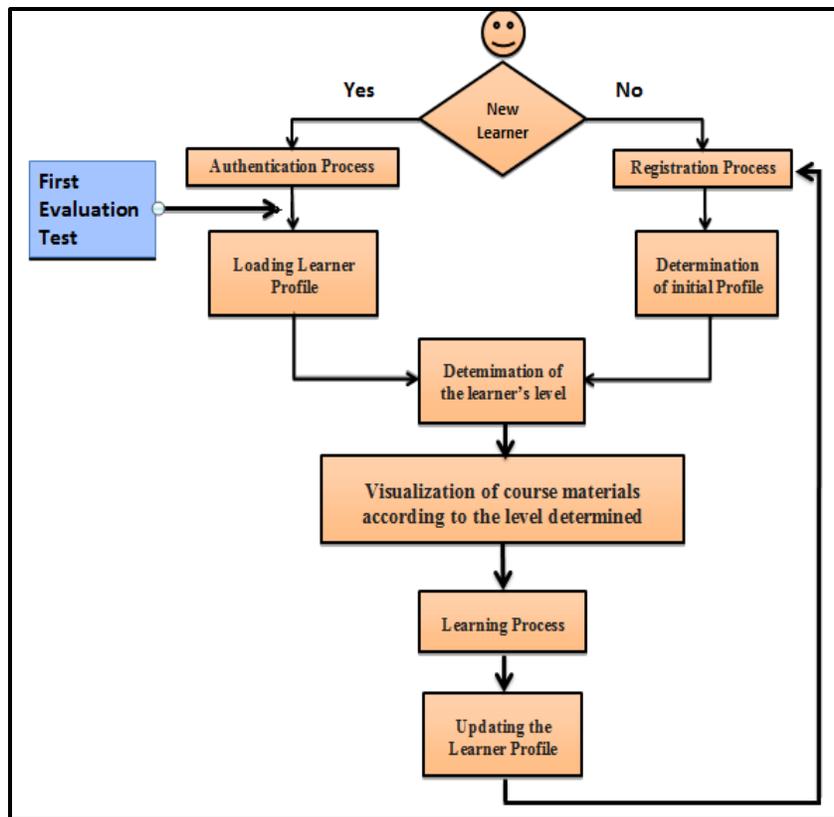


Fig. 4. Sequence diagram course access process.

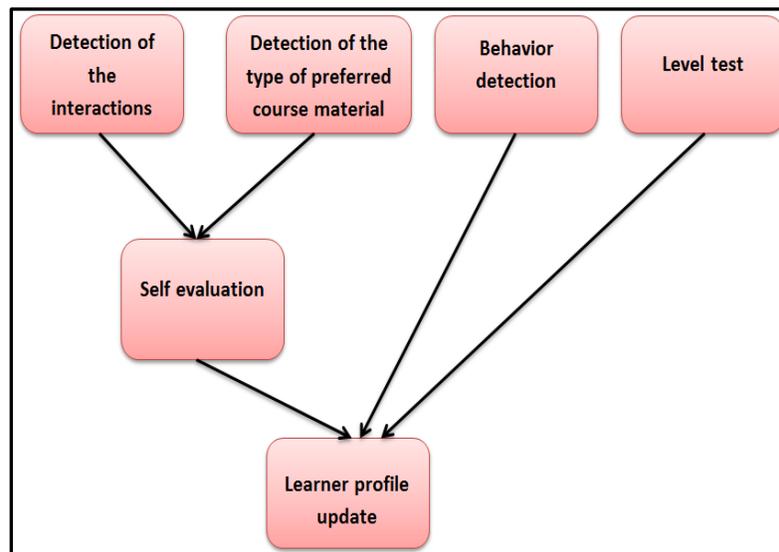


Fig. 5. Different stakeholders in updating the learner's profile.

Fig. 5 shows the different stakeholders that participate in updating the learner's profile. We will find in the following how these techniques take part in the update of the profile.

J. Level Test

It is a dynamic test in XML format controlled by a parser which chooses the type of question to ask the learner according to its characteristics and not a multiple choice questions test (MCQ). It will serve to validate learner's degree of knowledge of the concept. It provides the system with valuable information to adjust the next course. If the test is positive, the system will allow the learner to go to the next level, otherwise he must review the old courses until the latter is validated.

Consider the following example assuming that the learner is in level B and that he wishes to pass the level C test. If the system detects via the elaborate ontology that the learner has not answered the questions of level A correctly, then it is directly downgraded to level A even if it has already validated previously, level B.

It should be noted that the test is composed of several questions with multiple levels of difficulty, which will determine the actual score of the learner in a given test. For example, each section of the test will have a specific number of points according to the difficulty of its questions.

K. Interaction

The interaction of the learner with one of the proposed resources can be decisive in updating his profile. This is, for example, the slowness of the learner during the reading of a course. This slowness will be determined by the time spent in this activity, or by hesitation or change of a response relative to one or more questions during the test. This technique will also be applied to educational materials such as staying inactive for

a certain time, which leads to taking this parameter into account in order to determine the interaction of the learner.

L. Behavior

The behavior of the learner during a session can play a determining role in the detection and updating of the profile if it provides the necessary elements for this operation.

For example, an undecided learner can consult several resources in a time that the system may deem insufficient to assimilate a notion. It can also be happened during passing a test if he reviews the course to ensure that he has responded well. These data are used by the designed system to determine the behavior of a given learner.

After this overview on the ways in which the system updates the profile and detects behavior and interaction. We will discuss the historical component that stores information about a learner and his activity.

M. History

A learner model must store all the relevant information about learners, including knowledge and attitude [6].

Our environment keeps all information about a given learner (navigation, read resources, documents consulted, videos viewed, test past) to be exploited at any time by our ontology. The history of the course allows the learner to know his background. After each learner action, the browsing history is updated.

The following figure (Fig. 6) shows all the processes and learning phases that our ontology controls, including loading and updating the profile, determining the level and learning style of the learner.

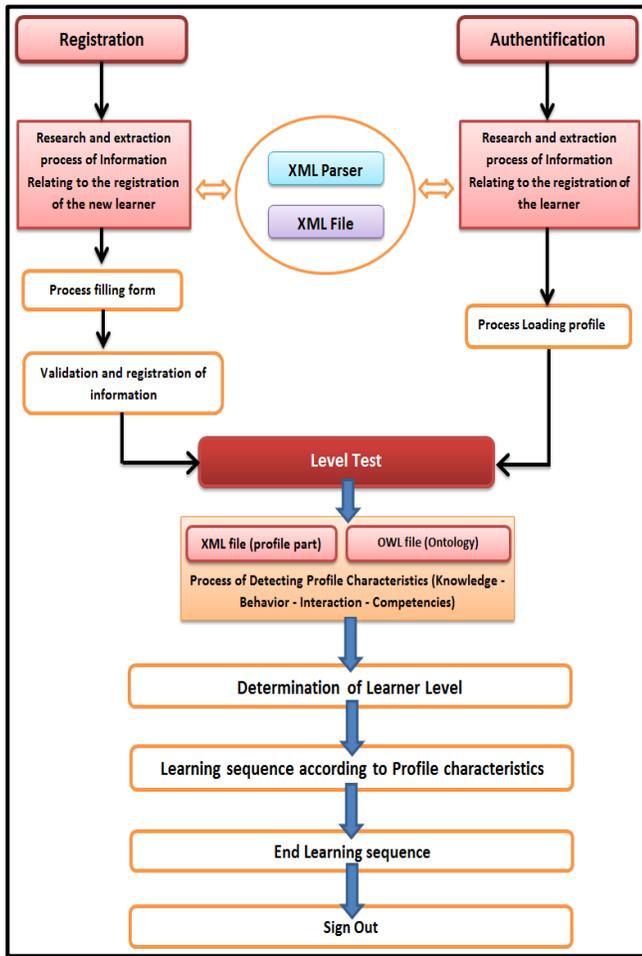


Fig. 6. Sequence diagram of the learning process by our ontology.

IV. DEVELOPMENT ENVIRONMENTS AND LANGUAGES

This section aims to describe the implementation elements of the different packages of the proposed deployment architecture. The following table (Table 1) summarizes our choices for the creation and manipulation of the proposed ontology.

TABLE I. CHOICE OF TOOLS FOR THE CREATION AND PROCESSING OF OUR ONTOLOGY

Tool / Language	Choice
Creation/edition of ontologies	PROTEGE Editor
Program access to ontologies	JENA API
Inference and reasoners	SPARQL Query Engine (ARQ)
Web language for Ontology (OWL)	OWL 2
Data storage and handling technology	RDF/XML
Application Server	GlassFish

V. ARCHITECTURE OF THE DEVELOPED SYSTEM "PERFECT-LEARN"

"Adaptive hypermedia systems are hypermedia systems which reflect some features of the user in a user model and use this model by adapting various visible aspects of the system to the user" [7].

The architecture of our system reflects the organization of the various elements, it includes (software, hardware, humans and information) and the relationships between these elements. This structure follows a series of strategic decisions taken during the design of this system.

The implementation of an ontology controlling learning as well as the storage and accessibility of learner information requires the use of a few techniques that are still little used and all contribute to the adaptation of the learner's profile for better learning.

Our application is developed with the NetBeans development environment, and for its deployment we used the web server GlassFish and this by generating an archive file associated with the Web application where the application is saved and the resources it needs. The overall architecture of our system is as follows (Fig. 7):

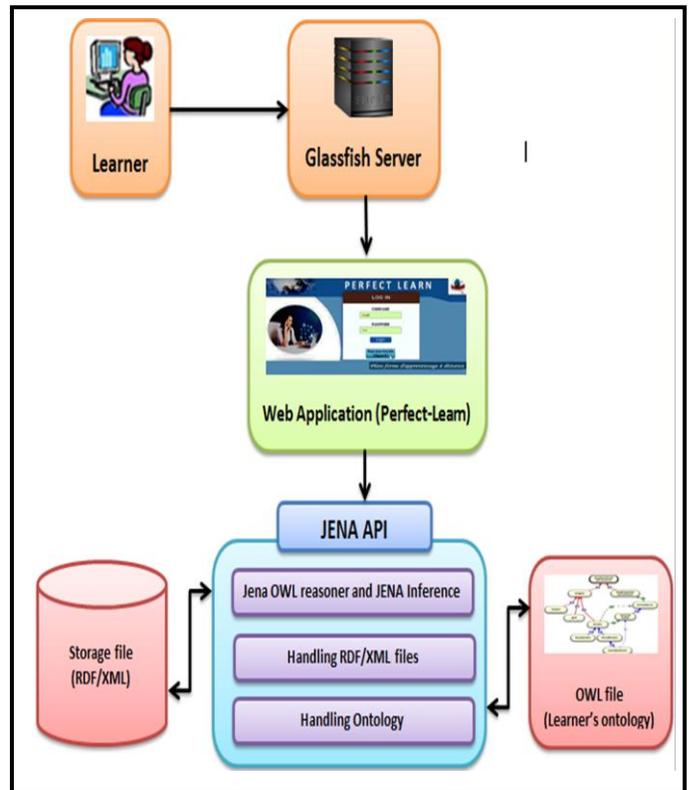


Fig. 7. Overall architecture of our system.

The prototype proposed in Fig. 7 is a tool for easily exploring the different logical structures of a set of documents in XML and RDF/XML format (RDF/XML is a syntax defined

by the the World Wide Web Consortium (W3C) to express an RDF graph as an XML document). It is built entirely in Java using a set of JENA APIs to manage access to RDF/XML documents. The latter offer tools for describing data and which can be of any type.

The JENA Framework (an open source Semantic Web framework for Java. It provides an API to extract data from and write to RDF graphs) is designed in a modular architecture. It offers several modules to meet the different needs of efficient manipulation of RDF data as well as those of ontology.

The JENA inference subsystem is designed to allow a range of inference engines or reasoners to be plugged into Jena. Such engines are used to derive additional RDF assertions which are entailed from some base RDF together with any optional ontology information and the axioms and rules associated with the reasoner. The primary use of this mechanism is to support the use of languages such as RDFS (Resource Description Framework Schema) and OWL which allow additional facts to be inferred from instance data and class descriptions. However, the machinery is designed to be quite general and, in particular, it includes a generic rule engine that can be used for many RDF processing or transformation tasks.

The overall structure of the inference machinery is illustrated below (Fig. 8).

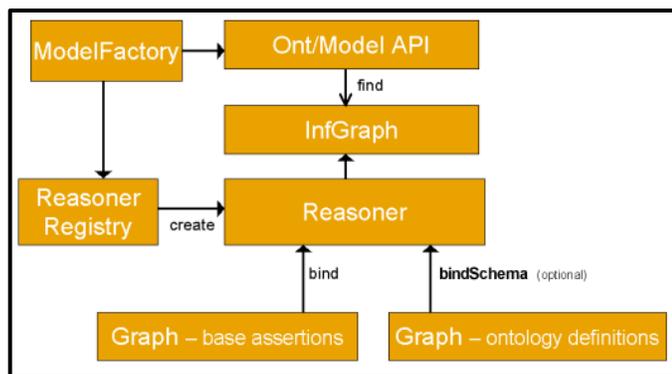


Fig. 8. Overview of inference support.

Applications normally access the inference machinery by using the ModelFactory to associate a data set with some reasoner to create a new Model. Queries to the created model will return not only those statements that were present in the original data but also additional statements that can be derived from the data using the rules or other inference mechanisms implemented by the reasoner.

As illustrated in Fig. 8, the inference machinery is actually implemented at the level of the Graph, so that any of the different Model interfaces can be constructed around an inference Graph. In particular, the ontology API provides convenient ways to link appropriate reasoners into the OntModels that it constructs. As part of the general RDF API we also provide an InfModel, this is an extension to the normal Model interface that provides additional control and access to an underlying inference graph.

Once you have an instance of a reasoner it can then be attached to a set of RDF data to create an inference model. This can either be done by putting all the RDF data into one Model or by separating into two components - schema and instance data. For some external reasoners a hard separation may be required. For all of the built in reasoners the separation is arbitrary. The prime value of this separation is to allow some deductions from one set of data (typically some schema definitions) to be efficiently applied to several subsidiary sets of data (typically sets of instance data).

A. Ontology Access Techniques

In order to concretize our approach, we had to use our ontology in our system “PERFECT-LEARN”. To do this, it had to be loaded from the JAVA code (using the JENA API):

```
Model m =
ModelFactory.createOntologyModel(OntModelSpec.OWL_
DL_MEM_RDFS_INF);
Model model =
ModelFactory.createMemModelMaker().createModel(null);
ReadFile OWL. The Namespace of our ontology must be
specified
InputStream in = (InputStream)
FileManager.get().readModel( m, inputFileNames );
if ( in == null )
throw new NotFoundException("Not found:
"+inputFileNames );
return load(in, "RDF/XML" );
InputStream in = (InputStream)
FileManager.get().readModel( model, inputFileNames );
if ( in == null ) {
throw new IllegalArgumentException("File: " +
inputFileNames
+ " not found");
}
model.read(in, "RDF/XML");
```

The central point of access is the “OWLOntologyManager”, which is used to create, load and access ontologies.

We first created an OWLOntologyManager object that will be used to load the ontology. Using the loadOntologyFromOntologyDocument () method, which takes the ontology local path as its parameter, it loads it into the ontology variable.

For the storage of learner information, we chose to create an RDF file for each learner to facilitate data management and retrieval via the loaded ontology. Our system manages any activity and records it in order to present it and exploit it. The query language SPARQL is present through its query engine ARQ RDF files. To run this engine, the following SPARQL command is used:

```
$ sparql Usage: [--data URL] [exprString | --query file]
```

The following figure shows; for example ; a part of the RDF file of a given learner ” who is already registered and that our ontology exploits.

```
1 @prefix apa: <http://localhost:8080/APA/OntoApp.owl#> .
2 @prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
3 @prefix foaf: <http://xmlns.com/foaf/0.1/> .
4
5 apa:Apprenant
6   rdf:prenom foaf:Said ;
7   foaf:Passe_test <Passe_test> ;
8   foaf>Password <Stringpass> ;
9   foaf:domain <domain> ;
10  foaf:eta <eta> ;
11  foaf:login <Stringlogin> ;
12  foaf:niveau <niveau> ;
13  foaf:nom apa:Apprenant .
```

Fig. 9. RDF file of a learner using the Friend of a Friend (FOAF) ontology.

We have opted for the use of FOAF (Friend Of A Friend) which is an RDF ontology that describes people and the relationships they have with each other. Used as reference by hundreds of other vocabularies, it is a central element of the Semantic Web. We preferred FOAF to reuse what already exists. FOAF contains all the properties necessary for the description of the persons and in our case the learners to know: Name, First name (Fig. 9).

The choice of the JENA APIs allowed us to manipulate the RDF, RDFS and OWL documents. They provide the necessary tools for the management and storage of the information circulating in our system. Among these tools is the inference engine which allows reasoning on the ontologies that handles the RDF/XML files. A validation of these files is then required via the W3C RDF validator. It ensures that an OWL document respects the RDF syntax, which already gives an initial indication of the validity of an ontology.

In our case, and for the part reserved for the management and storage of learners' information, use is made of the JENA APIs which enable certain operations on the learners' data to be performed such as updating, modifying, deleting, etc.

As for the OWL file, it is used to store ontology schemas and instances and not learner data such as their names, first names, levels, degree of interaction, etc. These are stored in RDF files.

The following is an excerpt from the RDF file relating to the storage of information:

```
public static void LearnerToOntoApp
(String firstname, String nickname ,String eta, String domain,
String Level,
String Passe_test, String login, String pass)
{ String ins="http://localhost:8080/APA/OntoApp.owl#";
String foaf="http://xmlns.com/foaf/0.1/";
String rdf="http://www.w3.org/1999/02/22-rdf-
syntax-ns#";
Model model = ModelFactory.createDefaultModel();
model.setNsPrefix("foaf", foaf);
model.setNsPrefix("apa", ins);
model.setNsPrefix("rdf", rdf);
Resource reso1 = model.createResource(ins+"Learner");
Resource reso2 =
model.createResource(foaf+prenom);
Resource reso3 = model.createResource(eta);
Resource reso4 = model.createResource(domain);
Resource reso5 = model.createResource(level);
Resource reso6 = model.createResource(Passe_test);
Resource reso7 = model.createResource(login);
Resource reso8 = model.createResource(pass);
Property prop1 = model.createProperty(foaf+"name");
Property prop2 =
model.createProperty(rdf+"nickname");
Property prop3 = model.createProperty(foaf+"eta");
Property prop4 =
model.createProperty(foaf+"domain");
Property prop5 = model.createProperty(foaf+"level");
Property prop6 =
model.createProperty(foaf+"Passe_test");
Property prop7 = model.createProperty(foaf+"login");
Property prop8 =
model.createProperty(foaf+"Password");
model.add(reso1,prop1,reso1).add(reso1,prop2,reso2).add(reso1,prop3,reso3)
.add(reso1,prop4,reso4).add(reso1,prop5,reso5).add(reso1,prop6,reso6)
.add(reso1,prop7,reso7).add(reso1,prop8,reso8);
System.out.println("-----");
model.write(System.out,"N3");
System.out.println("-----");
System.out.println("-----");
Try
{ Writer writer = new FileWriter("Learner"+name+".rdf");
model.write(writer,"N3");
}
catch(Exception a){
System.out.println("Erreur : Generation of RDF\n Plus Precis
:"+a.getMessage());
}
System.out.println("Generation of RDF");
System.out.println("-----");
}
```

The result of executing the RDF file is as follows:

```
@prefix apa: <http://localhost:8080/APA/OntoApp.owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
apa:Learner
rdf:nicknamefoaf:Said ;
foaf:Passetest<Passe_test> ;
foaf:Password<Stringpass> ;
foaf:domain<domain> ;
foaf:eta<eta> ;
foaf:login<Stringlogin> ;
foaf:niveau<level> ;
```

B. 8.2 - Self-Assessment Process

The self-assessment process occurs throughout the learning session. Its role is to detect the interaction and behavior of the learner in a given activity. It provides the system with accurate and valuable information about the learner during his pedagogical activity in order to adapt and update his profile for a better learning.

Self-assessment process allows learning systems to interact with the learner. This process takes advantage of the existence of the created ontology (OntoApp) to evaluate learners during their learning sequences (Fig. 10).

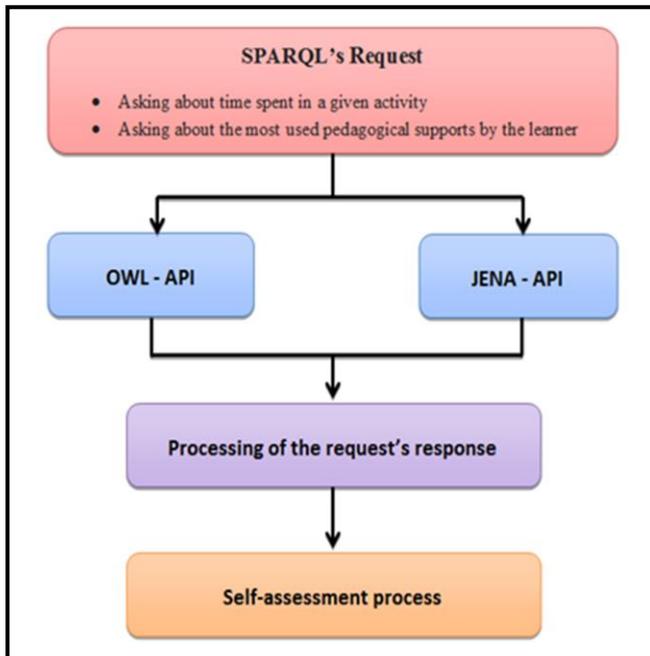


Fig. 10. Stakeholders in the self-assessment process.

The self-assessment process is based on the detection of time spent in an activity and the type of course material most

used by the learner (Fig. 10). It uses the OWL API, the JENA API, and the SPARQL query engine to query the RDF/XML data warehouse in coordination with the OWL (Ontology file) to retrieve the data it needs to self-Learning during his learning sequence and performing the pre-programmed measurements to adjust the learning to the new learner profile.

VI. CONCLUSION

The main objective of this paper is the modeling of the learner and the adjustment of the learning process taking into consideration the learner's profile. This paper tries to go beyond the classical methods of knowledge modeling. Our contributions focus on the following elements: behavioral analysis and evaluation, the detection of learning styles, the development of the learner's profile that takes into account the knowledge, preferences and attitude of the learner. Finally, the paper ends by the realization of an adaptive learning system that allows the adaptation of the pedagogical content according to the current needs of the learner while self-evaluating the learner during the learning sequence thanks to the use of ontologies.

We are aware that our work could be completed and evolved. The next steps are decisive because they consist of confronting the evolving needs of the learners.

We believe that this ontology can detect more parameters, in particular with the contribution of techniques and tools of the semantic Web to better design a CEHL with a large number of satisfied learners.

REFERENCES

- [1] FARIDA, DAHMANI. Modélisation basée ontologies pour l'apprentissage interactif-Application à l'évaluation des connaissances de l'apprenant. 2010. Thèse de doctorat. Université Mouloud Maameri de Tizi Ouzou.
- [2] BEHAZ, Amel et DJOUDI, Mahieddine. Approche de Modélisation d'un Apprenant à base d'Ontologie pour un Hypermédia adaptatif Pédagogique. In : CHIA. 2009.
- [3] EL MEZOUARY, Ali, BATTOU, Amal, OHOUD, Mohsine Ben, et al. The Educational Semantic Web and Associated Technologies for Adaptability in Adaptive Learning Systems. International Journal of Computer Applications, 2011, vol. 32, no 9.
- [4] FERNÁNDEZ-LÓPEZ, Mariano, GÓMEZ-PÉREZ, Asunción, et JURISTO, Natalia. Methontology: from ontological art towards ontological engineering. 1997.
- [5] KORCHI, Adil, EL IDRISSE, Najiba EL AMRANI, Lahcen OUGHDIR, et al. 'A modeling learner approach in a computing environment for human learning based on ontology', International Journal of Computer Engineering & Technology (IJCET) Volume 6, Issue 9, Sep 2015, pp. 21-31, Article ID: IJCET_06_09_003.
- [6] CARMONA, C. et Conejo, R., —A Learner Model in a Distributed Environment, Adaptive Hypermedia and Adaptive Web-Based Systems, Third International Conference, AH 2004, Eindhoven, The Netherlands, Proceedings, 2004, pp. 353-359.
- [7] DE KOCH, Nora Parcus. Software Engineering for Adaptive Hypermedia Systems-Reference Model, Modeling Techniques and Development Process. 2001.

Suitable Personality Traits for Learning Programming Subjects: A Rough-Fuzzy Model

Abdul Rehman Gilal¹, Jafreezal Jaafar², Mazni Omar³, Shuib Basri⁴, Izzatdin Abdul Aziz⁵, Qamar Uddin Khand⁶,
Mohd Hilmi Hasan⁷

^{1,2,4,5,7}Department of Computer and Information Sciences, Universiti Teknologi PETRONAS, Malaysia

^{1,6}Department of Computer Science, Sukkur IBA University, Pakistan

³School of Computing, Universiti Utara Malaysia

Abstract—Programming is a cognitive activity which requires logical reasoning to code for abstract presentation. This study aims to find out the personality traits of students who maintain the effective grades in learning programming courses such as structured programming (SP) and object oriented programming (OOP) by gender classification. Data were collected from three universities to develop, validate, and generalize the Rough-Fuzzy model. Genetic and Johnson algorithms were applied under Rough set theory's (RST) principles to extract the decision rules. In addition, Standard Voting, Naïve Bayesian, and Object Tracking procedures were applied on the generated decision rules to find the prediction accuracy of each algorithm. Mamdani's Fuzzy Inference System (FIS) was used for mapping the decision rules' condition (input) to decision (output) based on fuzzy set theory (FST) to develop the model. The results highlighted that certain personality compositions can be suitable for scoring good grades in programming subjects. For instance, a female student is capable enough to improve the programming skills if she is composed of introvert and sensing personality traits. Therefore, it is important to investigate an appropriate personality composition for programming learners.

Keywords—Software development; personality; programming; rough sets; fuzzy sets

I. INTRODUCTION

Learning to program has always been a hard activity for students. Compare to other courses, programming courses are usually difficult and often have the high dropout rates [1]. Nevertheless, programming skills let students to find a bright future too. In recent years, students' interest for learning programming languages has been increasing rapidly. But, everyone cannot perform well in programming. Shneiderman [2] maintains it that even a similar background of programmers cannot assure the similar performance. In the same vein, Brooks [3] also faced a huge variability in the achievements of introductory programming classes students.

Learning programming acquires cognitive and mental skills to design, code, and debug. Robins et al., [1] mention that writing a program includes various mental models. On the other hand, it is also proposed that mental models are formed by personalities and life experiences [4]. Personality is a complex natural phenomenon and one of the important human factors [5]. For personality, one group of psychology experts declares that personality is an inherited property which does not change but gets improved by time within the same personality type [6]. Whereas, other group of experts say

contrary statement that personality gets changed with time and age between 20 to 40 is stable [7]. Moreover, several questions can take place if personality gets changed or gets betterment into it. For example, which personality types are suitable for learning programming subjects or which personality gets changed into betterment for learning programming? Researchers also believe that many factors influence the type of personality: culture and gender [8]. According to our understanding, personality is also influenced by other personality types. It has certain natural equations which form interpersonal and mental skills. It is highlighted because it was found that certain personality types are flexible to work with each other and some are not [9], [10]. Certain studies have been conducted in the past which proposed several methods and models for finding effective personnel for programming by focusing personality. But, different results have been observed when those models were practically implemented [11]–[13]. Furthermore, several ambiguities have also been found in the literature of personality in software domain. For instance, Gorla and Lam [11] proposed extrovert trait of personality for programmer, whereas Capretz and Ahmed [14] proposed introvert personality trait for the same role.

It is believed that workforce for software industry is always prepared by the education institutes. Keeping all in view, this study was performed on the student population to measure the personality behavior while learning programming subjects. In other words, in order to know that learning programming language is not a random behavior but it has a natural relationship with personality. Therefore, the main objective of the paper was set: "to find out the personality traits which maintain the effective grades in learning programming courses: SP and OOP, by gender classification". Moreover, the future of this research may contribute in following ways:

- 1) It may help the students to select the programming courses (i.e., SP or OOP) as their semester or major course based on their personality type.
- 2) It may also help the subject teachers to design their course and focus particular personality types' students since the beginning of courses for the better outcomes.
- 3) It leaves a new idea for research community who wish to contribute in this area of software development.

The next section of this paper presents the related work for foundation of the study. The section after related work, methodology section, discusses the methods used for data collection and experimenting. The Section IV discusses the results emanated from this study in detail. Additionally, Section V discloses the threats to validity which can be considered for future work. In the end, the paper is concluded in the Section VI.

II. RELATED WORK

Personality has been researched in several fields of science. The following section is organized to show the importance of personality in the software development. This study has also set the gender as a mediating variable between personality and performance. Therefore, Section B (*i.e., Gender and Software Development*) presents the gender in software development. In the last section of related work, programmer role is discussed under the shades of personality.

A. Personality and Software Development

Personality refers to the internal psychological patterns such as feelings and thoughts which curve the behavior of a person. In simple words, personality traits are formed from internal forces. Numerous studies have been carried out in software domain which applied psychological frameworks, widely used in the domain of psychology, to understand the developer personality [15]. These theoretical frameworks include: 1) dispositional, 2) biological, 3) psychoanalytic, 4) neoanalytic, 5) learning, 6) phenomenological, and 7) cognitive self-regulation. Cruz et al., [15] also mentioned that the past research studies have not only used dispositional perspective abundantly so as to determine the personality traits and types in organizational psychology. But, they have also been used commonly in the field of software engineering to determine the most suited personalities to form ideal team for software development. Similarly, this study explores the key importance of personality perspective handy for learning programming courses.

Dispositional perspective of psychology that sheds light on trait and type approach depicts the fact that the personality deals with internal stable qualities that vary from individual to individual and it also influences behavior. American Psychiatric Association defined “trait” as “enduring patterns of perceiving, relating to, and thinking about the environment and oneself that are exhibited in a wide range of social and personal contexts.” Thus, the personalities of the people are determined by their personality patterns classified by psychological differences. Moreover, personality and trait can be distinguished as the former demonstrates different levels and degrees. Whilst, types are discreet, because they cannot be distinguished by levels and degrees [15].

There are some key theories pertinent to personalities that have been profusely implemented in psychological and computing research studies [16]. The most prominent among them are: Keirsey Temperament Sorter [17], Five-Factor Model (FFM) [18], also known as Big Five, and Myers-Briggs Type Indicator (MBTI) [19]. The distinctive point amongst these three personality theories is the way of the describing personality types. Keirsey Temperament Sorter accentuates on

the long term behavior of the individuals [20]. Whereas, Five-Factor Model (FFM) encompasses five distinctive personality traits such as: conscientiousness, agreeableness, openness to experience, extraversion, and agreeableness. On the contrary, MBTI mainly probes into what people think. According to Furham [21] both MBTI and Big Five personality tests are helpful when a researcher aims to examine behavioral and cognitive sides of individuals by correlating both the scales. However, there are many proponents of MBTI in the domain of software engineering as this theory has been widely used in the past research studies [10], [22]–[27]. Thus, keeping in view the wide acceptance of MBTI in terms of its effectiveness, the current study has used this theory.

MBTI primarily focuses on four pairs of the personality which can be further classified into 16 types. The four pairs are: Extroversion-Introversion (I-E), Sensing-Intuitive (S-N), Thinking-Feeling (T-F), and Judging-Perceiving (J-P). These four dimensions also beget sixteen possible combinations of personality types as shown in the following Table 1:

TABLE I. THE 16 MBTI PERSONALITY TYPES

ISTJ (1)	ISFJ (2)	INFJ (3)	INTJ (4)
ISTP (5)	ISFP (6)	INFP (7)	INTP (8)
ESTP (9)	ESFP (10)	ENFP (11)	ENTP (12)
ESTJ (13)	ESFJ (14)	ENFJ (15)	ENTJ (16)

Based on the performance and the score obtained, a person can be attributed with one of the 16 personality types cited in the above Table 1. For instance, a person scoring higher on Introversion (I) than Extroversion, Sensing (S) than Intuition (N); Thinking (T) than Feeling (F) and Judging (J) than Perceiving (P) would be categorized as an ISTJ.

B. Gender and Software Development

In social sciences, many research studies have explored personality and gender, either collectively or separately, to address the grave problems of teamwork in organizations and have achieved the acute success as well. However, this problem is still persistent in the field of software development since few researchers have ever tried to test personality and gender collectively to test the suitability of the team handy for software development. In this regard, Richards and Busch [28], Gilal et al., [10], and Rehman et al., [29] also assert that maturity level is yet to find in software development research. In the same vein, Trauth [30] also recommends that the need of improvement is required in the theoretical work on software development.

Study conducted by Gilal et al., [10] comes among the few studies which focused personality with gender. This study investigated the performance variation among software development team members caused by genders’ personality types. For instance, the male-dominated teams create reasons for females for being ineffective in teams if the personality type of female is with “E” trait. Furthermore, the study also revealed that the female-leader are more convenient with only female or majority-female (*i.e.*, having female in majority) groups. Whereas, male-leaders are acceptable with all kind of team compositions. Critically, this study was just based on

tabulated calculation and could not give any statistical or predictive evidences. However, the study also recommended some future research on gender with personality types to obtain appropriate conclusions. Moreover, Richards and Busch [28] study explored the gender and culture parameters to find their effects on the performance in IT workplaces. This study focused the knowing and doing gaps in software development workplaces. The researchers tried to find the effectiveness of diversity on the overall performance of the team. Moreover, authors acknowledged that these results are too weak to generalize that was one of the limitations of the study. But, these limitations can be overcome by inclusion of personality in the study. Because, it is also believed that inclusion of personality can help to achieve the efficiency, productivity, and quality [31].

C. Software Programmer and Personality

In the software development process, programmer has the key position for implementing the designs of system. The sensitivity of programmer's role lie in a fact that the programmer must be adept in syntax of the programming and good at analytical and logical sharpness for finding the code of the program with an ease. The lack of these qualities could make programmer to face the terrible failure. Because, coding phase has the crucial importance which is used to apply and identify data structures, control structure of the program and determines relevant variables [32]. Moreover, the past research studies have tried to empirically prove the relationship between personality and computer programming activities. Capretz [33] conducted experiments on Brazilian software engineering students to propose a personality profile for software developers. In his study, total 68 students participated and majority of them were male. Moreover, author concluded that ISFP, INTP, and ESTP personality types were significantly overrepresented among Brazilian software engineering students and, whereas, ENTP, ESTJ, and ENTJ personality types were significantly underrepresented among them. In the same vein, Martínez et al., [34] proposed a methodology for assigning roles to software developers. They divided the research experiments into two cases: training and testing with 12 and 16 participants respectively. The findings of the study revealed that ISTP personality type is best fitted for programmer role. Additionally, study conducted by Capretz and Ahmed [35] also highlighted the same objective in which software development tasks were contrasted with personality types. In their study, personality types were mapped with job requirements collected from newspaper, magazines, and online forums. At the end, authors recommended ISTJ and ISTP personality types for programmer role.

III. METHODOLOGY

This study presents the methodology section into two major subsections: Data collection and preprocessing and model development. Data collection and preprocessing section is all about the process of data collection: variables, algorithm, population and criteria of data collection for better understanding of the results of the study. Whereas, the second section highlights the whole process of the model from development to generalization.

A. Data Collection and Preprocessing

In order to achieve the objective of the study, data was collected from three universities: University Teknologi Petronas (UTP), Universiti Utara Malaysia (UUM), and Sukkur Institute of Business Administration (SIBA). Total size of the main dataset was 270, in which 110, 120, and 40 students participated from UTP, UUM and SIBA respectively. In the year of 2015, students who were learning software engineering subject during their bachelor from the universities participated voluntarily in the process of data collection. Software engineering class was chosen to collect data with the reason that in the all three universities, software engineering course is only offered after SP and OOP courses are already learnt by students. It is, because, to maintain the main objective of the study to see that which personality types maintain the effectiveness in learning programming languages in these two subjects. Moreover, in the all three universities, SP and OOP courses were of 4 credit hours per week. In these universities SP is the prerequisite course for OOP. Importantly, content and time duration (i.e., 16 weeks) of the courses were almost same in the all universities because these all three universities offer culture exchange program for international students. Therefore, they have to make a standard course contents and time durations.

MBTI instrument was used to measure the personality types of the participants. It stores the responses of personality in four pairs, as mentioned above, IE pair, SN pair, TF pair, and JP pair. Therefore, this study has 5 independent predictor variables (i.e., gender, IE, SN, TF, and JP) and 1 dependent outcome variable (i.e., improved; where this variable holds the final results whether or not the results of students in SP and OOP are improved). The following Table 2 shows the possible inputs which can be passed to study variables.

TABLE II. CONTROLLING THE INPUTS TO VARIABLE

Variable	Input
Predictor	
1. Gender	1=Male 2=Female
2. IE	1=introvert 2=extrovert
3. SN	1=sensing 2=intuiting
4. TF	1=thinking 2=feeling
5. JP	1=judging 2=perceiving
Outcome	
1. Improved	0= did not improve 1= improved

The calculation of the outcome variable was made from the obtained marks of students in the SP and OOP subjects. For example, if a student obtained grade "B" in SP and grade "A" in OOP then it means the student improved the grades. Similarly, if the student obtained grade "B" in SP and grade "C" in OOP then it represents that student could not manage to improve the results. Another possibility could also occur that student neither "improved" nor "did not improve". In that situation, the input was adjusted in "improved" if the grades

are still greater than “B” grade otherwise considered as “did not improve”. It was applied, because, all these universities consider that grade “B” or Grade Point Average (GPA) 3.0 and above are good to excellent academic levels. Appendix 1 is highlighting the chart of grading scales in the universities. The following algorithm script defines the process of assigning values to outcome variable.

```
for (i=1 to i<=k)// k is the total number of participants in the experiment
if (Grade_OOP(i) > Grade_SP(i)) // Grade_OOP contains the results of OOP subject of all
students and Grade_SP contains SP results
    improved(i)=1 // it means "improved"
else if (Grade_OOP(i) < Grade_SP(i))
    improved(i)=0 // it means "did not improve"
else // otherwise the Grade_OOP(i) = Grade_SP
    if (Grade_OOP(i) >= "B")
        improved(i)=1 // it is considered in "improved"
    else
        improved(i)=0 // otherwise considered in "did not improve"
```

Fig. 1. Determining the values for outcome variable.

B. Model Development

To develop the model, the combination of RST and FST was used together to build an efficient model [36]. In this study, RST is used to extract the rules (IF-THEN) and, which were, then used as rules' database to fuzzy controller with Mamdani inference. The details from rules extraction to implementation to Mamdani inference are presented in the next sections. Basically, the model development phase was based on several steps:

a) Rules Generation

Model development was started with extracting the useful rules for decisions. In this study, the rules generation and validation were performed by using ROSETTA toolkit: analyzing toolkit for tabular data within RST framework [37]. This step of modeling also helped to remove the redundancy and unimportant data through reduction process. Moreover, Genetic Algorithm (GA) and Johnson's Algorithm (JA) were applied on data. Because, Hvidsten [38] says that GA is one of the effective solutions to searching problems. On another hand, Johnson [39] stated that JA invokes a variation of a simple greedy algorithm to compute a single reduct only. Therefore, both algorithms were applied to use the most effective one's rules for decision. The following experiment objectives were set for this step:

1) Find the personality traits of students who obtained effective (good) or low grades in SP and OOP courses (where term effective (good) refers to grade B and above or GPA 3.0 and above).

2) Find the personality traits of students who managed to improve grades in learning OOP (or managed to be consistent in the effective grades while learning SP and OOP).

It is important to note that these experiments were performed on each dataset to extract the effective rules separately. But, the rules extracted from UTP dataset were compared with UUM and SIBA datasets for results generalization purpose. It means that UTP dataset was used for model development and data from UUM and SIBA were used for validation and generalization.

b) Rules Evaluations and Generalization

The measurement of effectiveness of rules was computed through hold-out methods. For that purpose, this study used 70% of data for training and remaining 30% for testing the prediction accuracy only on UTP dataset (as mentioned above). It means that selection of algorithm results was based on the prediction accuracy. Hence, higher the prediction accuracy will increase the efficiency of model. Standard Voting, Naïve Bayesian, and Object Tracking procedures were applied on generated rules to find the prediction accuracy of each algorithm.

Moreover, for further validation and generalization, datasets from UUM and SIBA were equally distributed as the size of 30% of testing dataset of UTP. Basically, UUM data was used to see whether or not the personality preferences remain the same within Malaysian university students (i.e., UTP and UUM, during learning SP and OOP). Similarly, it was also validated to see the personality preferences behavior with Pakistani students. Basically, the performances of the model were measured in two ways: prediction accuracy and based on ROC, Area under Curve (AUC) results. Hence, 70% was considered as a benchmark for the effective prediction accuracy. Because, according to Bakar [40], the predication results can be known effective if the prediction accuracy is at least 70%. Similarly, Hvidsten [38] also mentioned that the 70% prediction accuracy is acceptable for prediction modeling. Moreover, Fawcett [41] asserted that the model is perfect if the obtained AUC is 1. On another hand, the model can be accepted if the computed AUC is 0.5 or above otherwise the generalization is rejected (if $AUC < 0.5$). Therefore, the model results can be generalized if the prediction accuracy is at least 70% and AUC curve was computed 0.5 or above.

c) Fuzzy Inference System (FIS) Development

Once the rules were extracted and validated, the FIS; a system that maps the input to output based on fuzzy set theory, was used with Mamdani [42] inference system. The selection of Mamdani, instead of Sugeno inference system, was basically because of defuzzification process. In the same vein, Govindarajan [43] also states that Mamdani is mostly used in pure fuzzy systems. Moreover, the Mamdani FIS system was developed by following its four basic parts [44] by using Matlab:

1) *Fuzzifier*: With the help of membership function, it helped to convert the crisp inputs into fuzzy inputs. Linear triangular, one of the mostly used membership functions [44], was used to define membership functions.

2) *Rules*: IF-THEN statements that were already defined from RST experiments.

3) *Interface Engine*: A part which converted the fuzzy input sets to fuzzy output from defined rules database.

4) *Defuzzifier*: With the help of membership function, the fuzzy outputs were converted to crisp output. Centroid or Center of area method (COA), a popular approach, was used to perform defuzzification.

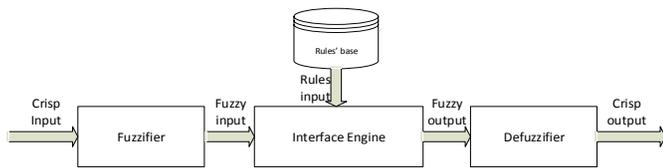


Fig. 2. Mamdani inference system.

Fig. 2 represents the general form of Mamdani inference system which independently acts for fuzzy controller development. On another hand, Fig. 3 is showing the overall process of rough and fuzzy approaches integration within the model development.

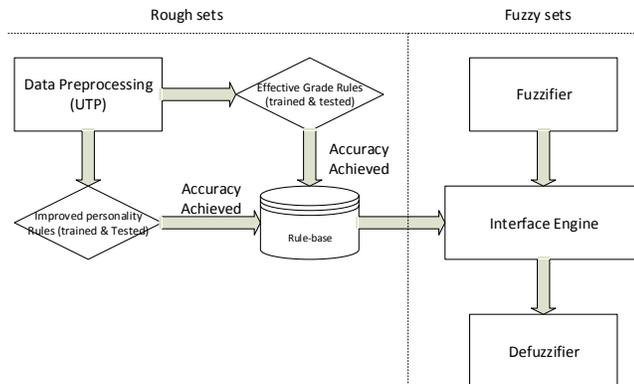


Fig. 3. Integration of rough and fuzzy approaches.

IV. RESULTS AND DISCUSSION

The first step was to decide which algorithm (i.e., GA or JA) results were effective for model development. The obtained accuracy of rules was set as a benchmark (i.e., 70%) for algorithm selection. Table 3 summarizes the overall results of experiments on the both algorithms.

TABLE III. EXTRACTED REDUCTS WITH OBTAINED ACCURACY

	Objective:1 (good grades)		Objective: 2 (improved grades?)	
	GA	JA	GA	JA
Redcuts	16	14	18	15
Standard Voting	71.25%	74.24%	69.50%	72.72%
Object Tracking	70.59%	73.43%	67.76%	71.80%
Naïve Bayesian	65.34%	69.10%	68.54%	70.23%

JA algorithm was found suitable in both objectives. For example, in both cases, JA algorithm produced less reducts, than GA, which create a lesser complexity in the model. Moreover, JA algorithm maintained the effective accuracy in the all mentioned classifying techniques. Only in Naïve Bayesian, JA algorithm reducts could not obtained the said benchmark accuracy. But, overall, JA got the acceptable accuracy in both objectives. Therefore, this model used reducts of JA algorithm for finalizing the rule-base for fuzzy controller.

A. Redcuts for Finding Personality Traits who Obtained Good and Low Grades in Learning SP and OOP

In order to maintain the main theme of this paper, it was first objective to see that what personality traits could earn good and low grades while learning SP and OOP. Each participant's results of SP and OOP programming were collected during experiment. In the first objective, gender, personality traits, and the results (of SP and OOP) were used to form the reducts without caring subject specification. Because, if a student managed to earn good grade in any subject it means that particular personality has capability to earn good grades. Therefore, it was even more precise discovery within dataset about effective personality traits. Table 4 shows the extracted reducts from JA algorithm on the experiments for finding good and low grades' personality traits.

TABLE IV. REDUCTS OBTAINED FROM JA ALGORITHM FOR THE FIRST OBJECTIVE

No	Redcuts	LHS Support	RHS Support	RHS Accuracy	LHS Coverage	RHS Coverage
1	Female AND Extrovert AND Judging => Good-grade OR low-grade	45	34, 11	0.7555 56, 0.2444 44	0.29 2208	0.2905 98, 0.2972 97
2	Female AND Introvert AND Sensing => Good-grade OR low-grade	30	22, 8	0.7333 33, 0.2666 67	0.19 4805	0.1880 34, 0.2162 16
3	Female AND Thinking => Good-grade OR low-grade	34	26, 8	0.7647 06, 0.2352 94	0.22 0779	0.2222 22, 0.2162 16
4	Extrovert AND iNtuiting AND Judging => Good-grade OR low-grade	56	43, 13	0.7678 57, 0.2321 43	0.36 3636	0.3675 21, 0.3513 51
5	Extrovert AND Feeling AND Judging => Good-grade OR low-grade	52	38, 14	0.7307 69, 0.2692 31	0.33 7662	0.3247 86, 0.3783 78
6	Male AND Extrovert AND Sensing AND Thinking AND Judging => Good-grade	5	5	1	0.03 2468	0.0427 35
7	iNtuiting AND Thinking => Good-grade OR low-grade	32	23, 9	0.7187 5, 0.2812 5	0.20 7792	0.1965 81, 0.2432 43
8	Male AND Introvert AND Feeling => Good-grade	15	15	1	0.09 7403	0.1282 05
9	Thinking AND Perceiving => Good-grade OR low-grade	4	3, 1	0.75, 0.25	0.02 5974	0.0256 41, 0.0270 27
10	Introvert AND Thinking => Good-grade OR low-grade	27	22, 5	0.8148 15, 0.1851 85	0.17 5325	0.1880 34, 0.1351 35
11	Female AND Extrovert AND Sensing AND Perceiving => low-grade	2	2	1	0.01 2987	0.0540 54
12	Male AND Feeling AND Perceiving =>	10	10	1	0.06 4935	0.0854 7

	Good-grade					
13	iNtuiting AND Perceiving => Good-grade	8	8	1	0.051948	0.068376
14	Introvert AND iNtuiting AND Feeling => Good-grade	12	12	1	0.077922	0.102564

In Table 4, the term *reducts* (also called *rules*) refers to the extracted information of condition where “IF” a situation occurs “THEN” what decision should be taken. “Left Hand Side (LHS) support” informs that how many objects belong to “IF condition” and, whereas, “Right Hand Side (RHS) support” shows the number of decision objects on the “IF condition” within whole training dataset. Moreover, in this objective, the decision variable contains only two values: “Good-grade” and “low-grade”, hence the RHS support has sometimes returned two numbers if the decision is divided. The rules which have two decision are called *bi-dimension* (i.e., rule no. 1). Otherwise, rules are called *uni-dimension* if it has only one decision (i.e., rule no. 6). Similarly, RHS accuracy is obtained to highlight the weight of decision within rule. It is simply obtained by dividing LHS support with the one value from RHS support. The accuracy of decision is always 1 if the rule has only one decision. In the same vein, LHS coverage shows the overall appearance of “IF condition” within the training dataset. Whereas, RHS coverage is almost similar to LHS coverage but it is only applied on decision part of *reducts* upon the class listed in the then part. For instance, the appearance of “low-grade” in the training set was 37 and, hence, the RHS coverage of rule no 1 for “low-grade” class is 0.297297.

First of all, total eight rules (i.e., Table 4 rule no. 1-5, 7, 9, and 10) out of fourteen were *bi-dimension* and, whereas, only remaining six rules were *uni-dimension*. Overall, most of *bi-dimension* rules seemed more towards “good-grade” if simply rely on the RHS support computation. Nevertheless, in this case, the high coverage of RHS support was not sufficient to decide the impact of these *bi-dimension* rules on one side of class when the dataset is not normally distributed. Therefore, the RHS coverage computation was necessary to consider defining the impacting class. More precisely, rule no. 1, 2, 5, and 7 were classified into *low-grade* class because their coverage was higher than *good-grade* class. On another hand, rule no. 3, 4, 9 and 10 were found more to *good-grades* class with the same reason. But, it should also be noted that rule no. 1, 3, and 4 did not show the higher difference between good and low grades. The following figures show the coverage of IF and THEN parts within graphs:

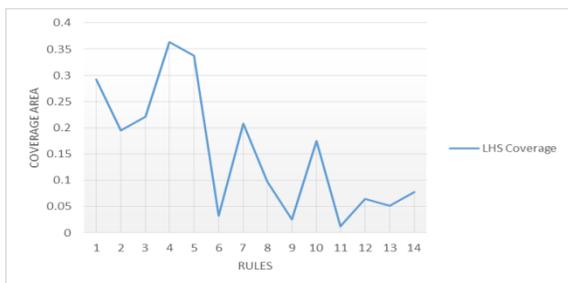


Fig. 4. “if-statement” coverage within training set for the first objective



Fig. 5. “Then” part coverage grades within training set for the first objective

Moreover, total seven rules (i.e., Table 4 rule no. 1-3, 6, 8, 11, and 12) were found on gender classification. It shows that 50% of rules were gender-based decisions. More specifically, four rules: 1, 2, 3, and 11, were listed for female decision and in which rule 1, 2, and 3 were *bi-dimension* and rule number 11 was *uni-dimension*. Based on RHS coverage computations, rule 1 and 2 were listed for *low grade* decision. Therefore, it was extracted from rules that if a female learner is E with J personality traits or I with S or E with S and P then she may get *low-grades* in learning SP and OOP. On another, thinking (T) females were found effective in earning *good-grades*. This was also found in our previous study [25], [45] that T-trait females are effective for programming jobs. Furthermore, combinations of ESTJ personality traits were found progressive for male learners. The ESTJ personality type is also appeared in the past studies [46] but the gender is missing. Hence, results from this study highlight that ESTJ can be progressive for male learners. In the same vein, I with F and F with P personality traits combinations were also found effective for male learners in SP and OOP classes. Lastly, N with P and I with N and F (rule no. 13 and 14) personality combinations were suitable for both genders for learning SP and OOP.

B. Reducts for Finding Consistent Effective Personality Traits in Learning SP and OOP

The section above underlined the personality traits of students that achieved good and low grades in learning SP and OOP subjects. On the other hand, this section was prepared to highlight those personality traits which maintained the good or low grades in the both subjects. For example, it does not happen always that the personality combination which managed the good grades in SP will manage the good grades in OOP too or other way around. Therefore, in order to see that whether the behavior of personality traits appeared identical in these both subjects or it has variations upon the subjects need. Table 5 summarizes the *reducts* extracted from the experiments for finding the answer on personality consistency.

The structure of the table is totally same as Table 4 but the description of decision class is different with “improve” and “didn’t improve” outputs. Where “improve” denotes that personality combination appeared in IF statement (or LHS) of the rule was found effective or improved in learning SP and OOP. Whereas, “didn’t improve” classify that the personality combination in the IF statement did not manage to improve the results in OOP.

TABLE V. REDUCTS OR FINDING CONSISTENCY OF PERSONALITY TRAITS IN LEARNING SP AND OOP

No	Reducts	LH S Support	RH S Support	RHS Accuracy	LHS Coverage	RHS Coverage
1	iNtuiting AND Judging => Improved OR didn't-improve	37	17, 20	0.4594 59, 0.5405 41	0.48 0519	0.3953 49, 0.5882 35
2	iNtuiting AND perceiving => Improved	3	3	1	0.03 8961	0.0697 67
3	Female AND Feeling AND Judging => Improved OR didn't-improve	22	12, 10	0.5454 55, 0.4545 45	0.28 5714	0.2790 7, 0.2941 18
4	Extrovert AND iNtuiting => didn't-improve OR Improved	28	16, 12	0.5714 29, 0.4285 71	0.36 3636	0.4705 88, 0.2790 7
5	iNtuiting AND Thinking => didn't-improve OR Improved	18	11, 7	0.6111 11, 0.3888 89	0.23 3766	0.3235 29, 0.1627 91
6	Female AND Introvert AND Sensing => Improved OR didn't-improve	13	8, 5	0.6153 85, 0.3846 15	0.16 8831	0.1860 47, 0.1470 59
7	Male AND Extrovert AND Thinking => didn't-improve OR Improved	10	5, 5	0.5, 0.5	0.12 987	0.1470 59, 0.1162 79
8	Female AND Extrovert AND Sensing AND Thinking => Improved	2	2	1	0.02 5974	0.0465 12
9	Male AND Introvert AND Sensing AND Thinking => didn't-improve	1	1	1	0.01 2987	0.0294 12
10	Male AND Introvert AND Sensing AND Feeling => Improved	4	4	1	0.05 1948	0.0930 23
11	Thinking AND perceiving => Improved OR didn't-improve	2	1, 1	0.5, 0.5	0.02 5974	0.0232 56, 0.0294 12
12	Female AND Extrovert AND perceiving => Improved	1	1	1	0.01 2987	0.0232 56
13	Male AND Extrovert AND Judging => didn't-improve OR Improved	17	8, 9	0.4705 88, 0.5294 12	0.22 0779	0.2352 94, 0.2093 02
14	Male AND Introvert AND perceiving => Improved	4	4	1	0.05 1948	0.0930 23
15	Male AND Extrovert AND Feeling AND perceiving => didn't-improve	1	1	1	0.01 2987	0.0294 12

Table 5 comprised total eight bi-dimension rules (i.e., 1, 3-7, 11, and 13) and seven rules (i.e., 2, 8, 9, 10, 12, 14, and 15) as uni-dimension. From those bi-dimension rules, number 1, 4, 5 and 7 were classified to “didn’t improve” class based on RHS support and coverage. It was found in these classified rules that combination of N trait with E or F or J (i.e., Table 5 rule no. 1, 4, and 5) trait does not guarantee the improvement

in learning OOP. Similarly, male with E and T personality traits appeared inconsistent in the training set. On the other hand, rule no. 3, 6, and 13 were computed for “improved” class. Moreover, rule no 11 remained uncertain at this stage because it was computed almost same for both decision classes. But, it was considered in “didn’t improve” class as it had very little higher coverage than “improved”. The following Fig. 6 displays the RHS coverage against the rules.



Fig. 6. RHS coverage for second objective from training set.

In this objective, gender was appeared highly impacting variable on the extracted rules. Total 10 rules (i.e., 3, 6-10, 12-15) were listed in gender classification; in which 4 rules (i.e., 3, 6, 8, and 12) belonged to female and remaining 6 (i.e., 7, 9, 10, 11, 13, 14, and 15) for male learners. Moreover, in the female rules, number 3 and 6 were bi-dimension and classified to “improved class”. Whereas, two more rules: 8 and 12 were uni-dimension with “improved” class. Based on these rules, one can say that female gender can produce consistently effective results in learning SP and OOP if she is composed of feeling (F) and judging (J) or introvert (I) and sensing (S) or extrovert (E) and sensing (S) or extrovert (E) and perceiving (P). On another hand, number 7, 9 and 15 rules were straightly classified to “didn’t improve” class in the male rules. In which, male with I, S, and T traits or E, F, and P traits or E with T traits’ composition were found inconsistent in learning SP and OOP subjects. Moreover, a male was found consistent or improved when I with S and F traits or I with P traits or E with J traits (i.e., Table 5 rule no 10, 13 and 14). Furthermore, in the past studies [46], [47], E or S or J traits are found effective for programming job. But, to what extent they are consistent and for which gender they are more suitable or what other traits should be aligned for better results. Therefore, these results can bring some interesting and new information for users.

C. Comparison between both Objectives

In the previous sections, personality traits were extracted either based on low and good grades or consistency in improving grades in learning SP or OOP subjects. The first objective helped to extract the personality traits that can produce effective or ineffective results in learning SP and OOP. Whereas, second objective extracted the personality traits which were consistent to achieve improvement in learning SP and OOP or other way around. However, this section is designed to compare the both objectives and to see that whether the obtaining effective results and consistency are with reasons or it is random. Table 6 contains the rules of both objectives after deciding bi-dimensional rules in its suitable classes.

TABLE VI. COMPARISON BETWEEN 1ST AND 2ND OBJECTIVES

	1st Objective (personality traits based on low and good grades)	Classified in	2nd objective (personality traits based on consistency in results)	Classified in
1	Female AND Extrovert AND Judging	low-grade	iNtuiting AND Judging	didn't improve
2	Female AND Introvert AND Sensing	low-grade	iNtuiting AND perceiving	Improved
3	Female AND Thinking	good-grade	Female AND Feeling AND Judging	Improved
4	Extrovert AND iNtuiting AND Judging	good-grade	Extrovert AND iNtuiting	didn't improve
5	Extrovert AND Feeling AND Judging	low-grade	iNtuiting AND Thinking	didn't improve
6	Male AND Extrovert AND Sensing AND Thinking AND Judging	good-grade	Female AND Introvert AND Sensing	Improved
7	iNtuiting AND Thinking	low-grade	Male AND Extrovert AND Thinking	didn't improve
8	Male AND Introvert AND Feeling	good-grade	Female AND Extrovert AND Sensing AND Thinking	Improved
9	Thinking AND Perceiving	good-grade	Male AND Introvert AND Sensing AND Thinking	didn't improve
10	Introvert AND Thinking	good-grade	Male AND Introvert AND Sensing AND Feeling	Improved
11	Female AND Extrovert AND Sensing AND Perceiving	low-grade	Thinking AND perceiving	didn't improve
12	Male AND Feeling AND Perceiving	good-grade	Female AND Extrovert AND perceiving	Improved
13	iNtuiting AND Perceiving	good-grade	Male AND Extrovert AND Judging	Improved
14	Introvert AND iNtuiting AND Feeling	good-grade	Male AND Introvert AND perceiving	Improved
15			Male AND Extrovert AND Feeling AND perceiving	didn't improve

It was already mentioned above that total fourteen (14) rules were obtained from first objective and fifteen from second objective. In the first objective, five (5) rules were classified into “low grade” and nine (9) into “good grade” class. Whereas, on another side, seven (7) rules were classified into “didn’t improve” and eight (8) into “improved” class in the second objective. Moreover, based on the results, it was found that in the first objective rule no. 2 and 11 (i.e., Table 6, “Female AND Introvert AND Sensing” and “Female AND Extrovert AND Sensing AND Perceiving”) were listed in “low

grade” class. But, both rules were appeared in the “improved” class in the second objective (see Table 6 rule no. 6 and 12 in the column of second objective). Similarly, rule no. 7 (iNtuitive AND Thinking) in the first objective was classified into “low grade” class and found into “didn’t improve” class in the second objective (rule no. 5). Whereas, rule no. 9 was in “good grade” class in the first objective but it was found in “didn’t improve” class in the second objective (i.e., rule no. 11). In the same way, rule no. 3 and 13 in the first objective (“Female AND Thinking” and “iNtuiting AND Perceiving”) were considered in the “good grade” class and found in the “improved” class of second objective (i.e., rule no. 8 and 2).

Finally, from three “good grade” personalities (i.e., rule no. 3, 9 and 13 in 1st objective), two had capability to improve (i.e., rule no. 2 and 8 in 2nd objective) and one did not improve (i.e., rule no. 11 in 2nd objective). On another hand, two “low grade” personality (i.e., rule no. 2 and 11 in 1st objective) combination were found improved (i.e., rule no. 6 and 12 in 2nd objective) and one could not improve (i.e., rule no. 7 in 1st objective and 5 in 2nd objective) in learning. Therefore, based on the personality pairs found common in the both objectives, it could be summarized that there are some personality traits which can grade good in programming subjects and remain consistent or other way around.

D. Rules Generalization

UTP dataset was used to extract rules for making rule-base for FIS system. Rules extracted from it fulfilled the demand of efficiency benchmark [38], [40]. But, to what extent these results can be utilized for finding effective personality traits for programming learners. Data from UUM and Sukkur IBA were used to find the generalization within Malaysia and out of it. Table 7 summarizes the results extracted from these datasets.

TABLE VII. GENERALIZATION OF RESULTS BASED ON UUM AND SUKKUR IBA DATASETS

Objective 1					
	sub-sets	Standard Voting	Object Tracking	Naïve Bayesian	ROC
UUM	uum1_ob1	69%	64%	65%	0.48
	uum2_ob1	73%	71%	69%	0.54
	uum3_ob1	70%	74%	71%	0.51
	uum4_ob1	76%	72%	71%	0.59
Sukkur IBA	iba1_ob1	54%	55%	51%	0.35
Objective 2					
	sub-sets	Standard Voting	Object Tracking	Naïve Bayesian	ROC
UUM	uum1_ob2	81%	76%	72%	0.65
	uum2_ob2	77%	79%	73%	0.63
Sukkur IBA	iba1_ob2	51%	55%	47%	0.41

It was clearly found that the results extracted from Malaysia cannot be generalized with Pakistan. In the first objective, one subset (i.e., uum_ob1) from UUM dataset was

appeared slightly lower than the benchmark of accuracy but it showed the accepted benchmark on ROC curve. But, generally the rules accuracy was above 70%. Hence, it can be inferred that these results can be used within Malaysian universities. Whereas, for further expansion in the model, rules can be extracted from other countries data.

E. Fuzzy Inference System (FIS) Development

It is already mentioned in the methodology section that Mamdani inference system was used for fuzzification and defuzzification. For fuzzification process, the input variables were simply used with defined ranges in their membership functions. But, for output variable, the “low-grade” and “didn’t improve” classes were merged into “ineffective” membership function and, similarly, “good-grade” and “improved” classes were merged into “effective” membership function. Additionally, the output variable was ranged from 0 to 1 in which it was considered “ineffective” if the computed range is less than or equal to 0.5 and it was set “effective” if the range is greater than 0.5. Fig. 7 displays the control on output variable. During rule defining process, two rules (“iNtuitive AND Thinking” and “iNtuitive AND Perceiving”) were appeared twice with the same computation results because they were listed in the both objectives. Hence, they were kept once in the rule-base with double weight.

Moreover, in order to ensure the performance, the controller was used within Simulink. An array of values was passed to it by using “from workspace” block and the response of all returns was also captured in the workspace for further verifications. Therefore, the datasets used in training and testing were supplied from workspace to the controller without passing decision or outcome variable. It is because, at this stage, the accuracy of the controller was being measured rather than the performance of the model. However, it was expected to achieve the performance, at least, the same like obtained accuracy (i.e., at least 70%) as the datasets were same. Table 8 presents the confusion matrix based on the real dataset outcomes and obtained from controller.

TABLE VIII. CONFUSION MATRIX FOR CONTROLLER PERFORMANCE

	Predicted ineffective	Predicted effective	
Actual ineffective	29	12	41
Actual effective	13	56	69
	42	68	0.7273

Finally, as mentioned above, the obtained prediction accuracy was much similar as it was obtained in the classification. Moreover, the sensitivity of the predicted results was 0.81 and, whereas, computed specificity was 0.71. Therefore, based on the results obtained from these experiments, the developed model can be considered satisfied for future use.

V. THREATS TO VALIDITY

Personality is a complex part of human factors which can be vague in shapes. It can be impacted from several internal and external factors: culture and language. Thus, the results from this study cannot be generalized other than Malaysian

universities. In order to generalize it, the model can be expanded with multicultural data for more rules. Moreover, only two subjects (i.e., SP and OOP) learning was measured to develop the model. It restricts the results of the model for other programming and development subjects. The model can be enriched if it is trend with several other subjects: Database languages or web development languages. Furthermore, only MBTI based personality compositions are offered in the model. It can be one of the limitations of the model. Therefore, this model can include the new rules based on personality assessment other than MBTI: Big Five or Keirsey Temperament Sorter.

VI. CONCLUSION

Some personality combinations have capability to improve their programming learning skills. Other way around, some personality compositions are weak in learning and improving the programming skills. It was also found that in some cases each gender (i.e., male or female) demands different composition. For example, compositions “Female AND Introvert AND Sensing” or “Female AND Extrovert AND Sensing AND Perceiving” have capabilities to improve the programming skills. This study concludes that learning programming subjects has direct relation with certain personality compositions. Hence, it is very much important to investigate an appropriate personality composition for programming learners. Moreover, personality is complex in nature, extensively hidden results can be produced if complex networks approaches are applied. We aim to extend this study with complex networks approaches. Currently, weighted degree centrality, betweenness centrality, and closeness centrality techniques are proposed for future work.

REFERENCES

- [1] Robins, J. Rountree, and N. Rountree, “Learning and teaching programming: A review and discussion,” *Comput. Sci. Educ.*, vol. 13, no. 2, pp. 137–172, 2003.
- [2] B. Shneiderman, “Software psychology: Human factors in computer and information systems (Winthrop computer systems series),” 1980.
- [3] R. E. Brooks, “Studying programmer behavior experimentally: the problems of proper methodology,” *Commun. ACM*, vol. 23, no. 4, pp. 207–213, 1980.
- [4] N. Halevy, T. R. Cohen, E. Y. Chou, J. J. Katz, and A. T. Panter, “Mental Models at Work Cognitive Causes and Consequences of Conflict in Organizations,” *Personal. Soc. Psychol. Bull.*, p. 146167213506468, 2013.
- [5] F. P. Brooks Jr, *The Mythical Man-Month, Anniversary Edition: Essays on Software Engineering*. Pearson Education, 1995.
- [6] C. J. Boyce, A. M. Wood, and N. Powdthavee, “Is personality fixed? Personality changes as much as ‘variable’ economic factors and more strongly predicts changes to life satisfaction,” *Soc. Indic. Res.*, vol. 111, no. 1, pp. 287–305, 2013.
- [7] P. T. Costa Jr and R. R. McCrae, “Personality stability and its implications for clinical psychology,” *Clin. Psychol. Rev.*, vol. 6, no. 5, pp. 407–423, 1986.
- [8] J. S. Karn and a J. Cowling, “Using ethnographic methods to carry out human factors research in software engineering,” *Behav. Res. Methods*, vol. 38, no. 3, pp. 495–503, Aug. 2006.
- [9] R. Feldt, L. Angelis, R. Torkar, and M. Samuelsson, “Links between the personalities, views and attitudes of software engineers,” *Inf. Softw. Technol.*, vol. 52, no. 6, pp. 611–624, 2010.
- [10] A. R. Gilal, J. Jaafar, M. Omar, and M. Z. Tunio, “Impact of Personality and Gender Diversity on Software Development Teams’ Performance,”

- in International Conference on Computer, Communication, and Control Technology (I4CT 2014), 2014, no. 2014 IEEE 2014, pp. 261–265.
- [11] N. Gorla and Y. W. Lam, “Who should work with whom?: building effective software project teams,” *Commun. ACM*, vol. 47, no. 6, pp. 79–82, 2004.
- [12] A. R. Gilal, J. Jaafar, S. Basri, M. Omar, and A. Abro, “Impact of software team composition methodology on the personality preferences of Malaysian students,” in 2016 3rd International Conference on Computer and Information Sciences (ICCOINS), 2016, pp. 454–458.
- [13] A. R. Gilal, M. Omar, J. Jaafar, K. I. Sharif, A. W. Mahesar, and S. Basri, “Software Development Team Composition: Personality Types of Programmer and Complex Networks,” in 6th International Conference on Computing and Informatics (ICOI-2017), 2017, pp. 153–159.
- [14] L. F. Capretz and F. Ahmed, “Making sense of software development and personality types,” *IT Prof.*, vol. 12, no. 1, pp. 6–13, 2010.
- [15] S. Cruz, F. Q. B. da Silva, and L. F. Capretz, “Forty years of research on personality in software engineering: A mapping study,” *Comput. Human Behav.*, vol. 46, pp. 94–113, 2015.
- [16] M. Omar, N. Katuk, S. L. S. Abdullah, N. L. Hashim, and R. Romli, “ASSESSING PERSONALITY TYPES PREFERENCES AMONGST SOFTWARE DEVELOPERS: A CASE OF MALAYSIA,” *ARNP J. Eng. Appl. Sci.*, vol. VOL. 10, N, no. FEBRUARY 2015, pp. 1499–1504, 2015.
- [17] M. Bates and D. Keirse, “Please Understand Me: Character and Temperament Types,” *Del Mar Prometh. Nemesis B. Co*, 1984.
- [18] R. R. McCrae and O. P. John, “An introduction to the five-factor model and its applications,” *J. Pers.*, vol. 60, no. 2, pp. 175–215, 1992.
- [19] I. B. Myers, M. H. McCauley, N. L. Quenk, and A. L. Hammer, *MBTI manual: A guide to the development and use of the Myers-Briggs Type Indicator*, vol. 3. Consulting Psychologists Press Palo Alto, CA, 1998.
- [20] L. J. Francis, C. L. Craig, and M. Robbins, “The relationship between the Keirsey Temperament Sorter and the short-form revised Eysenck Personality Questionnaire,” *J. Individ. Differ.*, vol. 29, no. 2, pp. 116–120, 2008.
- [21] A. Furnham, “The big five versus the big four: the relationship between the Myers-Briggs Type Indicator (MBTI) and NEO-PI five factor model of personality,” *Pers. Individ. Dif.*, vol. 21, no. 2, pp. 303–307, 1996.
- [22] A. D. Da Cunha and D. Greathead, “Does personality matter?: an analysis of code-review ability,” *Commun. ACM*, vol. 50, no. 5, pp. 109–112, 2007.
- [23] A. R. Gilal, J. Jaafar, M. Omar, S. Basri, and A. Waqas, “A Rule-Based Model for Software Development Team Composition: Team Leader Role with Personality Types and Gender Classification,” *Inf. Softw. Technol.*, vol. 74, pp. 105–113, 2016.
- [24] J. S. Karn, S. Syed-Abdullah, A. J. Cowling, and M. Holcombe, “A study into the effects of personality type and methodology on cohesion in software engineering teams,” *Behav. Inf. Technol.*, vol. 26, no. 2, pp. 99–111, 2007.
- [25] A. R. Gilal, J. Jaafar, M. Omar, S. Basri, A. Aziz, and I. Din, “A Set of Rules for Constructing Gender-based Personality types ’ Composition for Software Programmer,” *Lect. Notes Electr. Eng. by Springer*, 2015.
- [26] M. Omar and S.-L. Syed-Abdullah, “Identifying Effective Software Engineering (SE) Team Personality Types Composition using Rough Set Approach,” in *IEEE*, 2010, pp. 1499–1503.
- [27] A. R. Gilal, M. Omar, and K. I. Sharif, “A Rule-Based Approach For Discovering Effective Software Team Composition,” *J. Inf. Commun. Technol.*, vol. 13, pp. 1–20, 2014.
- [28] D. Richards and P. Busch, “Knowing-doing gaps in ICT: gender and culture,” *J. Inf. Knowl. Manag. Syst.*, vol. 43, no. 3, pp. 264–295, 2013.
- [29] M. Rehman, A. K. Mahmood, R. Salleh, and A. Amin, “Mapping job requirements of software engineers to Big Five Personality Traits,” in *Computer & Information Science (ICIS)*, 2012 International Conference on, 2012, vol. 2, pp. 1115–1122.
- [30] E. M. Trauth, “Theorizing gender and information technology research,” *Enycl. Gen. Inf. Technol.*, vol. 2, pp. 1154–1159, 2006.
- [31] D. Varona and L. F. Capretz, “Evolution of Software Engineers ’ Personality Profile,” vol. 37, no. 1, pp. 2–6, 2012.
- [32] S. Licorish, A. Philpott, and S. G. MacDonell, “Supporting agile team composition: A prototype tool for identifying personality (in) compatibilities,” in *Proceedings of the 2009 ICSE Workshop on Cooperative and Human Aspects on Software Engineering*, 2009, pp. 66–73.
- [33] L. F. Capretz, “Psychological Types of Brazilian Software Engineering Students,” *J. Psychol. Type*, no. 5, pp. 37–42, 2008.
- [34] L. G. Martínez, G. Licea, A. Rodríguez-Díaz, and J. R. Castro, “Experiences in Software Engineering Courses Using Psychometrics with RAMSET,” in *Proceedings of the Fifteenth Annual Conference on Innovation and Technology in Computer Science Education*, 2010, pp. 244–248.
- [35] L. F. Capretz and F. Ahmed, “Making Sense of Software Development and Personality Types,” *IT Prof.*, vol. 12, no. February, pp. 6–13, 2010.
- [36] J. Krupka and P. Jirava, “Rough-fuzzy Classifier Modeling Using Data Repository Sets,” *Procedia Comput. Sci.*, vol. 35, pp. 701–709, 2014.
- [37] A. Øhrn, J. Komorowski, and others, “Rosetta—a rough set toolkit for analysis of data,” in *Proc. Third International Joint Conference on Information Sciences*, 1997.
- [38] T. R. en Hvidsten, “Fault diagnosis in rotating machinery using rough set theory and ROSETTA,” 1999.
- [39] D. S. Johnson, “Approximation algorithms for combinatorial problems,” *J. Comput. Syst. Sci.*, vol. 9, no. 3, pp. 256–278, 1974.
- [40] A. A. Bakar, Z. Kefli, S. Abdullah, and M. Sahani, “Predictive models for dengue outbreak using multiple rulebase classifiers,” in *Electrical Engineering and Informatics (ICEEI)*, 2011 International Conference on, 2011, pp. 1–6.
- [41] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.
- [42] E. H. Mamdani, “Application of fuzzy algorithms for control of simple dynamic plant,” *Proc. Inst. Electr. Eng.*, vol. 121, no. 12, pp. 1585–1588, 1974.
- [43] A. Govindarajan, “A soft computing framework to evaluate the efficacy of software project management,” in *Intelligent Systems and Control (ISCO)*, 2015 IEEE 9th International Conference on, 2015, pp. 1–6.
- [44] A. Amindoust, S. Ahmed, A. Saghafinia, and A. Bahreinejad, “Sustainable supplier selection: A ranking model based on fuzzy inference system,” *Appl. Soft Comput.*, vol. 12, no. 6, pp. 1668–1677, 2012.
- [45] A. R. Gilal, J. Jaafar, M. Omar, S. Basri, and I. Din, “Balancing the Personality of Programmer: Software Development Team Composition,” *Malaysian J. Comput. Sci.*, vol. 29, no. 2, pp. 145–155, 2016.
- [46] D. Varona, L. F. Capretz, and Y. Piñero, “Personality types of Cuban software developers,” *Glob. J. Eng. Educ.*, vol. 13, no. 2, pp. 77–81, 2011.
- [47] G. Mourmant and M. Gallivan, “How Personality Type Influences Decision Paths in the Unfolding Model of Voluntary Job Turnover: An Application to IS Professionals,” *Business*, pp. 134–143, 2007.

APPENDIX 1

GRADE	MARKS	GRADE POINT	STATUS
A	85-100	4.00	Excellent
A-	80-84	3.67	
B+	70-79	3.33	Good
B	65-69	3.00	
B-	60-64	2.67	Pass
C+	55-59	2.33	
C	50-54	2.00	
D+	45-49	1.5	Fail
D	40-44	1.00	
F	0-39	0.00	

Usability of Government Websites

Mahmood Ashraf¹, Faiza Shabbir Cheema¹, Tanzila Saba², Abdul Mateen¹

¹Department of Computer Science

Federal Urdu University of Arts, Science and Technology, Islamabad, Pakistan

²College of Computer and Information sciences

Prince Sultan University Riyadh 11586 Saudi Arabia

Abstract—Usability of Government websites plays pivotal role in order to provide benefits and services to the citizens. This study presents a usability evaluation for investigating the Nielsen's usability attributes in Government websites. Based on the previous studies, a proposed website template is used in this study. This template is compared with a selected Government website. Thirty (30) participants performed three (3) representative tasks for each website. The results show that the user responses for the parameters of efficiency, memorability and pleasantness are improved for the proposed template. This effort is a part of the study that may lead to the principles for improving the usability of Government websites of Pakistan.

Keywords—Usability; statutory bodies websites; government websites

I. INTRODUCTION

The Web is currently the main source of providing computer services to reach a larger number of users having different characteristics [1]. After the commercialization has been started on the Internet, the interactive media segment of the Internet, organizations and people hustled to put Web pages and substance on quality of experience that people have is a concern, and one challenge is to ensure the usability of the site [2].

In order to provide government services to the public, Government websites are important windows [3] to its citizens. While visiting a website, usability and user experience are the major challenges [2]. To overcome these challenges, emphasis on usability and user experience is paramount. One way of achieving this is by measuring the easiness of the website's interface [4].

Usability is the quality attribute that measures the easiness of an interface [4]. Battleson et al. (2001) asserted that usability testing is the best approach to assess a website's usability [5]. The meaning of usability, overall, proposes that there are four common factors that influence usability of the interactive system clients, assignments, innovation and setting. The features studied by Bruno et al. (2005), were grouped into above four factors [6].

Usability evaluation concentrates on how well clients can learn and utilize an item to attain their objectives. It also states to how clients are content with that method. To collect this data, specialists use a variety of procedures that collect response from customers about a present site or ideas related to a new site [7]. The articulation, "test early and frequently", is especially fitting with regards to usability testing [7]. There are many methods for the evaluation of usability of the websites

like scenario, paper prototype, email, think aloud procedure, co discovery learning, eye tracking and user testing.

This paper reports a user study that was conducted to investigate whether the designed template fulfills the Nielsen's usability attributes. For this purpose a few tasks were set which were performed by thirty users and pre-test was filled before performing these tasks and post-test questionnaires were filled by the users after performing these tasks. Quantitative data was collected through the collection of results on the basis of filled questionnaires and qualitative data was also collected by observing the users.

II. LITERATURE REVIEW

The meaning of usability, overall, proposes that there are four mutual factors that influence the usability of the interactive system clients, responsibilities, tools and setting. The features investigated by [6], were come together into these four factors. Many researchers performed usability studies for websites, for mobile web browsing, for older adults. A usability study was performed in which users' mobile browsing experience was evaluated in comparison to desktop Web browsing [8]. Two usability studies were conducted to study whether there were variances in how older adults work together with the Web and whether changes in content size would influence execution [9].

Another research principally focused on Web usability and older adults and set up that even when Web experience is organized, older adults still revealed less Web capability than younger adults. They bolstered the theory that Web aptitude is fundamentally impacted by how clients took in the Web - or their total time spent in community oriented learning situations (gaining from and with others) - instead of exactly to what extent or how regularly they have utilized it [10].

As individuals give careful consideration to the possibility of open administration, government sites end up noticeably vital window to the resident, which can help the legislature to give open administration [3]. In many countries, many researchers performed usability evaluation of many government website. For example: Theory of usability engineer and information architecture was applied on the Canadian Government website [11]. Eleven websites in Zhejiang Province which were of government were taken as samples. Based on the prior research Quasi-experimental study was applied to tell usability issues in these websites and then comprehensive them to a boarder extent [3]. A user study was performed by Yuan and Zhongling having 24 students as participants. Every one of them was understudies with at any

rate essential site surfing knowledge. Be that as it may, with a specific end goal to make the outcome more powerful and precise, those were chosen who have never utilized government gateways in Zhejiang Province some time recently [3].

In order to understand how to make data vibrant and stress-free to be comprehensible theories have been discussed with respect to following aspects content concise not complicated, structure clear not disorder, data arrangement according to client's judgment not creators' logic, award importance to client familiarity and sensitivity [11]. In order to understand how to provide facts more meaningfully, special attention should be given on the followings. Masterful nature of data conveyance, landing page plan, great structure of peruse and route, without influencing download speed, utilize illustrations, outlines, multi-media and different methods for articulation, well-disposed data interface [11]. As per this examination, utilizing the above hypotheses, the data design of Canada government site has a few components as the accompanying: Rational utilization of shading to express data and give route, help clients comprehend and judge the data substance whenever, diversity of data request and meet the full needs of clients, all-round and multi-dimensional show of data, content fitting [11].

Electronic Government sites of the focal administration of Indonesia was assessed to investigate the ideas and models for gigantic substance sites for creating nations by exploring the qualities and how it spurs the clients in perusing the destinations to look for data and utilize others administrations [12]. Study demonstrated that the clients of Indonesian e-Government are originated from various social foundation and diverse era hole. They have distinctive recognition and fulfillment to the substance of the e-Government sites. To enhance the ease of use, availability and the adequacy of e-taxpayer supported organizations for residents, it is important to accumulate the feeling of the e-Government sites clients by polls frequently to offer a suitable plan and give helpful data [12].

Usability of five websites (Singapore, Korea, Japan, Hong Kong and Malaysia) of Asian countries is performed by Jati and Dominic [13]. Thirty governmental Websites were examined after launching e-government program in Jordan to evaluate ease of access, ease of use, clearness, and approachability to civilians' demand [14]. In view of the assessment comes about and through a manual testing, it was discovered that most of the Websites don't utilize a similar outline, where there ought to be steady and ought to be utilizing similar guidelines and elements [14].

III. METHOD

A. Design

To perform this user study one website of Statutory Body Website has been chosen to compare it with the designed template.

- Designed Template of SBW (shown in Fig. 1).
- Frequency Allocation Board (FAB) (<http://www.fab.gov.pk>) (shown in Fig. 2).

To evaluate the usability attributes for these two websites, user study-III Performa was prepared having introduction form, consent form, three representative tasks were designed which were performed by each user for each website. These tasks were performed by each user for each website in different order.



Fig. 1. Designed template of SBW.



Fig. 2. FAB.

Can you find the following?

- Phone numbers to contact this organization.
- The act of this organization. (This organization is established through the act of Parliament and Act is available on the website.)
- The vision statement of the organization.

To check the usability of the SBWs, a pre-test questionnaire was prepared to get the general information about the participants.

Post-test questionnaire was prepared to check the usability attributes for these websites. This questionnaire includes the three components of Nielson's attributes of usability which were asked for each website from the users. Each question has two more sub questions so that the in depth information can be taken regarding usability components.

B. Participants

30 participants have performed the tasks for the evaluation of attribute of usability for two websites in different order. Each user performed the tasks for each website. Out of 30 participants 40% were males and remaining were females.

C. Experimental Design

Test was conducted on different dates. At a time only one user was taken to test the usability. Websites were randomly evaluated for usability in different order. Users of different age group and different educational background were taken so that the results cannot be biased. Branded core i3 desktops having Windows 7 were provided to all users with an internet facility to perform this user study.

D. Procedure

The participants began the study by first completing the consent form. The previously mentioned four tasks were performed by each user for each website in different order, which is as follow:

IV. COMPILATION OF RESULTS

After the performance of the user study, the data were collected and results were carefully analyzed.

A. Pre-Test data Analysis

According to these results, out of 30, 6 users have qualification MS and 5 have bachelor degree, 17 have master’s degree. There was no PhD user and only 2 had intermediate qualification. All the information of pretest results have been summarized in Table 1.

B. Post-Test Data Analysis (Frequency Allocation Board (FAB))

30 users performed the requested tasks. Two users strongly agreed that it is quick to perform these tasks. 8 users agreed upon that it is quick to perform these tasks. 18 users disagreed and 2 users strongly disagreed that it is quick to perform these tasks. Out of 30, 1 user strongly agreed that home page tasks are quick to perform. 10 users agreed upon while 1 user remained neutral, 17 users disagreed and 1 strongly disagreed. Out of 30, 1 user strongly agreed that other page tasks are quick to perform. 5 users agreed upon while 10 users remained neutral, 13 users disagreed and 1 user strongly disagreed for the quickness of other page tasks as shown in Fig. 3.

One user strongly agreed that it is easy to remember. 8 users were agreed upon that it is easy to remember. 3 users remained neutral while 18 users disagreed and no one strongly disagreed that it is easy to remember these tasks. Out of 30, users strongly agreed that home page tasks are easy to remember. 10 users agreed upon while 5 users remained neutral, 13 users disagreed. Out of 30, 2 users strongly agreed that other page tasks are easy to remember. 2 users strongly agreed that other page tasks are easy to remember, while 15

users remained neutral, 10 users disagreed and 1 user strongly disagreed for the easy to remember of other page tasks.

TABLE I. PRE-TEST DATA ANALYSIS RESULTS

Qualification	Matric	0
	Intermediate	2
	Bachelors	5
	Masters	17
	MS	6
	PhD	0
Age	Below 20	0
	20-30 years	8
	30-40 years	18
	40-50 years	4
	Above 50	0
Gender	Male	12
	Female	18
Occupation	Student	8
	Government Officer	12
	Private Officer	4
	None of these	7
Use Internet	Daily	18
	Once a week	9
	Once in a month	1
	Once in a year	0
	Sometimes	2
Used Government Websites	Yes	21
	No	9
Purpose to use Government Website	To get any information	14
	To see job advertisement	12
	To avail any service	9
	For research purpose	4
	For any other purpose	6

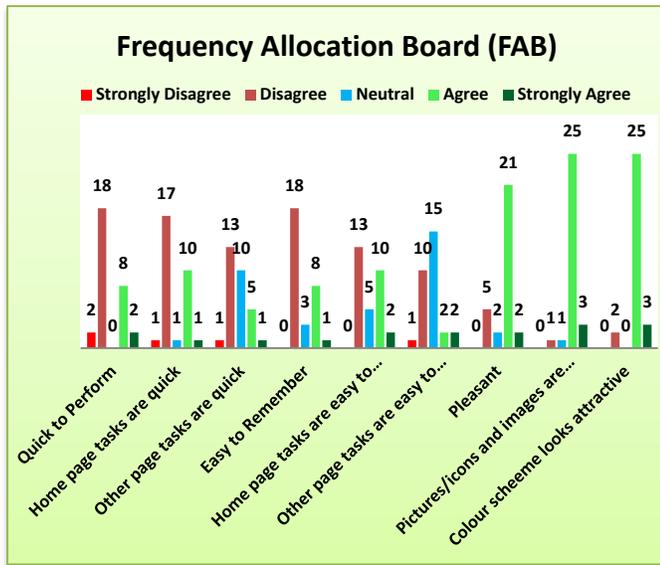


Fig. 3. NAU results of FAB.

Two users strongly agreed that it is feeling pleasant. 21 users agreed upon that it is pleasant. 2 users remained neutral while 5 users disagreed and no one strongly disagreed that it is Feeling pleasant. Out of 30, 3 users strongly agreed that pictures, images, icons are beautiful. 25 users agreed upon while 1 user remained neutral, 1 user disagreed, and no one strongly disagreed. Out of 30, 3 users strongly agreed that color scheme looks attractive. 25 users strongly agreed that color scheme looks attractive, while no one remained neutral, 2 users disagreed and no user strongly disagreed.

C. Post-Test Data Analysis (Designed Template of Statutory Body Website (SBW))

Thirty users performed the desired tasks mentioned as Annex-A. 27 users strongly agreed that it is quick to perform these tasks. 3 users agreed upon that it is quick to perform these tasks. Out of 30, 28 users strongly agreed that home page tasks are quick to perform and 2 users agreed upon. All the users strongly agreed that other page tasks are quick to perform as shown in Fig. 4.

Twenty eight users strongly agreed upon that it is easy to remember and two users agreed upon that it is easy to remember these tasks. Out of 30, 27 users strongly agreed that home page tasks are easy to remember and 3 users agreed upon. Out of 30, 27 users strongly agreed that other page tasks are easy to remember and 3 users agreed that other page tasks are easy to remember.

Twenty eight users strongly agreed that it is feeling pleasant while two users agreed upon that it is pleasant. Out of 30, 26 users strongly agreed that pictures, images, icons are beautiful and 4 users agreed upon. Out of 30, 25 users strongly agreed that color scheme looks attractive and 5 users agreed that color scheme looks attractive. It is important to note that no user clicked against neutral, disagree and strongly disagree for all the questions of usability attributes.

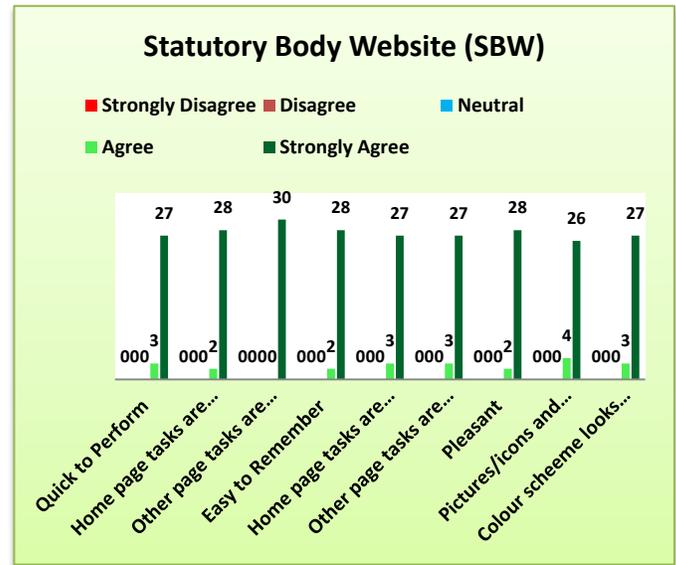


Fig. 4. NAU results of SBW.

V. RESULTS ANALYSIS

A. Further Data Analysis of all Websites

Quantitative results have been collected for this user study, which were evaluated. In user study-I all five components of Neilson’s Usability Attributes (NAU) were tested and it was realized that users faced more problems in three components. Therefore, in the second and third user study these three components were tested; Quick to perform, Easy to perform and Pleasantness. For this two more questions were attached against each attribute. Results were compiled for each attribute and shown in Fig. 5.

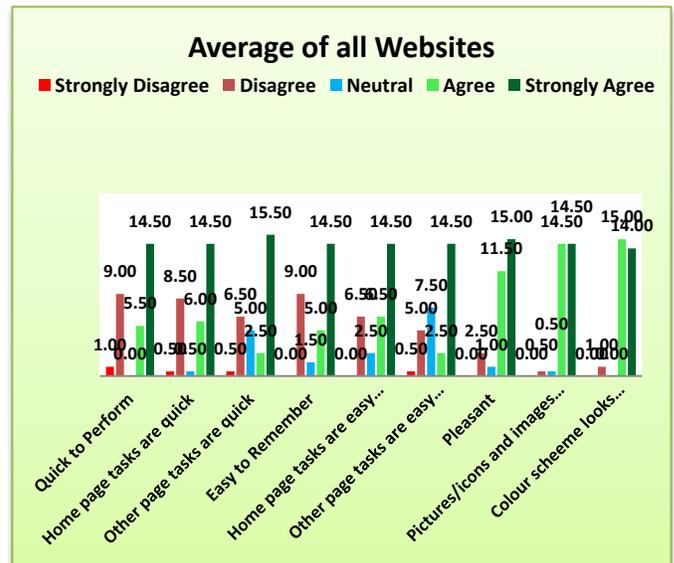


Fig. 5. Average of NAU results of all websites.

B. Quick to Perform

The average results for quick to perform are 1.00 for Strongly Disagree, 9.00 for disagree, nothing for neutral, 5.5

for agree and 14.50 for strongly agree. Similarly for quickness of home page tasks the results are 0.50 for Strongly Disagree, 8.50 for disagree, 0.5 for neutral, 6.0 for agree and 14.5 for strongly agree respectively. For other page tasks the results of quick to perform are 0.50 for Strongly Disagree, 6.50 for disagree, 5.0 for neutral, 2.50 agree and 15.5 for strongly agree.

C. Easy to Remember

The average results for easy to remember are nothing for Strongly Disagree, 9.0 for disagree, 1.50 for neutral, 5.0 for agree and 14.5 for strongly agree. Similarly for remembrance of home page tasks the results are nothing for Strongly Disagree, 6.5 for disagree, 2.5 neutral, 6.5 for agree and 14.5 for strongly agree respectively. For other page tasks the results of easy to remember are 0.5 for Strongly Disagree, 5.0 for disagree, 7.50 neutral, 2.5 agree and 14.5 for strongly agree.

D. Feel Pleasant

The average results for pleasantness are Zero for Strongly Disagree, 2.5 for disagree, 1.0 for neutral, 11.5 for agree and 15.0 for strongly agree. Similarly for pleasantness of icon, images the results are nothing for Strongly Disagree, 2.5 for disagree, 2.5 neutral, 14.5 for agree and 14.5 for strongly agree respectively. For pleasantness of color scheme the results of are nothing for Strongly Disagree, 1.0 for disagree, nothing for neutral, 15.0 agree and 14.0 for strongly agree.

E. Mean and Standard Deviation

The statistical analysis on the sampled data of both websites is calculated and summarized in the following Table 2.

TABLE II. MEAN AND STANDARD DEVIATION OF ALL WEBSITES

#		FAB		SBW	
		Mean	SD	Mean	SD
Q#.1	Quick to Perform	2.67	1.14	4.90	0.30
Q#.1 (a)	Home page tasks are quick	2.77	1.05	4.93	0.25
Q#.1. (b)	Other page tasks are quick	2.73	0.89	5.00	0.00
Q#.2	Easy to Remember	2.73	0.96	4.90	0.30
Q#.2. (a)	Home page tasks are easy to remember	3.03	1.02	4.97	0.18
Q#.2. (b)	Other page tasks are easy to remember	2.80	0.87	4.93	0.25
Q#.3	Pleasant	3.67	0.83	4.97	0.18
Q#.3. (a)	Pictures/icons and images are beautiful	4.00	0.52	4.87	0.34
Q#.3. (b)	Color scheme looks attractive	3.97	0.60	4.83	0.37

VI. CONCLUSION

This study analyzes and verifies the influence of three components of usability; efficiency, memorability and

pleasantness. A template has been prepared on the basis of the results of user study-I and user study-II. Designed template and the website having poor usability results from previous two studies were selected and tested by the users. Results were compiled and observed that the usability on the designed template of Statutory Bodies Websites has been improved. The results show that the user responses for the parameters of efficiency, memorability and pleasantness are improved for the proposed template. Therefore, we can design rules for the Government of Pakistan on the basis of this work.

REFERENCES

- [1] Dias, A. L., Fortes, R.P.D.M., Masiero, P.C., Watanabe, W.M., & Ramos, M.E. (2013). "An approach to improve the accessibility and usability of existing web system." Proceedings of the 31st ACM international conference on Design of communication. Greenville, North Carolina, USA, ACM: 39-48.
- [2] Davis, P. A. & Shipman, F. M. (2011). Learning usability assessment models for web sites. Proceedings of the 16th international conference on Intelligent user interfaces. Palo Alto, CA, USA, ACM: 195-204.
- [3] Yuan, L., & Zhongling, L. (2010). Experimental evaluation on government portal website's usability to 11 government websites of Zhejiang province. doi:10.1109/ICISE.2010.5688953
- [4] NNGROUP. (2012, January 4, 2012). "Nielsen Norman Group." Retrieved 11th May, 2015, from <http://www.nngroup.com/articles/usability-101-introduction-to-usability/>.
- [5] Battleson, B., Booth, A., & Weintrop, J. (2001). Usability testing of an academic library Web site: A case study. Journal of Academic Librarianship, 188-198.
- [6] Bruno, V., Tam, A., & Thom, J. (2005). Characteristics of web applications that affect usability: a review. Proceedings of the 17th Australia conference on Computer-Human Interaction: Citizens Online: Considerations for Today and the Future. Canberra, Australia, Computer-Human Interaction Special Interest Group (CHISIG) of Australia: 1-4.
- [7] Usability. (n.d.). In Usability. Retrieved August 24, 2015, from <http://www.usability.gov/>
- [8] Shrestha, S. (2007). Mobile web browsing: usability study. Proceedings of the 4th international conference on mobile technology, applications, and systems and the 1st international symposium on Computer human interaction in mobile technology. Singapore, ACM: 187-194.
- [9] Chadwick-Dias, A., McNulty, M. & Tullis, T. (2003). Web usability and age: how design changes can improve performance. Proceedings of the 2003 conference on Universal usability. Vancouver, British Columbia, Canada, ACM: 30-37.
- [10] Chadwick-Dias, A., Tedesco, D., & Tullis, T. (2004). Older adults and web usability: is web experience the same as web expertise? CHI '04 Extended Abstracts on Human Factors in Computing Systems. Vienna, Austria, ACM: 1391-1394.
- [11] Zhou, X. (2009). Usage-Centered Design for Government Websites - A Practical Analysis to Canada Government Website. doi:10.1109/ICIC.2009.84
- [12] Nariman, D. (2010). E-Government Websites Evaluation Using Correspondence Analysis. doi:10.1109/CISIS.2010.13
- [13] Jati, H., & Dominic, D. D. (2009). Quality Evaluation of E-government Website Using Web Diagnostic Tools: Asian Case. doi:10.1109/ICIME.2009.147
- [14] Al-Soud, A. R., & Nakata, K. (2010). Evaluating e-g overnment websites in Jordan: Accessibility, usability, transparency and responsiveness. doi:10.1109/PIC.2010.5688017

InstDroid: A Light Weight Instant Malware Detector for Android Operating Systems

Saba Arshad

Department of Computer Science
COMSATS Institute of Information Technology
Islamabad, Pakistan

Munam Ali Shah

Department of Computer Science,
COMSATS Institute of Information Technology
Islamabad, Pakistan

Rabia Chaudhary

Department of Computer Science,
Bahria University
Islamabad, Pakistan

Neshmia Hafeez

Department of Computer Science,
COMSATS Institute of Information Technology
Islamabad, Pakistan

Muhammad Kamran Abbasi

Department of Distance Continuing & Computer Education
University of Sindh
Hyderabad, Pakistan

Abstract—With the increasing popularity of Android operating system, its security concerns have also been raised to a new horizon in past few years. Different researchers have introduced different approaches in order to mitigate the malware attacks on Android devices and they succeed to provide security up to some extent but these antimalware techniques are still resource inefficient and takes longer time to detect the malicious behavior of applications. In this paper, basic security mechanisms, provided by Google Android, and their limitations are discussed. Also, the existing antimalware techniques which lie under the basic detection approaches are discussed and their limitations are also highlighted. This research proposes a light weight instant malware detector, named as InstDroid, for Android devices that can identify the malicious applications immediately. Through experiments, it is shown that InstDroid is an instant malware detector that provides instant security at low resource consumption, power and memory, in comparison to other well-known commercial antimalware applications.

Keywords—Android; static; resource efficient; power consumption; memory; detection rate; accuracy

I. INTRODUCTION

Smart phones have become a necessary part of everyday life. From businessman to a common person, everyone uses smart phones to perform different tasks depending upon their needs. Android devices provides attractive and easy to use features to the users due to which they are known as most popularly used devices from previous few years [1]. Android phones store the critical data related to the personal as well as professional life of a person. This data can be in the form of important transaction details, pictures, SMS and official encrypted files. It is important to ensure the security of such data in smart phones. Large number of malwares had been designed to infect and intrude into the smart phones in order to exploit the privacy of the user [2]. The mobile malware

designers exploit the vulnerabilities that exist in the Android operating system. Android operating system is an open source platform that allows the installation of third party applications from App-stores other than Google play store for example PandaApp [3] and GetJar [4]. This openness becomes the opportunity for malware developers to harm the user's data and is the reason for several issues such as invalid access from one resourceful application to the other (information leakage), permission escalation, repackaging application to infuse malicious code and Denial of Service (DoS) attacks.

In order to mitigate these issues, researchers have developed lot of detection systems by using different approaches to ensure the security up to some extent. The basic approaches used by malware detection approaches includes static analysis and dynamic analysis. Static analysis techniques monitor the behavior of application without running the application on device. It scans all the code of application without running the application due to which it is not able to detect the runtime malicious behavior of applications. In dynamic analysis technique, run time behavior of application is monitored by executing the application on emulator or real device for a specific time period. These analysis techniques enable the antimalware systems to identify the malicious applications and protect the Android devices.

Android smartphone devices are usually resource constrained. They have limited battery power and storage. Due to this reason, detailed static and dynamic analysis cannot be performed on Android devices. In order to overcome this limitation, researchers have developed cloud based malware detection systems. Although these security systems shift the workload from mobile device to cloud server, but the service becomes expensive and network dependent. If the detailed analysis at server takes longer time, it is possible that during

this time period, the malicious application might get the control over device and compromise the device. An efficient and very light weight system is the necessity of time which can provide protection to Android devices against known malware types and their variants at the instant when the application is installed on the device at very low resource consumption.

In this research, InstDroid, a light weight malware detection system, is proposed that can provide instant detection of malicious applications as soon the user will install the application. It immediately identifies the malicious applications through quick scan and notifies the user about it. The heavyweight Android malware tools consume a lot of power and memory while the smart phones are constrained by resources. InstDroid is able to detect the malware using very negligible amount of hardware resources of Android devices, thus not affecting the performance of the device.

Rest of the paper is organized as follows: Section II discusses about basic security mechanisms provided by Google Android to the Android devices and user's data. Basic approaches for malware detection, static and dynamic analysis, and deployment systems are discussed in Section III. Section IV describes about the proposed malware detection system, InstDroid. The experimental results are explained in Section V and Section VI concludes the paper and future work is also discussed in this section.

II. BASIC SECURITY MECHANISMS & THEIR LIMITATIONS

This section discusses the basic security mechanisms provided by Google Android and their limitations. These security mechanisms include permission framework, application sandboxing and Bouncer, shown in Fig. 1.

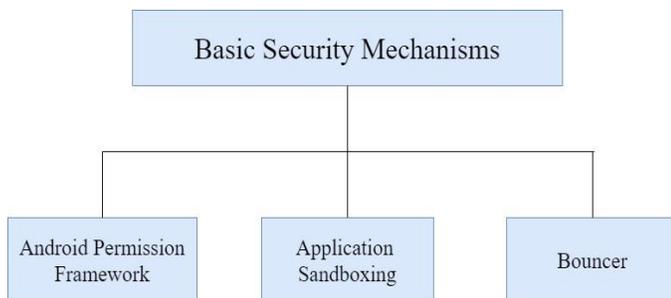


Fig. 1. Basic security mechanisms provided by Google Android.

A. Android Permission Framework

By default, an Android application has no permissions linked with it until the application requires special resources in order to operate. Different permissions have different purposes associated with them but they are used in order to limit the access of the application to the critical resources of device such as camera, SMS storage and Bluetooth permissions, etc. After careful inspection of these permissions, it is up to the user whether he wants to install the application or not [12]. There are four major categories of permissions: Normal, Dangerous, Signature and SignatureOrSystem [22]. Normal permissions are low level permissions that allows the (requesting) application to access the restricted application level features with only minimum level risk attached to other applications, the system, or the user. Dangerous permissions are high risk

level permissions and can be consequently used to harm the user's device and data. Signature and SignatureOrSystem permissions are only used by the system applications or the applications which are added by the manufacturer. Any user application requesting such permissions can be malicious. Although, permission system provides information to users about applications behavior up to some extent but due to lack of technical knowledge about these permissions and their use, by the applications, users usually ignore the permissions and simply install the applications. This makes Android permission mechanism completely ineffective to provide security against the access of unnecessary resources by newly installed application, which might be malicious.

B. Application Sandboxing

Android uses application sandboxing mechanism which separates the application associated data and code implementation from other applications. Each Android application runs within its separate space or sandbox, having no conflict with other applications or interaction, unless a particular application has been assigned special privileges to communicate with other applications. For better protection of Android application's data, Android kernel executes the Linux Discretionary Access Control (DAC) to efficiently manage and protect the device from getting misused. Each application process is protected with an assigned unique ID (UID) within its isolated sandbox [13]. The isolated application communicates with each other through a method known as Inter-Component Communication (ICC) or Binder. Android middleware allows the ICC between different components of the application. The ICC very smoothly takes care of transferring the request from user to the destination applications. After that applications can access the components or services of other applications as a service [12]. This ICC process is used by malware applications too in order to control the other applications and perform malicious activities on the device. Privilege escalation or permission escalation attacks were actually possible because of the loopholes that exist within the Android operating system, in order to get access to the assets that are hidden or protected from the user of application. This series of attacks can result into the leakage of fatal information because of the unauthorized access of resources to the application than the intended access of resources. Android applications might have such components that have been added into it through external resources. In this case these exported components can be misused in order to get the access to critical permissions [11].

C. Bouncer

Bouncer is a malware detection tool deployed at Google Play Store for the analysis of all the applications available at Google Play Store. The main purpose of the bouncer is to provide a security check looking for malicious software containing malware, spyware, and Trojans. This kind of applications can be used to intrude the privacy of the user, selling it to the blackmailers or using it for more harmful purposes. Bouncer keeps on analyzing the applications continuously. If any application is detected as malware, it is instantly removed from the Play Store. Although, Bouncer performs its job very well but still there exist some malware

applications on Google Play Store that remains undetected by Bouncer, reported in a research [5].

III. MALWARE DETECTION APPROACHES

In spite of the security mechanisms provided by Google Android, malware attacks are increasing every year [6]. Lot of research has been done to protect the Android devices from malware attacks. Major approaches used for the malware analysis includes static analysis and dynamic analysis.

A. Static Analysis

Static analysis techniques monitor the behavior of application without running the application on device. Kirin [7], Drebin [8] and RiskRanker [9] are well known examples of antimalware techniques which performs static analysis to explore the static features of Android malware. It scans all the code of application but cannot detect dynamic loading of malware code. Also, the encrypted malicious code remains undetected. In [10] authors have categorized static analysis based malware detection techniques as signature based malware detection, permission-based malware detection, and dalvik byte code malware detection. The signature-based detection technique extracts the signatures of the applications and then matches it with the database of known malware signatures [9]. *AndroSimilar* [11] and *DroidAnalytics* [12] are signature based detection systems.

Permission based detection is a light weight malware detection method which also falls under the category of static analysis. In [13], authors have proposed the system which performs analysis on permissions declared in the Android manifest file and then analyzes if the application is over privileged or not. In the manifest file of the application, they extract three major features i.e. permissions, intent filters, process number and a total number of predefined permissions. On basis of these features, they compare it with the list of already known keywords. They tested 365 samples on the total to determine the efficiency of the proposed system. The proposed system almost provides 90% detection rate. In [14], [15] and [16], authors have also used permission based detection method.

Dalvik byte code analysis performs the instruction level code analysis to find out the malicious behavior of the applications. But it occupies more storage space due to the instruction level analysis of the code and hence consuming more power resources, therefore making it less likely to be more productive on resource constrained devices like smart phones [17]-[19].

B. Dynamic Analysis

Dynamic analysis technique provides run-time monitoring of the applications. *TaintDroid* [20], *DroidRanger* [5] and

DroidScope [21], use the dynamic analysis to monitor the run-time behavior of the application. Dynamic analysis can detect the dynamic malicious payloads.

DroidDolphin [22] uses dynamic analysis that takes support of GUI-based testing, big data and machine learning for the detection of Android malwares. API calls are monitored by *API Monitor* [23] during execution of apk. Logs are collected by installing instrumented apk file on virtual device of Android. Sandboxing is done through *DroidBox* [24] for having dynamic logs. Testing tool, *Monkeyrunner*, is combined with *APE* [25], that is used for GUI based event simulation. Events are represented by n-grams and features are given as input to *Support Vector Machine* [26] algorithm that classifies the applications. Emulation and testing phases become complex for future testing because of large data set.

CopperDroid is presented in [27] that works on top of QEMU and performs dynamic analysis. Behaviors are analyzed by system calls tracking and centric analysis. The *CopperDroid* analyzes malware by information extraction from manifest file. The *CopperDroid* was evaluated for two sets of malwares and there is no static analysis involved.

Although dynamic analysis overcomes the limitations of static analysis, but it can only analyze the code which executes during monitoring interval and is not able to detect malicious code which does not execute during monitoring period.

C. Cloud Based Detection

These analysis approaches, static and dynamic, can be used at either mobile device or at cloud for detection of malwares. As mobile devices are resource constrained due to which malware detection systems cannot perform detailed and effective analysis on mobile devices. To develop an effective and accurate malware detection system, researchers have deployed the analysis and detection mechanism at clouds.

A cloud based intrusion detection and response framework was developed and discussed in [28], that analyzes behavior of a device and in case of unusual events, it performs different appropriate actions. This framework can work with minimum resources and can produce real and accurate detection and responses for registered devices. A key point of this architecture is to copy user inputs in real time. Proxy settings are configured by installing a software and proxy server replicates the conversation between internet and device and sends it to emulated environment for malware detection and analysis. A light weight agent is also involved for gathering info, sending it to emulated environment and waiting for responses and actions. Proposed framework was deployed to Android-equipped HTC Droid Incredible devices but attack graph does not automatically take actions in an emulated phone environment, like computer systems.

TABLE I. CLOUD-BASED ANDROID MALWARE DETECTION TECHNIQUES

Ref.	Year	Implementation	Limitations
[28]	2011	Working prototype	Android-equipped HTC Droid Incredible devices and attack graph does not work for emulated devices
[29]	2014	Framework	Need device user, app store and security professionals' association
[30]	2012	Security system	Cloud can be crashed because of single component failure
[31]	2012	Architecture	Needs number of detection engines
[32]	2014	Security Mechanism	Mobile interference is less due to of cloud services
[33]	2015	Experimental	Requires different configurations

TABLE II. RESOURCE UTILIZATION ANALYSIS FOR ANDROID MALWARE DETECTION

Ref.	Year	Implementation	Evaluated Parameters	Limitations
[34]	2013	Prototypes	Battery level, FPR, cutoff drop value	Cut off values may affect the results
[35]	2016	Experimental	Accuracy, FPR, FNR	Specific pattern for resource utilization was not considered
[36]	2014	Prototype	FP percentage	Need user efforts and time to create profiles

In [29], authors proposed a cloud based detection and prevention approach. When a user makes request for any application, the request is sent to known libraries. If the application is found in libraries then it is declared as safe or malicious, on the basis of classification of that application. If application is not found in libraries then application is declared as unknown and send to malware detector that downloads the application. The malware detector performs both static and dynamic analysis and declares the application as safe or malicious for users on the basis of classification results. All these operations are performed at cloud, that keeps resources of mobile devices conserved. Mobile devices just deal with libraries for finding application classification, as safe or malicious. The major limitation of this technique is that it is highly dependent on the Internet services and cloud system. If any component at cloud fails to perform its operations, security will not be provided. This approach requires mobile users, app stores and IT security professional's association.

Qian *et al.* [30] proposed a cloud based security system which provides security to Android devices by detecting malwares, pours out harmful application and provides data backup facility. Android devices have an agent/client that communicates with the cloud. Connection between client and server should be fair enough for sending malicious applications to cloud. Authors presented agent and server modules to elaborate the system clearly. Different features were implemented that provide security. VPN builds connection between device and cloud for user safety. A transparent proxy is used to communicate data between internet and proxy server that provides security to users. Malicious applications can also send information to suspicious addresses. Push function is used to discard illegal packets that are sent to devices. Management server has facility to detect malicious applications by running different algorithms that may be available in market or may use static, dynamic zero-day analysis programs in an emulated environment or can be executed on the PC. Backing up of data is also maintained at cloud. Proposed system uses limited device resources but the service might be expensive for the users.

The security system proposed in [31], contains a host that works with the cloud provided services and it has a vast range of signature database. Different detection modules can be made run simultaneously. Virtualization helps a lot to detect malware

and large number of users can be scaled over the network. Proposed system provides services such as creating a clone of the device and a proxy in cloud is used for identifying memory, system calls invoked on run time. Different open source antiviruses are used to detect malwares. Host agent is a process that is installed on the device. It performs inspection on files and compares the files against a cache of files. If file is absent in cache it is sent to the cloud for further analysis and recovery actions are taken accordingly. After analysis, it is placed on local and cloud caches. This approach needs number of detection engines to provide large detection exposure.

According to the research performed in [32], proposed system consists of three modules. First module classifies applications as light, heavy, medium, very light and very heavy, based on the signatures, permissions and services etc. Second module has local server that creates all user's feedback. Package name for feedback, date of report, IMEI number for report receiving and report that has '1' and '0' values for good and bad applications. In third module, filters are applied to applications for permission set and the generated report is sent to server. Algorithm is used for malware detection and works on confidence index. If confidence index is greater than 50 %, there is possibility of malware if not then application is considered to be safe. Mobile resource consumption is less due to the use of cloud services.

Table 1 shows cloud-based detection for malicious applications in Android. Cloud-based detection requires internet availability, detection engines, files uploading on cloud which consumes large amount of power. Major limitations of such techniques include that any component failure at cloud may affect the whole detection system. Mobile or host device have to wait for the cloud response in order to provide security on Android devices.

D. Resource Utilization Based Detection

Although cloud based detection systems allow deep analysis of applications but at the cost of heavy servers and they are dependent on cloud server's response. Also, the power consumption at mobile device increases if the device is at large distance from the server and communicates with cloud server for detection purpose. Many researchers have developed malware detection systems to overcome the power consumption limitations of cloud based detection systems.

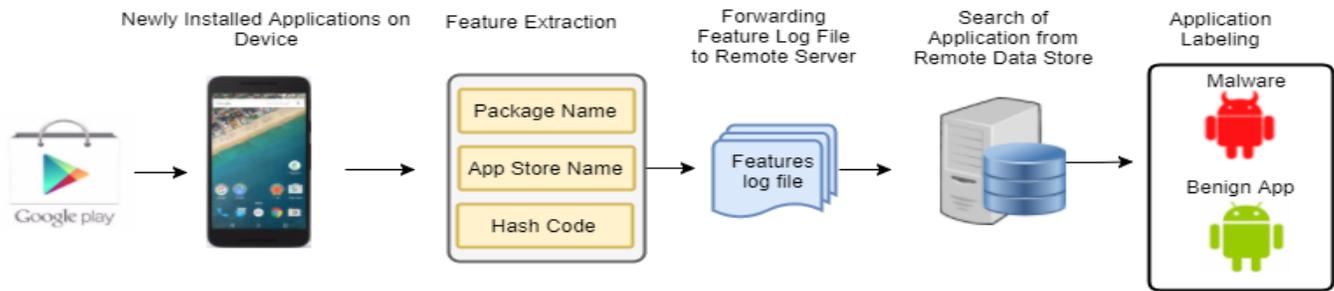


Fig. 2. Workflow diagram of InstDroid.

In [33], authors have observed effectiveness of two techniques for malware detection. Prototypes were developed for Android platform. Techniques include normal and location specific power profiles for phones. Experiments were performed to detect malware and minimizing power consumption. Authors used SMS spam and user tracking simulators for the evaluation of techniques. Normal power profile technique takes power utilization as a time function. Normal battery consumption rate is measured initially after which the system starts monitoring the power drainage pattern. Location power profile works over an extended time, based on the location i.e. whether playing games at home or using browser at airport etc. A program was written by authors to measure power utilization for working models. For first discussed technique cut off value may affect results of prototype. For second discussed technique, anomalies were predicted just for two locations.

Canfora *et al.* [34] proposed a malware detection technique that detects presence of malicious applications by analyzing the device resources such as memory, CPU, and network. Proposed methodology has three components: numerical feature set related to application behavior, a procedure in which applications are executed in a balanced environment and performs data collection, method for analyzing the collected data. Monkey tool was used as a debugger. Data is analyzed by using machine learning techniques.

Three different detection techniques are mentioned in [35] that are used in Android malware detection for testing and data collection. These techniques include location based detection, time based detection and a hybrid, combination of both. The basic idea of these techniques is to investigate the usage of battery profiles to detect malwares. Battery usage will be more in case of malware attack. In first technique, profiles are created for normal battery usage, based on the user location, because battery usage may vary depending upon location. Second technology creates profile, based on time in which user uses the Android device. Third technology involves hypothesis that user uses Android device differently at different locations in different timings. SMS spam and location tracking simulations are performed by authors. Data collection and location based detection is done by standalone prototype. Data needs to be segmented after assortment correspondent to fall in battery level between two data points and average rate of charge per second. Standard deviation is calculated for each segment by standalone project. Abnormal battery usage is observed when a new segment is created for a location. Segments are also monitored for hours but during period of

6 hours, segments produce better detection results. When both these techniques are combined, false positive rate is reduced. A program is written to measure battery usage of the prototype by authors. Random values for location and time data segments were taken and tested for two simulators. Profile creation for specific location involves user presence at that location at different time.

Table 2 shows different techniques that are developed for enhancing the resource efficiency in terms of power. Keeping in view all the limitations of malware Antimalware techniques, discussed in literature, an instant malware detection system is proposed that can provide instant security against known malware families and their known variants, at low resource consumption.

IV. INSTDROID: THE MODEL

This research proposes a light weight and instant malware detection system for Android devices. This instant malware detector immediately detects the malwares and provides instant protection to Android devices from known malware types. This light weight Android security system consumes very negligible amount of hardware resources of resource constrained Android devices. Fig. 2 depicts the workflow of Instant malware detection system. When an Android user installs any application, InstDroid instantly initiates the detection mechanism and secures the Android devices.

A. Features

Features used for the detection of malicious applications are:

- 1) *Hash Code*: Hash code generated for application.
- 2) *Package Name*: Package name of application.
- 3) *Application Store Name*: Name of market from which the application is installed.

B. Working

Initially, when a user installs the application from Application store, InstDroid gets activated. It generates the hash code of application and extracts the features from the application code statically. Features extracted from the application includes package name and name of application store from which application is downloaded. These features are then forwarded to the remote server which is responsible for making decision about the application's behavior. Remote server contains the database of malware applications. When it receives the application's hash code, package name and App-store name from InstDroid client application, it immediately

looks into the malware database. An application is declared as malicious if one of the two conditions occurs:

a) Any record in the database contains the same package name and App-store name, sent by InstDroid client application.

b) Any record in the database contains the same hash code sent by InstDroid client application.

If the application package name and App-store name or hash code is not found in the remote server's database then the application is declared as legitimate.

Once the application is declared as legitimate or malicious, the decision is forwarded to the InstDroid client application which informs user about the application's behavior immediately. Fig. 2 describes the work flow of the proposed system.

V. EVALUATION

This section provides the experimental results which we have performed for evaluation of InstDroid. We have used Drebin's dataset of malicious application for identification of malware applications, as this dataset is claimed to be the largest dataset of malware applications.

A. Power Consumption

In the first experiment we have measured the power consumed by InstDroid and compared it with the real antimalware applications such as 360 Security [36], Avira Antivirus [37] and Avast Antivirus [38]. These antivirus applications are commercially available in Google official marketplace.

In most of the detection systems, the security service keeps on running in the background all the time which consequently affects the performance of the device and causes the resource drainage. InstDroid is a light weight detection system which is developed to overcome the limitations of the existing systems. It gets activated only when any application is installed on the device, performs detection mechanism and then stop running in the background. This is how the power consumption at real Android device is very low in comparison to the other malware detectors. Fig. 3 depicts the comparison between InstDroid and other antimalware applications. It can be observed that InstDroid consumes significantly low power in comparison to other devices.

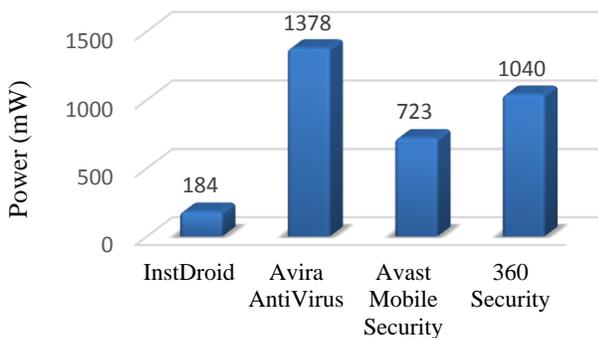


Fig. 3. Comparison of power consumed by different antimalwares.

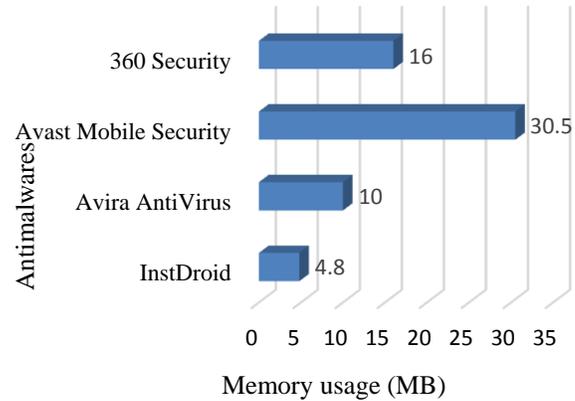


Fig. 4. Comparison of memory usage by different antimalwares.

B. Memory Consumption

The memory consumption and CPU usage of any application is directly proportional to the performance of the device. The large sized antivirus tools provide the efficient scanning of the applications on the cost of reduced performance and battery derail age of the device. The proposed system provides a very light weight mechanism for detecting the malicious properties as it requires very low amount of storage space to perform malware detection. Due to this low resource usage feature of InstDroid, performance of the device is not affected.

In this experiment, InstDroid is evaluated on the basis of memory consumption and the results are compared with the other well-known antimalware Android applications. Fig. 4 depicts the comparison of memory consumption by different antimalware systems. It can be seen that InstDroid is more resource efficient than the other antimalware tools.

C. Detection Time

Time taken by the antimalware system is also an important parameter for the evaluation. In this experiment, InstDroid is evaluated on the basis of detection time, time taken by the security system to detect the malicious behavior of application. Total time taken by the InstDroid to complete the detection process is compared with other antimalware applications. Fig. 5 shows the comparison of detection time between different anti-malwares. It can be seen that InstDroid is faster than all the other applications, just like its name – an instant malware detector.

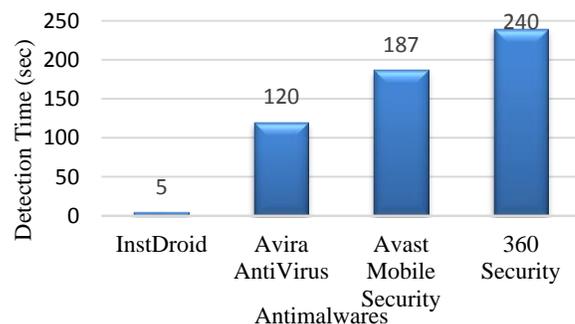


Fig. 5. Comparison of detection time between different antimalwares.

D. Detection Accuracy

In this experiment, the detection accuracy of antimalware system is measured and is compared with other commercial antimalware applications. This experiment is performed on 100 different malware applications and the detection accuracy of antimalware systems is observed, depicted in Fig. 6. Experimental results show that InstDroid achieves highest accuracy.

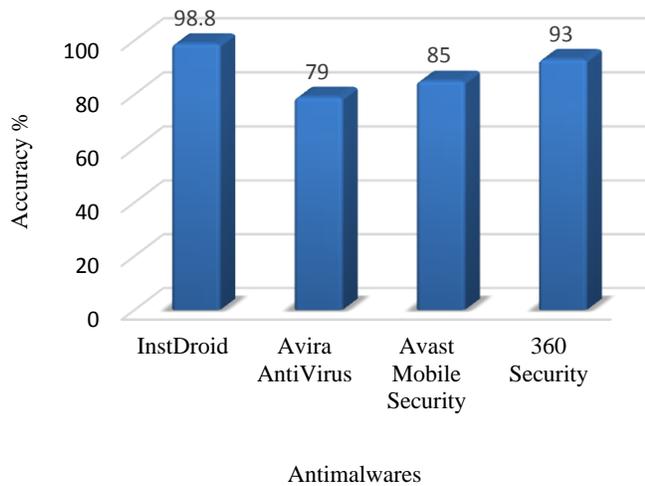


Fig. 6. Comparison of detection time between different antimalwares.

VI. CONCLUSION AND FUTURE WORK

With the increasing popularity of Android operating system, its security concerns have also been raised to a new horizon in past few years. Different researchers have introduced different approaches in order to mitigate the malware attacks on Android devices and they succeed to provide security up to some extent but they are still resource inefficient and takes longer time to detect the malicious behavior of applications. If any malware gets installed on the device, it is possible that it effects the device before the antimalware tool knows about the malicious behavior of application. InstDroid is the instant malware detection system which becomes active at the instant when application is installed on the device and in no time, it notifies about the application's classification to the user. It is a light weight malware detector that barely occupies the space of few megabytes and consumes significantly low power in comparison to other antimalware applications.

In future, we aim to enhance the dataset of malware applications so that InstDroid can detect the new malware families and their variants immediately. InstDroid can be integrated with other antimalware systems in a modular form, for instant detection of all the known malwares and their variants. As an example, different malware types and attacks are usually recorded in different countries. For such case, InstDroid can be used with addition of cache mechanism. In such a scheme, the data set of malwares, specific to the country, can be stored in cache for quick detection. This will provide instant detection of malwares and protection against them at low resource consumption.

REFERENCES

- [1] "Gartner Says Worldwide Sales of Smartphones Grew 7 Percent in the Fourth Quarter of 2016," 2017. [Online]. Available: <http://www.gartner.com/newsroom/id/3609817>. [Accessed: 28-Apr-2017].
- [2] "Mind the (Security) Gaps: The 1H 2015 Mobile Threat Landscape - Security News - Trend Micro USA." [Online]. Available: <http://www.trendmicro.com/vinfo/us/security/news/mobile-safety/mind-the-security-gaps-1h-2015-mobile-threat-landscape>. [Accessed: 08-Dec-2015].
- [3] "Android.PandaApp.com | Free Your Mobile Life!" [Online]. Available: <http://android.pandaapp.com/>. [Accessed: 01-Aug-2017].
- [4] "GetJar - Download Free Apps, Games and Themes APK." [Online]. Available: <https://www.getjar.com/>. [Accessed: 01-Aug-2017].
- [5] Y. Zhou, Z. Wang, W. Zhou, and X. Jiang, "Hey, You, Get Off of My Market: Detecting Malicious Apps in Official and Alternative Android Markets," in Proceedings of the 19th Annual Network and Distributed System Security Symposium, 2012, no. 2, pp. 5–8.
- [6] "Trend Micro Q2 Security Roundup Report | Androidheadlines.com." [Online]. Available: <http://www.androidheadlines.com/2015/08/trend-micro-q2-security-roundup-report.html>. [Accessed: 08-Dec-2015].
- [7] W. Enck, M. Ongtang, and P. McDaniel, "On lightweight mobile phone application certification," in of the 16th ACM conference on ..., 2009, pp. 235–245.
- [8] D. Arp, M. Spreitzenbarth, H. Malte, H. Gascon, and K. Rieck, "Drebin: Effective and Explainable Detection of Android Malware in Your Pocket," in Symposium on Network and Distributed System Security (NDSS), 2014, pp. 23–26.
- [9] M. Grace, Y. Zhou, Q. Zhang, S. Zou, and X. Jiang, "RiskRanker: Scalable and Accurate Zero-day Android Malware Detection Categories and Subject Descriptors," in Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services, 2011, pp. 281–293.
- [10] S. Arshad, M. Ahmed, M. A. Shah, and A. Khan, "Android Malware Detection & Protection: A Survey," Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 2, pp. 463–475, 2016.
- [11] P. Faruki, V. Ganmoor, V. Laxmi, M. S. Gaur, and A. Bharmal, "AndroSimilar: Robust Statistical Feature Signature for Android Malware Detection," in Proceedings of the 6th International Conference on Security of Information and Networks - SIN '13, 2013, pp. 152–159.
- [12] M. Zheng, M. Sun, and J. C. S. Lui, "DroidAnalytics: A Signature Based Analytic System to Collect, Extract, Analyze and Associate Android Malware," 2013.
- [13] R. Sato, D. Chiba, and S. Goto, "Detecting Android malware by analyzing manifest files," Proc. Asia-Pacific Adv. Netw., vol. 36, pp. 23–31, 2013.
- [14] B. Sanz, I. Santos, C. Laorden, X. Ugarte-Pedrero, P. G. Bringas, and G. Álvarez, "PUMA: Permission usage to detect malware in Android," Adv. Intell. Syst. Comput., vol. 189 AISC, pp. 289–298, 2013.
- [15] M. Qiao, A. H. Sung, and Q. Liu, "Merging Permission and API Features for Android Malware Detection," in 2016 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), 2016, pp. 566–571.
- [16] K. A. Talha, D. I. Alper, and C. Aydin, "APK Auditor: Permission-based Android malware detection system," Digit. Investig., vol. 13, pp. 1–14, 2015.
- [17] Y. Aafer, W. Du, and H. Yin, "DroidAPIMiner: Mining API-Level Features for Robust Malware Detection in Android," in Security and Privacy in Communication Networks, Springer, Cham, 2013, pp. 86–103.
- [18] E. R. Wognsen, H. S. Karlsen, M. C. Olesen, and R. R. Hansen, "Formalisation and analysis of Dalvik bytecode," Sci. Comput. Program., vol. 92, no. December 2012, pp. 25–55, 2014.
- [19] W. Zhou, Y. Zhou, X. Jiang, and P. Ning, "Detecting repackaged smartphone applications in third-party Android marketplaces," in Proceedings of the second ACM conference on Data and Application Security and Privacy - CODASKY '12, 2012, pp. 317–326.
- [20] W. Enck, P. Gilbert, S. Han, V. Tendulkar, B.-G. Chun, L. P. Cox, J. Jung, P. McDaniel, and A. N. Sheth, "TaintDroid: An Information-Flow

- Tracking System for Realtime Privacy Monitoring on Smartphones,” ACM Trans. Comput. Syst., vol. 32, no. 2, pp. 1–29, Jun. 2014.
- [21] L. Yan and H. Yin, “Droidscope: seamlessly reconstructing the os and dalvik semantic views for dynamic Android malware analysis,” in Proceedings of the 21st USENIX Security Symposium, 2012, p. 29.
- [22] W.-C. Wu and S.-H. Hung, “DroidDolphin: A Dynamic Android Malware Detection Framework Using Big Data and Machine Learning,” in Proceedings of the 2014 Conference on Research in Adaptive and Convergent Systems, 2014, pp. 247–252.
- [23] “API Monitor: Spy on API Calls and COM Interfaces (Freeware 32-bit and 64-bit Versions!) | rohitab.com.” [Online]. Available: <http://www.rohitab.com/apimonitor>. [Accessed: 22-Aug-2016].
- [24] “DroidBox.” [Online]. Available: <https://github.com/pjlantz/droidbox>. [Accessed: 22-Aug-2016].
- [25] S. Chang, “Ape: A smart automatic testing environment for Android malware,” Comput. Sci. Inf. Eng. Natl. Taiwan Univ. Taiwan, 2013.
- [26] A. Andrew, “An Introduction to Support Vector Machines and Other Kernel-based Learning Methods,” in Kybernetes, 2013.
- [27] K. Tam, S. Khan, A. Fattori, and L. Cavallaro, “CopperDroid: Automatic Reconstruction of Android Malware Behaviors.,” NDSS, 2015.
- [28] A. Houmansadr, S. A. Zonouz, and R. Berthier, “A cloud-based intrusion detection and response system for mobile phones,” in 2011 IEEE/IFIP 41st International Conference on Dependable Systems and Networks Workshops (DSN-W), 2011, pp. 31–32.
- [29] N. Penning, M. Hoffman, J. Nikolai, and Y. Wang, “Mobile malware security challenges and cloud-based detection,” in 2014 International Conference on Collaboration Technologies and Systems (CTS), 2014, pp. 181–188.
- [30] H. Qian and Q. Wen, “A cloud-based system for enhancing security of Android devices,” in 2012 IEEE 2nd International Conference on Cloud Computing and Intelligence Systems, 2012, pp. 245–249.
- [31] R. S. Khune and J. Thangakumar, “A cloud-based intrusion detection system for Android smartphones,” in 2012 International Conference on Radar, Communication and Computing (ICRCC), 2012, pp. 180–184.
- [32] M. Patil and M. Shelke, “Revisiting Defense against Malwares in Android using Cloud Services,” Int. J. Appl. or Innov. Eng. Manag., vol. 3, no. 3, 2014.
- [33] B. Dixon and S. Mishra, “Power Based Malicious Code Detection Techniques for Smartphones,” in 2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, 2013, pp. 142–149.
- [34] G. Canfora, E. Medvet, F. Mercaldo, and C. A. Visaggio, “Acquiring and Analyzing App Metrics for Effective Mobile Malware Detection,” in Proceedings of the 2016 ACM on International Workshop on Security And Privacy Analytics - IWSPA '16, 2016, pp. 50–57.
- [35] B. Dixon, S. Mishra, and J. Pepin, “Time and Location Power Based Malicious Code Detection Techniques for Smartphones,” in 2014 IEEE 13th International Symposium on Network Computing and Applications, 2014, pp. 261–268.
- [36] “360 Security - Antivirus Boost - Android Apps on Google Play.” [Online]. Available: <https://play.google.com/store/apps/details?id=com.qihoo.security>. [Accessed: 16-May-2017].
- [37] “Avira Antivirus Security - Android Apps on Google Play.” [Online]. Available: <https://play.google.com/store/apps/details?id=com.avira.android>. [Accessed: 16-May-2017].
- [38] “Mobile Security & Antivirus - Android Apps on Google Play.” [Online]. Available: <https://play.google.com/store/apps/details?id=com.avast.android.mobilesecurity>. [Accessed: 16-May-2017].

A 7-Layered E-Government Framework Consolidating Technical, Social and Managerial Aspects

Mohammed Hitham M.H, Dr. Hatem Elkadi H.K, Dr. Sherine Ghoneim S.G
Faculty of Computer and Information System
Cairo University
Cairo, Egypt

Abstract—E-Government has been hype for the last 2 decades and still several implementations do not reach the intended success. Different definitions and consequently different models of operations and assessment were developed. This required the formulation of various frameworks describing the different perceptions and understandings of e-Government. The different frameworks proposed tend to agree on a set of elements, but each framework seems to have one or few different elements, depending on the perception of the framework founder. Also, entire categories (or dimensions) of elements seem to be left out. Through a literature review and field survey, the authors identified challenges of an e-Government initiative, categorized in five dimensions: technical, adoption, organizational, strategy and cultural. Not all categories were covered in any of the existing government frameworks. This would prove to be awkward in the formulation of new government initiatives or in the assessment of existing ones and evolution plan. In an effort to represent the majority of the factors and elements involved in most e-Government initiatives, the authors present a proposed seven-layer-framework for e-government. The layers included are: 1) end user access layer, 2) e-government layer, 3) organization layer, 4) national infrastructure layer, 5) strategic layer, 6) social cultural layer, and 7) national execution layer. The proposed model is compared with existing models and demonstrates that it covers all the aforementioned dimensions.

Keywords—E-government; framework; e-government; challenges; decision support system (DSS)

I. INTRODUCTION

An e-government initiative is rather a complex endeavor, way beyond the technological complexity commonly presented by ICT professionals and vendors. Social, legislative and managerial aspects are major dimensions that are usually left out during the formulation of the E-Government implementation plans, and in the majority of the proposed frameworks. An integrative framework is required to provide a bird's eye overview of the diversity of factors and dimensions involved. In this quest, the different dimensions and factors pertaining to e-government initiatives were first identified; existing frameworks reviewed and then the proposed framework formulated. The rest of this paper is organized as follows: section two presents a survey of related work; identifying and categorizing the issues and challenges facing e-government, and the coverage provided by each work. Section 3 discusses the e-government existing frameworks and

the common features or layers covered in each, the Challenges addressed by each and finally the unaddressed ones. Section 4 presents the proposed e-government framework, discussing the significance of each element and the challenges it addresses. Section 5 presents a comparison between the proposed e-government framework and the other frameworks, identifying the layers reported in previous frameworks, as compared to the proposed one. Finally, Section 6 summarizes the paper findings and presents planned objectives and possible extensions based on this work.

II. CHALLENGE FACING E-GOVERNMENT IMPLEMENTATION

Through a review of the research literature concerned with the challenges facing e-government, we could categorize these challenges into five categories as shown in Table 1:

Technical [1]-[19], adoption [1]-[5], [7], [9], [11]-[20], organizational [1], [3]-[6], [8], [10]-[12], [14]-[17], [19]-[21], strategy [1], [3]-[5], [11]-[15], [19], [20] and cultural [1]-[6], [11]-[18], [22]. Many researchers have categorized challenges that faced e-government projects implementations, as summarized in Table 2.

III. EXISTING FRAMEWORKS

In our effort to identify the different e-government framework elements, the frameworks proposed by several authors are reviewed [14], [18], [23]-[27]. These frameworks are analyzed from two perspectives: Existing Layers (architecture) and the Categories of Challenges covered.

A. Framework Layers

The majority of existing frameworks covered included mainly the following main layers:

1) Infrastructure layer

According to (RoslindKaur, 2006) [23], the infrastructure includes network, backup and redundancy, and storage and according to (Mundy, D. and B. Musa, 2010) [18], the e-government framework includes infrastructure such as network, IT education, and IT administration. According to (Zakareya Ebrahim, 2005) [24] and (Sharma, 2003) [25] and the infrastructure includes servers, LAN, internet, and extranet.

The ICT infrastructure layer is Omni-existent in different frameworks: (Hatem Ben Sta, 2014) [14].

TABLE I. CATEGORIZED ISSUES AND CHALLENGES FACING E-GOVERNMENT

Type of challenges	
Technical [1]-[19]	<ol style="list-style-type: none"> 1. Lack IT infrastructure 2. security and privacy 3. training 4. Project management 5. information quality 6. system quality 7. requirement: incomplete, change 8. data or system integration 9. re-engineering process, 10. Limited skills of employees, 11. Support web in different language.
Adoption [1]-[5], [7], [9], [11]-[20]	<ol style="list-style-type: none"> 1. Limited Funding, 2. lack of resources, 3. top management support, 4. Web Content, subscription to the internet 5. and cost of telecommunication infrastructure 6. trust and confidence in e-government 7. encouragement of citizens to use and 8. participate to e-government, 9. ICT policy 10. Marketing e-government to citizens.
Organization [1], [3]-[6], [8], [10]-[12]	<ol style="list-style-type: none"> 1. Administration reforms 2. internal policy 3. Resistance to change 4. Lack of competencies on organization level 5. collaboration, 6. IT Management 7. Objective and Motivation 8. Consistent evaluation and monitoring.
Strategic [1], [3]-[5], [11]-[15], [19], [20]	<ol style="list-style-type: none"> 1. Overall vision and mission 2. strategic framework, 3. strategic information management (SIM) 4. ICT strategy 5. Objective and goal 6. Principle 7. Focus area.
Cultural [1]-[6], [11]-[18], [22]	<ol style="list-style-type: none"> 1. Awareness 2. Internet experience, 3. IT literacy 4. Social influence. 5. ,Education, 6. Genders 7. Citizens expect.

TABLE II. E-GOVERNMENT CHALLENGES CATEGORIES ADDRESSED BY RESEARCH

Researchers	Challenges				
	Technical	Organization	Cultural	adoption	Strategic
Al-Khoury [20]	√	√		√	√
Alshehri, M [1]	√	√	√	√	√
Al-Shafi, S. and V. Weerakkody [2]	√		√	√	
AL-Naimat, A.M., M.S. Abdullah, and M.K [3]	√	√	√	√	√
Elkadi, H.	√	√	√	√	√

Abdelmoniem, E.M., S.A. Mazen, and E.E. Hassanein [4], [5]					
Al-Hagery, M.A.H. Alsohybe, N.T. Al-Wazir, A.A. and Z. Zheng. Al-Wazir, A.A. and Z. Zhen. [6], [7]-[10]	√	√	√	√	
Halligan, J. and T. [11]	√	√	√	√	√
Ramli, R.M.[12]	√	√	√	√	√
Mutula, S.M. and J. Mostert [13]	√		√	√	√
Rijadi, D.A. and E. Satriya [14]	√	√	√	√	√
Nkwe, N.[15]	√	√	√	√	√
Kumar, R. and M.L. Monga, A [16], [17]	√	√	√	√	
Mundy, D. and B. Musa [18]	√		√	√	
Chowdhury, H., M. Habib [19]	√		√	√	√
Ebrahim, Z. and Z [24]	√	√		Financial	
Hwang [28]	√	√			Legal

TABLE III. EXISTING FRAMEWORKS COVERAGE OF CHALLENGE CATEGORIES

Frameworks/challenges	Technical	Organizational	Cultural	Adoption	Strategy
RoslindKaur, 2006 “Malaysia”	Cover only: -IT infrastructure. - Security and privacy. - Data and system integration.	Cover only: -Collaboration.	-	Cover only: -Resources (access devices, hardware/software)	
Mundy, D. and B. Musa, 2010 “Nigeria”	Cover only: -Network infrastructure -IT management.		-	Cover only: -Resources (access devices, hardware/software)	
Zakareya Ebrahim, 2005	Cover only: -IT infrastructure such as LAN, servers. -Data and system integration. -Data Management		-	Cover only: -Resources (access devices, hardware/software)	
Sharma, 2003	Cover only: -IT infrastructure such as LAN, servers. -Data and system integration -Data Management		-	Cover only: -Resources (access devices, hardware/software)	
Hatem Ben Sta, 2014 “Tunisia”	Cover only: -IT infrastructure	Cover only: -Collaboration.	-	Cover only: -Resources (access devices, hardware/software) -Funding	
Harijadi, D.A. and E. Satriya, 2000 “Indonesia”	Cover only: -IT infrastructure	Cover only: -Administrator reforms.	-		
Kütt, A.andJ. Priisalu, 2014 “Estonia”	Cover only: -IT infrastructure -Data and system integration		-	Cover only: -Resources (access devices)	
Rashty, B.C.a.D, 2002 “Finance General AccountantOffice-israel”	Cover only: -IT infrastructure. - Security and privacy.	Cover only: -Collaboration.	-	Cover only: -Resources (access devices).	
Government, 2007 “Australian”	Cover only: -IT infrastructure. -IT Security	Cover only: -Re-engineering	-		
Abdelkader, 2006, New Zeland	Cover only: -IT infrastructure.	Cover only: - Re-engineering		Cover only: -Resources (access devices)	

2) E-government layer

This layer focus on integration of different organization data and services into one stop called web portal. E-government framework includes e-government layer, according to (Harijadi, D.A. and E. Satriya, 2000), (RoslindKaur, 2006), (Sharma, 2003), and (Zakareya Ebrahim, 2005).

3) Data layer

This layer contains integration database from different organizations government that use to support decision making. According to (RoslindKaur, 2006), (Sharma, 2003), and (Zakareya Ebrahim, 2005), e-government framework includes data layer.

4) Application layer and information layer

This layer contains e-government application such as ERP, and knowledge share information between organizations.

According to Zakareya Ebrahim, 2005; RoslindKaur, 2006 and Hatem Ben Sta, 2014 e-government include application and information layer.

B. Covered Challenges

The review of existing e-government frameworks coverage of the challenges identified in Section 2, each framework focused on specific issues and dropped others. Table 3 summarizes the challenges covered by each of these frameworks. From the table, it can be realized that the IT infrastructure challenges are covered in all frameworks and that the adoption and organizational challenges are dealt with at different levels in most of the frameworks.

On the other hand, the strategy and cultural challenges were not covered in the reviewed frameworks. So, it has been stipulated that corresponding layers/components needs to be introduced into our proposed e-Government framework to deal with the shortcomings emanating from neglecting the related challenges widely revealed during actual implementations.

IV. PROPOSED FRAMEWORK

In this section, the proposed framework for e-government is presented and discussed. The first layer represents access layer that includes government users and channels of access. By using these channels, the e-government's web portal integrates all data, information, and services from several departments that are protected by authentication layer which represents e-government layer. The e-government layer connects to organization layer that manipulates and integrates data, process and applications within the organization body to make information and services available to e-government portal and provide effective and efficient government services. In the bottom layer, the national infrastructure layer (New) was introduced to reach out to all government ministries. The National Infrastructure layer includes technical, legislative and regulatory aspects necessary for the proper function of e-Government (e.g. Law of Access to Information, electronic payment and banking, eID and signature). All layers connect to the strategy layer (New) responsible for the formulation of the national strategy and blueprint of e-Government. All layers also connect to socio-cultural layer (New), to account for the social and cultural aspects and specificities of the users and

staff, including awareness and readiness. A National Execution layer (New) is introduced to coordinate e-government projects implementations across organizations and ensure its abidance to the national strategy. A corresponding chief information officer function was introduced to the organization layer to ensure this coordination. In this framework the following additional layers have been introduced: national infrastructure layer, national strategy, socio-cultural layer, and national executive body. The organization layer was amended with additional roles/functions (business process, organization chief information officer and Decision Support) as shown in Fig. 1. Each layer of the proposed framework will be discussed in the following section.

A. Access Layer

This layer involves the channels that users can access the various government services. This layer has two components which are end user and communication channel that will be discussed as follows:

1) End user

This layer identifies the e-government user categories: citizens, government employees, businesses, other government department and another community member such civil society organizations.

2) Communication channel

Government user can access various services through multiple communication channels (e.g. website, Mobile phone). This layer helps identify the access standards and technologies to information and services for each government user group across different channels.

B. E-Government Layer

This layer is about integrating data from various organizations into a web-portal of government services; in to one-stop e-government port.

This layer has three components which are e-government interface (portal), authentication layer and a Service Oriented Architecture enterprise Government services bus (ESB) according to [23], [29].

1) E-government interface (web portal)

This component focuses on integrating the websites from different organizations in one website called e-government web portal. This component allows user to obtain information or services through a single window, improving access to services, reducing waiting time, saving cost and improving the quality of services (Ho, 2002; Gant and Gant, 2001; Sharma and Gupta, 2002).

2) Authentication layer "portal authentication layer"

Authentication is a process used for several methods to identify government users that can allow them to access system and information, once the users have authenticated [29]; they can be able to use applications that have privileges to use them. Additional multiple authentication layers may be added for extra protection in the government environment.

3) Services oriented architecture (SOA) government enterprise services bus (ESB)

E-government requires collaboration between government organizations and non-government organizations through using various systems. These systems use different data format, language, storage type and technologies thus issues of heterogeneity and interoperability of systems, like Jordan [30], thus, an integrated platforms are needed to enhance sharing of information and services between government organizations and non-government organizations. So, SOA can be used meeting these challenges [31]. ESB is an enterprise application platform that helps governments to develop open architecture, standards-based on integration solution and implementation (SOA).

C. Organization Layer

The organization layer covers the organizational infrastructure, data and information, business processes, applications and information management that coordinates with the national e-government executive body. One of the most important purposes of this layer is to increase efficiency and effectiveness. The national e-government executive body, one of the most purposes of this layer is to increase efficiency and effectiveness. According to survey of e-government frameworks weren't covered many challenges such as evaluation and monitoring challenge (in organizational challenges) and project management challenge (in technical challenges), so we propose organization (ministry) layer to overcome of these challenges. The organization layer is required to relate to the national strategy, national infrastructure and social cultural layer. The government services can generally be distributed between different government organizations. Organizational adjustments are required for the adoption of ICT and inter-departmental coordination [32]. In addition to ICT adoption, two main types of challenges exist: regulatory and internal resistance to change. With the implementations of e-government projects changes to organizational culture, legislation, policies, human resource and organizational structure [32]-[36] have to be performed.

1) Organization chief information officer

According to [4], [5], [19] many countries are suffering from uncontrolled the execution of the e-government projects, so these layers are responsible to control the executions of e-government project in each organization (ministry). According to (Seligman, 1999) OCIO provides leadership for organization overall information technology (IT), IT Architecture, change management, determine priority and strategy of implementation.

2) Business process layer

Business process layer aims at mapping existing and updating processes as well as managing them [37].

According to Taylor et al., 1911; Deming et al., 1982 and Juran et al., 1988, the first step in gaining control over the organizations is to know and understand the basic processes. According to our survey that has been conducted, we found that there are many organizations in some countries which are

still based on paper works [1], [4]-[10]. Therefore, these governments need to transform process from papers to computerize.

3) Application layer

The application layer includes the legacy systems that need to be integrated into online services delivery, new online systems, back-office systems, messaging and directory services [38], [39] as well as Decision support system (DSS) that are required to be integrated with the web portal. According to the conducted survey, there are several problems in decision taking in many governments [40]-[42]. Accordingly, a DSS framework for e-government has been proposed as well. The DSS framework contains six components (processes), data collection, data mining and OLAP, information processing, government knowledge, inference engine and information visualization.

4) Data layer and information layer

The ministries and organizations have separated database. This creates an obstacle to data exchange and service integration. Standardization of data and information formats, metadata, and data dictionaries are the elements of this layer [43], [44].

This layer includes citizens and employees' data and profiles, government data and data warehouses.

5) Infrastructure layer (local)

This layer consists of local technical infrastructure, infrastructure policy and infrastructure management, such as storage backup, and accessibility.

D. National E-government strategy layer

According to Heeks (2006), e-government strategy is defined as plan and guide how to transfer the government to response to e-government challenges. It is required to achieve organizational objectives. According to Abdelbaset, 2009 [27] the e-government strategy contains 'vision, objective, principles, focus area, building block, prioritized initiatives and implementation plan'. The e-government strategy was not included in the reviewed frameworks. From [36], [45], [46] it could be determined that major elements needs to be included in any government strategy:

1) Increasing public ICT awareness.

2) Formulation of a clear vision for prioritized e-government implementation.

3) Assess the success and failure factors before engaging in the implementation of a national e-government project.

4) Identify the regulations and laws required for the secure exchange of information.

5) Develop e-government standards including data standards, technical standards, application standards, business process standards, and privacy and security standards.

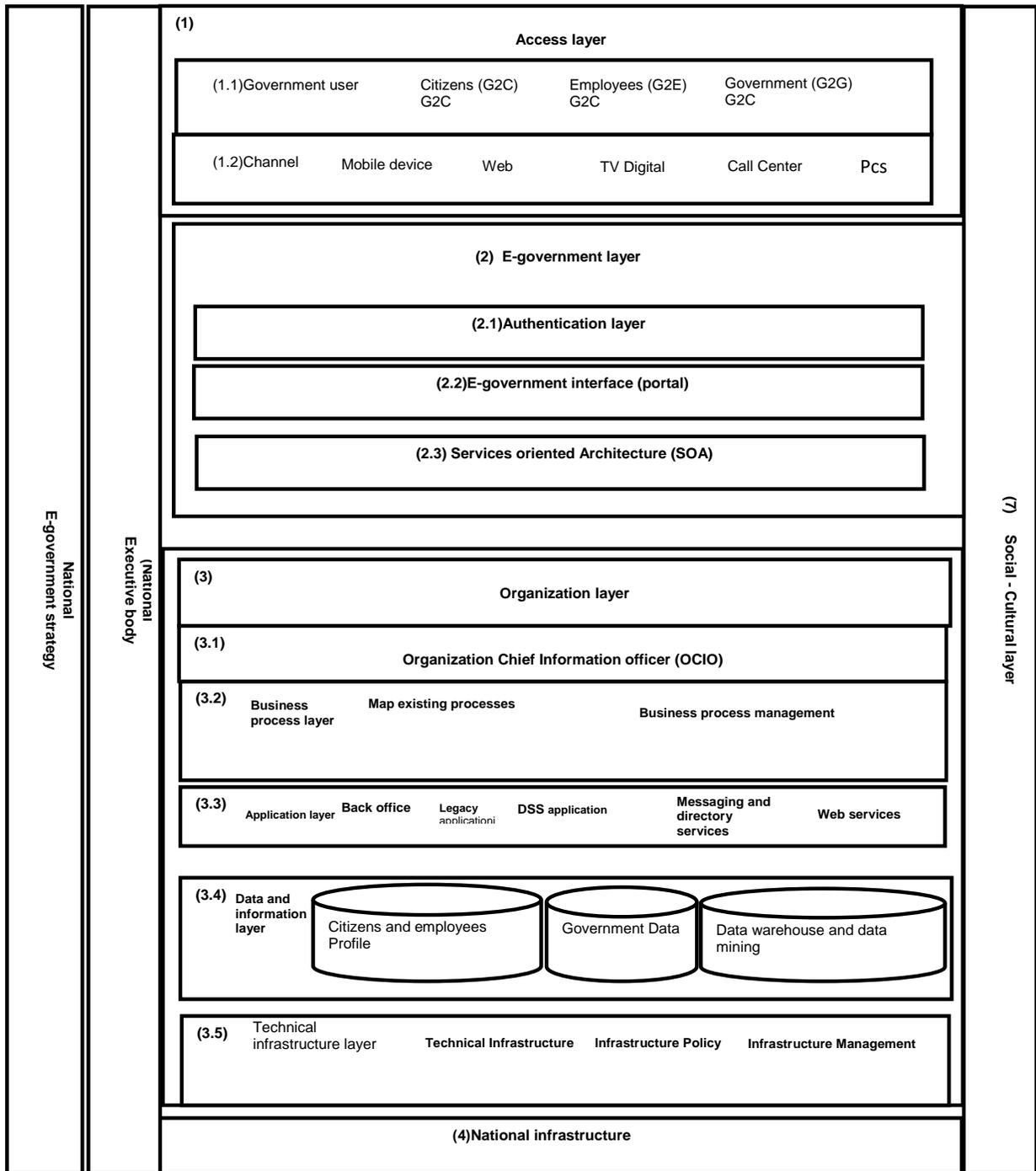


Fig. 1. Proposed e-government framework.

TABLE IV. COMPARISON BETWEEN PREVIOUS AND THE PROPOSED E-GOVERNMENT FRAMEWORK

Comparison	Type of Framework: case study or general (researcher)	Framework layers														
		Infrastructure	Data store	Application	Data layer	Information layer	Legal framework	Web portal	User access layer	E-business	Busines process	OCIO	National executive body	Strategy	Web service integration	Social - Cultural
Actual implementation FK	Malaysia [23]	√	√	√	√			√	√							
	Tunisia [47]	√		√			√		√							
	Indonesia [14]	√					√	√								
	Nigeria [18]								√							
	Estonia [48]	√	√	√												
	Israel [49]	√		√				√	√							
	New Zealand [50]	√		√	√	√			√		√					
	Australia [51]	√						√	√							√
Theoretical FK	Zakareya[24]	√			√	√		√	√	√						
	Sharma [25]	√						√	√							
	Y.N[27]	√														
Proposed Framework		√	√	√	√	√	√	√	√	√	√	√	√	√	√	√

E. National Executive Body (NEB)

Top management support challenge (in adoption challenge) not covered in e-government frameworks that have been reviewed, so this layer is added to proposed framework. .NEB is the administrator of e-government project as the whole. It is linked to the president of council of Ministers to ensure access support. NEB is responsible for ensuring organization interoperability, data and information sharing, maintain information security and privacy controls across the ministries.

F. Social Cultural Layer

From Table 3, it can be noticed absence of cultural challenges from e-government frameworks, so this layer is added to our framework. This layer aims to evaluate readiness of society to use e-government such as determine percent of computer literacy and readiness local and national infrastructure to provide government services. This layer treats the identified lack of readiness and awareness in developing countries and cultural obstacles such as discrimination between male and female, IT literacy and education. This layer impacts on access layer and is influenced by organization layer, national infrastructure and National strategy.

G. National infrastructure layer

National infrastructure includes the essential elements of communication infrastructure such as systems, processes and net-work. The purpose of this layer is to increase availability, and integration of services through the internet. It consists of the national network and national policy. National infrastructure is influenced by national strategy and impacts on organization layer and social cultural layer.

V. COMPARISON TO OTHER FRAMEWORKS

In this section, the proposed framework is compared with previous e-government frameworks as summarized in Table 4, the listed frameworks are categorized as theoretical (academic) and country-specific implementation frameworks. It can be clearly noted that the strategy, the CIO and the national executive body (management), as well as the Social-Cultural layers (dimensions) were not addressed by any of the reviewed frameworks, whereas the proposed model addresses them as well as all the other layers in other frameworks.

VI. CONCLUSION AND FUTURE WORK

This paper is presented an identification of the different

challenges facing e-government implementations. A review of existing e-government frameworks are presented, their structures and the challenges they handle. A set of challenges unhandled by exiting frameworks are identified and introduced corresponding layers to handle them.

There are some common layers between most e-government frameworks such as application layer, web portal layer, user interface, and infrastructure layer. In this framework the following layers have been introduced: national infrastructure layer, national strategy, socio-cultural layer, and national executive body. The organization layer was amended with additional roles/functions (business process, organization chief information officer and decision support).

This framework has been divided into seven layers which are:

- 1) Access layer.
- 2) E-government layer.
- 3) Organization layer.
- 4) National E-government strategy layer.
- 5) National infrastructure.
- 6) National executive body.
- 7) Socio-Cultural layer.

Eventually, this framework has been evaluated by comparing it with previous frameworks: theoretical and implementation framework. It has been found; the proposed framework includes all layers in previous frameworks as well as has additional layer.

Similar to other frameworks, the current one does not either provide for an execution plan nor for an implementation structure to guarantee success.

Future work would cover the formulation of a high level implementation plan and organizational structure capable of implementing the foreseen plan.

Also, the authors foresee the application of the proposed model to existing e-government implementations, which may reveal the implicit existence of some of the proposed additional layers, while not explicitly represented in the reported frameworks.

REFERENCES

- [1] Alshehri, M. and S. Drew, Challenges of e-government services adoption in Saudi Arabia from an e-ready citizen perspective. *Education*, 2010. 29(5.1).
- [2] Al-Shafi, S. and V. Weerakkody, Factors affecting e-government adoption in the state of Qatar. 2010.
- [3] AL-Naimat, A.M., M.S. Abdullah, and M.K. Ahmad. The Critical Success Factors for E-Government Implementation in Jordan. in *Proceedings of the 4th International Conference on Computing and Informatics*, University of Utara, Malaysia. 2013.
- [4] Elkadi, H., Success and failure factors for e-government projects: A case from Egypt. *Egyptian Informatics Journal*, 2013. 14(2): p. 165-173.
- [5] Abdelmoniem, E.M., S.A. Mazen, and E.E. Hassanein, Governance of Post-Construction Activities in IS Development Projects. *International Journal of Computer Science Issues(IJCSI)*, 2012. 9(5).
- [6] Ramli, R.M., Malaysian e-government: issues and challenges in public administration. *International Proceedings of Economic Development and Research* 2012. 22: p. 61-74.
- [7] Al-Hagery, M.A.H., Basic Criteria for the Purpose of Applying E-Government in the Republic of Yemen. *International Journal of Research & Reviews in Computer Science*, 2010. 1(3).
- [8] Alsohybe, N.T., The implementation of e-government in the Republic of Yemen: An empirical evaluation of the technical and organizational readiness. 2007.
- [9] Al-Wazir, A.A. and Z. Zheng, Factors Influencing E-government Implementation in Least Developed Countries: A Case Study of Yemen. *Developing Country Studies*, 2014. 4(7): p. 20-29.
- [10] Al-Wazir, A.A. and Z. Zheng, E-government development in Yemen: assessment and solutions. *J Emerg Trends Comput Inf Sci*, 2012. 3(4): p. 512-518.
- [11] Halligan, J. and T. Moore, E-government in Australia: The challenges of moving to integrated services. 2004.
- [12] Ramli, R.M., Malaysian e-government: issues and challenges in public administration. *International Proceedings of Economic Development and Research*, 2012. 48(2): p. 19-23.
- [13] Mutula, S.M. and J. Mostert, Challenges and opportunities of e-government in South Africa. *The Electronic Library*, 2010. 28(1): p. 38-53.
- [14] Harijadi, D.A. and E. Satriya. Indonesia's Road Map to e-Government: Opportunities and Challenges. in *APEC high-level symposium on e-government*. 2000.
- [15] Nkwe, N., E-government: challenges and opportunities in Botswana. *International journal of humanities and social science*, 2012. 2(17): p. 39-48.
- [16] Kumar, R. and M.L. Best, Impact and sustainability of e-government services in developing countries: Lessons learned from Tamil Nadu, India. *The Information Society*, 2006. 22(1): p. 1-12.
- [17] Monga, A., E-government in India: Opportunities and challenges. *JOAAG*, 2008. 3(2): p. 52-61.
- [18] Mundy, D. and B. Musa, Towards a framework for e-government development in Nigeria. *Electronic Journal of E-government*, 2010. 8(2): p. 148-161.
- [19] Chowdhury, H., M. Habib, and I. Kushchu. Success and failure factors for e-Government projects implementation in developing countries: A study on the perception of government officials of bangladesh. in *The Proceedings of Euro mGov*. 2006.
- [20] Al-Khoury, A.M., eGovernment strategies the case of the United Arab Emirates (UAE). *European Journal of ePractice*, 2012. 17: p. 126-150.
- [21] Alraimi, K.M., Towards a Sustainable e-Government Infrastructure Initiative: The Case of Yemen. *School of Engineering Korea Advanced institute of Science and Technology*, 2009.
- [22] Al-hashmi, A. and S. Suresha, Evaluating the Awareness of E-government in the Republic of Yemen. *International Journal of Computer Applications*, 2013. 67(16): p. 41-45.
- [23] Kaur, R., Malaysian e-government implementation framework, 2006, University of Malaya.
- [24] Ebrahim, Z. and Z. Irani, E-government adoption: architecture and barriers. *Business process management journal*, 2005. 11(5): p. 589-611.
- [25] Sharma, S.K. and J.N. Gupta, Building blocks of an e-government: A framework. *Journal of Electronic Commerce in Organizations (JECO)*, 2003. 1(4): p. 34-48.
- [26] Sta, H.B., Malaysian e-government: issues and challenges in public administration 2014.
- [27] Chen, Y., et al., E-government strategies in developed and developing countries: An implementation framework and case study. *Journal of Global Information Management*, 2006. 14(1): p. 23.
- [28] Hwang, M.-S., et al., Challenges in e-government and security of information. *Information & Security*, 2004. 15(1): p. 9-20.
- [29] Yan, M. and F. Zhi-hua, Research on Web/Portal Authentication Technology [J]. *Microelectronics & Computer*, 2004. 8: p. 021.
- [30] Saleh, Z.I., R.A. Obeidat, and Y. Khamayseh, A Framework for an E-government Based on Service Oriented Architecture for Jordan. *International Journal of Information Engineering and Electronic Business*, 2013. 5(3): p. 1.

- [31] Keen, M., et al., Patterns: Implementing an SOA using an enterprise service bus. IBM Redbooks, 2004. 336.
- [32] Nograšek, J., Change management as a critical success factor in e-government implementation. Business Systems Research, 2011. 2(2): p. 13-24.
- [33] Ojo, A., et al., Human Capacity Development for e- Government. UN University International Institute for Software Technology, Macau, UNU-IIST Report, 2007(362).
- [34] Kanungo, S. and V. Jain, Organizational culture and e-government performance: An empirical study. E-Government Services Design, Adoption, and Evaluation, 2012: p. 141.
- [35] Almutairi, N., The Impact of Organizational Culture on the Adoption of E-Management" Evidence from Public Authority for Applied Education and Training (PAAET) in Kuwait". International Journal of Business and Management, 2014. 9(9): p. 57.
- [36] Evans, D. and D.C. Yen, E-government: An analysis for implementation: Framework for understanding cultural and social impact. Government Information Quarterly, 2005. 22(3): p. 354-373.
- [37] Röglinger, M., J. Pöppelbuß, and J. Becker, Maturity models in business process management. Business Process Management Journal, 2012. 18(2): p. 328-346.
- [38] Markellou, P., A. Panayiotaki, and A. Tsakalidis. E-Government and Applications Levels. in Proceedings of IADIS Conference e-Society, Portugal. 2003. Citeseer.
- [39] Cavanaugh, E., Web services: Benefits, challenges, and a unique, visual development solution. white paper, Feb, 2006. 10.
- [40] Riad, A., et al. Effective and Secure DSS for E-Government. in WSEAS International Conference. Proceedings. Recent Advances in Computer Engineering Series. 2012. WSEAS.
- [41] Rao, G.K. and S. Dey, Decision support for e-governance: a text mining approach. arXiv preprint arXiv:1108.6198, 2011.
- [42] Van der Aalst, W.M., Using Process Mining to Bridge the Gap between BI and BPM. IEEE Computer, 2011. 44(12): p. 77-80.
- [43] Alasem, A., An overview of e-government metadata standards and initiatives based on Dublin Core. Electronic Journal of e-Government, 2009. 7(1): p. 1-10.
- [44] Weibel, S., et al., Dublin core metadata for resource discovery, 1998.
- [45] West, D.M., E-government and the transformation of service delivery and citizen attitudes. Public administration review, 2004. 64(1): p. 15-27.
- [46] Rabaiah, A. and E. Vandijct, A Strategic Framework of e-Government: Generic and best Practice". Leading Issues in e-Government Research, Academic Publishing International Ltd, 2011: p. 1-32.
- [47] Bouchnak, Tunisia's e-Government Experience. 2013.
- [48] Kütt, A. and J. Priisalu. Framework of e-government technical infrastructure. Case of Estonia. in Proceedings of the International Conference on e-Learning, e-Business, Enterprise Information Systems, and e-Government (EEE). 2014. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).
- [49] Rashty, B.C.a.D., The Five Layers Model of e-Government. Ministry of Finance General Accountant Office, 2002. 23(2): p. 207-235.
- [50] Abdelkader, A., A cooperative intelligent decision support system for contingency management. Journal of Computer Science, 2006. 2(10): p. 758-764.
- [51] Government, A., The Australian Government Business Process Interoperability Framework. 2007.

Validating a Novel Conflict Resolution Strategy Selection Method (ConfRSSM) Via Multi-Agent Simulation

Alicia Y.C. Tang

College of Computer Science and Information Technology
Universiti Tenaga Nasional
43000 Kajang Selangor Malaysia

Ghusoon Salim Basheer

College of Graduate Studies
Universiti Tenaga Nasional
43000 Kajang Selangor Malaysia

Abstract—Selecting a suitable conflict resolution strategy when conflicts appear in multi-agent environments is a hard problem. There is a need to develop a method that can select a suitable strategy which guarantees low cost in terms of the number of messages and time ticks. This paper focuses on conflicts over agents' individual opinion and decision making by taking into account an agent's features such as collaborative, autonomous, and local communication. The significance of this research is two-fold. Firstly, this research attempts to prove the significance of giving agents the ability to select an appropriate strategy in different conflict states depending on conflict specifications such as conflict strengths and confidence levels of the conflicting agents. Secondly, the study developed a new method named as ConfRSSM for reducing the communication cost and time taken for selecting a conflict resolution strategy. The approach ignores some conflict states, and replaces complex strategies by a simpler one, in some conflicting cases. Results show ConfRSSM reduces the number of messages and time ticks and thus improving the entire conflict resolution process.

Keywords—Multi-agent, conflict resolution strategy; conflict states; confidence level; simulation

I. INTRODUCTION

In Multi-Agent Systems (MAS), conflicts occur when two agents have dissenting opinions on the same subject [1]. The general model for resolving a conflict is either by avoiding or solving it by using conflict resolution algorithms, or negotiation protocols [2]. In distributed, dynamic and complex environments, conflict resolution is often essential because of computational and communication bottleneck, as a result, conflict resolution is a huge challenge in multi-agent systems, and agents need to resolve conflicts in a distributed manner without global knowledge [2]-[4]. In MAS, conflict considered as a failure or a synchronization problem [4]. Choosing the most appropriate conflict resolving approach ensures proper operation of the multi-agent system. The capability of strategy selection can enhance MAS's flexibility and adaptability to dynamic and uncertain environments [5]. A significant challenge in the research on agent's conflict is the question of how to select an appropriate conflict resolution strategy.

Indeed, there is no one strategy that works best for all situations [6]. Some conflict states can be solved without using complex computational strategies such as negotiation. Strategies such as ignoring, submitting or forcing that need less

computational complexity are sufficient. For this reason, developing methods for choosing among conflict resolution strategies is considered an important matter. Existing work on conflict resolution suffers from the following deficiencies:

- No technique available for detecting the confidence level of conflicting agents that takes in consideration three integrated factors, trust, certainty, and evidences. Some research builds a system that detects evidence depending on past experience [7], while other researches exploit the relation between evidence and certainty [8]. Some of the work associate certainty with the number of collected evidence. Some research evaluates trustworthiness depends on two sources of information: direct trust evidence and third party witness [9], while others build systems that detect evidence depending on the reputation of the agents [10]. It is argued that there is no formal technique for detecting agents' confidence levels that integrates trust, certainty and evidences.
- Researchers did not provide any model to detect conflict strengths and conflict classification. Conflict classification allows for identification and design of different methods for resolving conflict. Some research classified conflicts into two types: Potential conflict and real conflict [11], while others classified conflicts into two main classes: Physical conflicts and knowledge conflicts [2]. There is no model to detect the strength of an agent's conflict.
- Researchers have not discussed the relationship between conflict specifications and conflict resolution strategy selections. They did not provide a method to select a suitable conflict resolution strategy that solves conflicts among agents in all conflict states.
- In learning style detecting field, research only considers learners' responses to a specific questionnaire and detects learning styles from learner's behaviors and actions. These systems do not exploit other information such as the learners' social surrounding to detect learning styles. There is no model for learning style detection that considers the opinions of student's social surrounding.

The paper is organized as follows: Section II provides the background of this work with the research problems and objectives. Section III presents the research methods. Section IV discusses the ConfrSSM simulation. Result discussion is presented in Section V, and Section VI concludes the paper.

II. BACKGROUND

A. Previous Work

Belief-Desire-Intention (BDI) agents typically have various goals they are tracking of simultaneously. In some states, the goals are inconsistent, choices made about how to pursue each of these goals may well result in a set of conflicting intentions. Conflict Resolution (CR) is the fundamental process for coordinated agent attitude. Conflict resolution includes conflict detection that involves searching for solutions, and reaching an agreement through communication among agents [12]. The capability of strategy selection can enhance MAS's flexibility and adaptability to dynamic and uncertain environments. There are several issues that must be addressed to achieve this goal. A uniform representation of a different strategy is for the comparison and evaluation processing. A metal-level reasoning mechanism for strategic decision making, a set of specifications involving requirements for a domain that agents use to evaluate substantial strategies, and the ability for adaptation to improve the decision making required to select a strategy [5].

Adler et al. [13] allowed an agent to select a specific strategy from many other strategies such as priority agreement, negotiation, arbitration, and self-modification. In their work, if there is heavy network traffic, an agent selects the arbitration strategy to resolve conflict, but if there is light traffic, the agent selects negotiation or another strategy. Liu et al. [3] mentioned the importance of allowing agents to select an appropriate conflict resolution strategy based on many factors such as conflict's nature (if there is a conflict in goal, plan or belief), the agent's autonomy level, and the agent's solution preferences.

This research provides the main framework that comprised of Agent Confidence Detection Technique (AgConfDT) that detects agent's confidence levels, and a Conflict Strength Detection Model (CSDM) that detects conflict strengths. This information is used by a Conflict Resolution Strategy Selection Method (ConfrSSM) for selecting a suitable conflict resolution strategy. Then a new model for learning style detection is used for system validation and evaluation. AgConfDT includes an exploration of the three different confidence factors (trust, certainty, and evidences). It emphasizes important objects by integrating these factors in order to better understand the agents' specifications since the technique can detect the agent's confidence in the absence of any required information. Results show that the proposed technique eliminates untested opinions, such that the confidence levels of conflicting agents can be detected in all cases although in the absence of some confidence factors. CSDM detects the disagreement degree among the conflicting agents, a conflict ratio as input for the model, and the output is the conflict strength. In resolving a conflict, ConfrSSM uses

the confidence levels of conflicting agents and a conflict strength to select a suitable strategy.

Finally, we propose a new model for learning style detection. The model detects students' learning styles depending on social surrounding's opinions. The run-time model enables us to evaluate the strategy performance in various computing and networking environments. Simulation results show that the proposed model provides more accurate detection of a student's learning style. This part forms the basis of the discussion of this paper.

In a Learning Management System (LMS), individuals have different learning preferences that help them learn better. These preferences are named learning styles. Many educational theorists and researchers consider learning style as an important factor that affects the learning process. Recently, more attention has given to the use of multi-agent systems in many distributed applications. Studies in multi-agent systems include the inquiry for rational, autonomous and flexible behavior of entities, and their interaction and coordination in different areas [14]. The foundation of multi-agent systems play a significant role in the growth of teaching systems, because the basic issues of teaching and learning could be easily resolved by multi-agent systems [15].

B. Research Problems and Objectives

The objectives of this work are:

- a) To propose an integrated model for detecting agents' confidence levels that considers certainty, trust, and environmental evidence [16].
- b) To propose a model for detecting the conflict's strength that considers the number of conflicting agents and conflicting issues in conflict states [17]-[19].
- c) To propose a new model for detecting learner's learning style that considers learner's social surrounding opinions; for validating the entire framework [20], [21].
- d) To formulate a novel selection strategy method for a conflict resolution in multi agent systems [24].
- e) To validate (d) using agent-based simulation.

Items (d) and (e) are the focus of this paper.

III. RESEARCH METHOD

The research focuses on the formalization of a three frameworks: agent confidence detection, conflict resolution strategy selection, and learning style detection. The confidence detection model starts with identifying three factors that involve agent trust, agent certainty, and an evidence. The findings are used in modeling conflict resolution strategy selection. Learning style detection was selected as the domain for validating the framework. Besides analyzing common dimensions when detecting learning styles, social surrounding opinions to deliberate the detection of learning styles were added to the model. This component solicits information from parents and teachers. Conflicts may occur due to these opinions. The outcome of the study is a conflict resolution strategy selection framework that addresses agent confident, and conflict strength to select the most suitable conflict resolution strategy, as shown in Fig. 1.

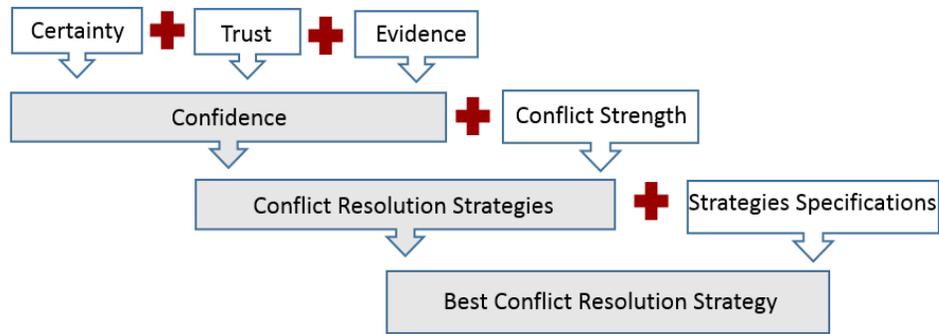


Fig. 1. The proposed conflict resolution strategy selection framework.

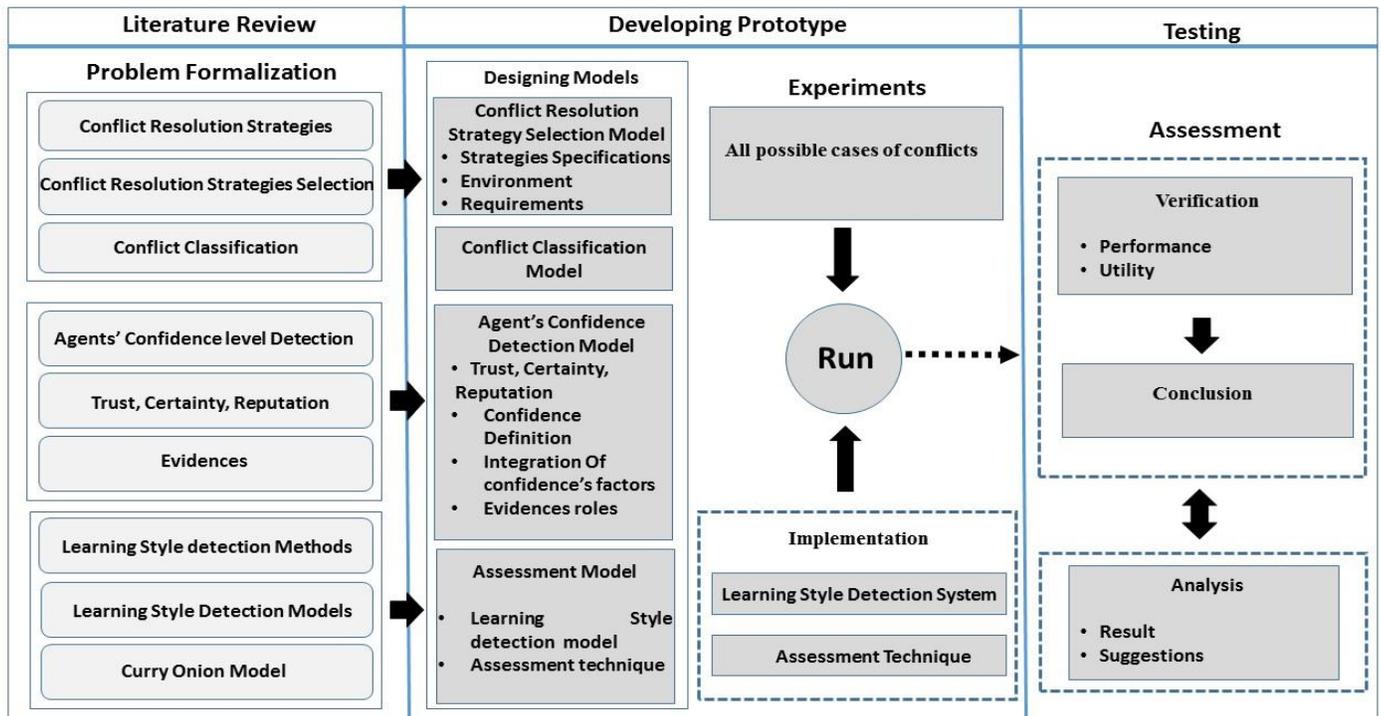


Fig. 2. The entire research framework.

The entire framework is shown in Fig. 2. This paper presents and discusses the prototype development with the simulation results (middle column of Fig. 2). Learning style detection was used as the platform to validate the confidence model and conflict resolution selection strategy method.

A. Implementation Essentials

On the LMS part, most systems only considered learners' responses to a specific questionnaire and detect learning styles from learner's behaviors and actions. These systems do not exploit other information such as the learners' social surrounding to detect learning styles. The proposed method involves collecting four different opinions, three opinions from the student's social surrounding, (parent, teacher, and friend). The fourth opinion is collected from the student agent. VARK model questionnaires [22], [23] that deal with multiple

students' personal activities and behaviors were also distributed.

B. Data Gathering

Data collection involves identifying a group of students and their social surrounding (parents, teachers, and friends). Students are required to attempt the VARK questionnaire, and their social surroundings are required to attempt different questionnaires.

C. Testing

To test the proposed ConfRSSM and AgnConfD models, the visual environments were created. Each visual environment represents a learning style detection scenario, which include agents that cooperate and achieve tasks that involve numerous parameters and settings. The process starts by creating four

agents in Matlab, first agent represents a student, while other three agents represent his/her social surroundings, each of these four agents use different questionnaires for detecting a student's learning style. To simulate agents' confidence level detection, each agent uses the developed questionnaires for detecting a student's learning style. After detection, an evaluation agent will collect all agents' opinions. The evaluation agent detects conflict states and a conflict strength for each state. Based on the AgnConfD technique, the confidence level of an agent is calculated. To simulate ConfrSSM, conflicts appeared in the first simulation were used. From here, conflict strength can be detected.

IV. CONFRSSM SIMULATION

A. Simulation Environment

The simulation is presented as a scenario of agents to select conflict resolution strategy by exploiting the ConfrSSM. The scenario includes four agents (student, parent, friend, and lecturer agents) in a learning management system. Experiments were conducted to explain how conflict states among agents are resolved, and how the conflict resolution strategy is selected based on the confidence level of conflicting agents. Many tests were generated to show the different in the number of messages and time ticks that are needed for resolving many types of conflicts.

Two experiments were conducted. The first experiment attempts to resolve conflicts by using a unique strategy (Negotiation and Arbitration). In the second experiment, the conflict resolutions are equipped with ConfrSSM. Multiple tests were run to explore the effects of environmental setting on the success of conflict resolution by a minimum number of messages and time ticks. The interface window shown in Fig. 3 is used to collect agents' opinions and calculate the conflict strengths among them. The GUI consists of two buttons, the first one collects agents' opinions, while the second button detects conflict strengths. Fig. 4 shows the results of collecting four agents' opinions and the detected conflict strengths.

An interface window (Fig. 5) was created to calculate and display the number of messages and time ticks needed for resolving conflicts among agents by using Negotiation and Arbitration strategies. The GUI consists of three columns: the first one receives the number of conflicting agents, number of proposals, and a CR strategy. The second and third columns display the number of messages and time ticks needed for resolving conflicts using the selected strategy. Fig. 6 receives the confidence values of the evaluation agents, detects and displays a suitable conflict resolution strategy for each conflict state. The GUI consists of five columns: the first one receives the agents' confidence levels from user, while the second column displays the conflict strength for each conflict state. The third column displays the conflict resolution strategy for each conflict state. The fourth and fifth columns display the number of messages and time ticks needed for each conflict resolution operation.

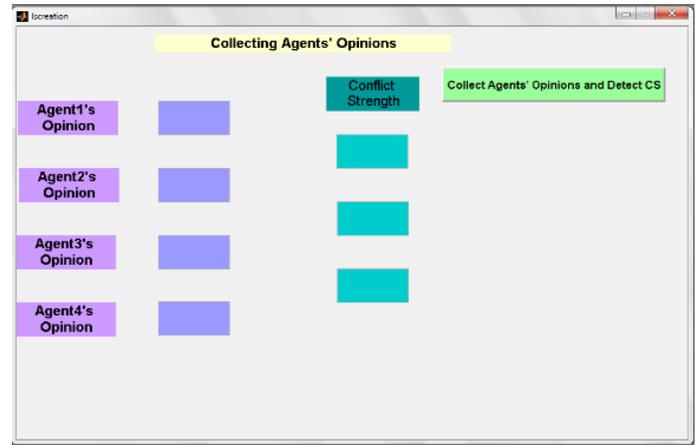


Fig. 3. The interface that collects agents' opinions and detects conflict strength.

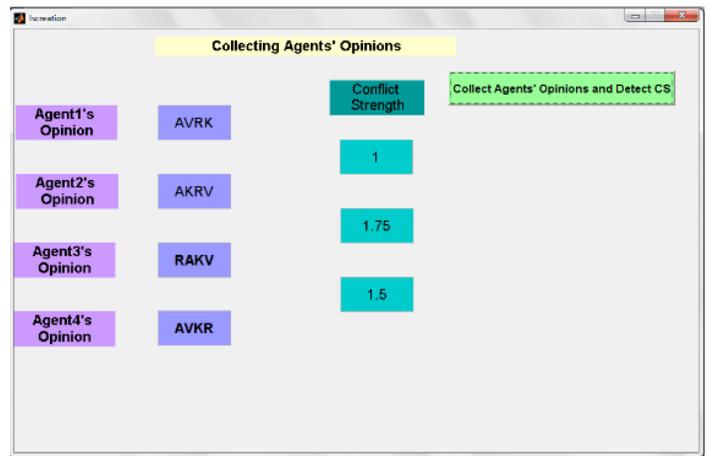


Fig. 4. The interface shows the collected agents' opinions and the detected conflict strength.

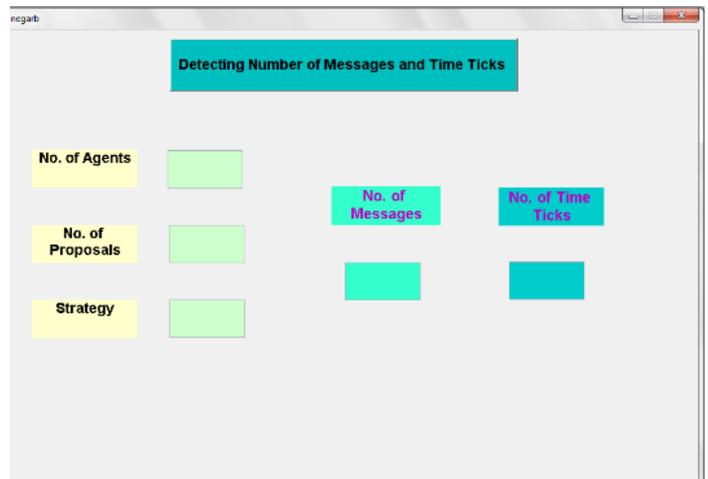


Fig. 5. The interface to calculate the number of messages and time ticks for CR strategies.



Fig. 6. Main simulation interface to run ConfrSSM.

B. Observing the Number of Conflicting Agents

Conflict states include conflicts between two agents, and conflicts among three or four agents. In the learning style detection scenario, four agents are used, a Conflicting Agent Set, CAS, is defined as a set of pairs of conflicting agents, i.e., if a_i conflicts with a_j , then $CAS = \{(a_i, a_j)\}$. Assuming that the four agents have varying levels of confidence, the following cases are apparent:

Case 1: When a conflict occurs between two agents, $(a_i, a_j) \in CAS$, both of them have High Level of Confidence (HLC), $Conf_{a_i} = Conf_{a_j}$.

Case 2: When a conflict occurs between two agents, $(a_i, a_j) \in CAS$, both of them have Low Level of Confidence (LLC), $Conf_{a_i} = Conf_{a_j}$.

Case 3: When a conflict occurs between two agents, $(a_i, a_j) \in CAS$, one of them has High Level of Confidence (HLC) and other have Low Level of Confidence (LLC), $Conf_{a_i} > Conf_{a_j}$ or or $Conf_{a_i} < Conf_{a_j}$.

In the simulation, there is a conflict resolution strategy selection agent a_{SS} that is responsible for the selection of an appropriate conflict resolution strategy in each conflict state.

C. Variables Setting

Each variable is defined as follows:

Conflict Strength: If there is a conflicting agent set (CAS); that conflicts about specific issues I , each conflict state has a strength of conflict, weak conflict or strong conflict. For each pair of conflicting agents $(a_i, a_j) \in CAS$, their conflict strength is represented by CS_{ij} .

Determining the Conflict Strength (CS) among Conflicting Agents: Each agent of the conflicting agents has a specific opinion about a student's learning style. The detected learning style of the student could be VARK, KVAR, and ARKV ... etc., each two conflicting agents are conflicts about the number of style (issues).

Determining the Dissenting Issues: There are issues which serve as "conflicts" among the agents. It is defined as the ratio of the number of dissenting issues to the total number of issues in one conflicting state.

$$D = \frac{\text{Number of Dissenting Issues}}{\text{Total Number of Issues}}$$

Determining the Conflict Ratio: Defined as a ratio of the number of conflicting agents to the total number of agents in a one conflicting set.

$$CR = \frac{\text{Number of Conflicting Agents}}{\text{Total Number of Agents}}$$

Learning Style Generator: This generates learning style for each agent (a_s , a_p , a_f and a_T). The generator uses a random function to produce a learning style (LS). Example: for a_s as KVRA, for a_p as RKVA, and for a_T as VKAR.

The Domain Style (VARK): This defines the patterns of the learning style domain and their four levels:

- High level mode (HLM): The first style in the detected learning style.
- First moderate level mode 1 (MLM1): The second style in the detected learning style.
- Second moderate level mode 2 (MLM2): The third style in the detected learning style.
- Low level mode (LLM): The fourth style in the detected learning style.

Calculating the Dissenting Degree: For each conflict state, the dissenting degree is calculated using the formula: if TI is a number of a total issues in the system, and i is the number of issues that agents are conflicting about it, then,

$$DD = i/TI$$

Calculating the Conflict Ratio: For each conflict state, the conflict ratio is High value if more than 50% of agents in the system conflict with the rest of the agents.

Calculating the Conflict Strength (CS): For each conflict state, the CS is calculated using the formula:

$$CS = \mu CR + \mu DD$$

Calculating the Number of Messages for each Strategy: For each test, the number of messages is calculated after selecting suitable strategies that resolved the conflict. Five strategies are available in the ConfrSSM method: Negotiation, Arbitration, Ignoring, Submitting, and Forcing.

D. Experiments

1) Test Cases without ConfrSSM

Experiment 1: When the conflicts resolution is not equipped with ConfrSSM. Conflict resolution strategies used: Negotiation and Arbitration.

Test 1: This test measures the number of messages that are needed for conflicts resolution among different number of agents, with two, three, four and five proposals, when the conflict resolution strategy is Negotiation.

Fig. 7 shows the number of messages for Negotiation. All conflicts are resolved by a unique strategy (Negotiation), without any consideration to the confidence level of conflicting agents or conflict strength among them.

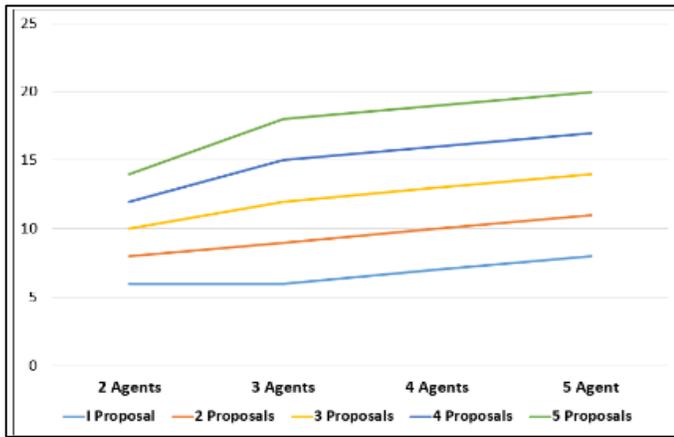


Fig. 7. The number of messages for conflict resolution using negotiation strategy.

Test 2: This test measures the number of messages needed for conflicts resolution among different number of agents, when the conflict resolution strategy is Arbitration.

Fig. 8 shows the number of messages for Arbitration. All conflicts are resolved by a unique strategy (Arbitration), without any consideration to the confidence level of conflicting agents or conflict strength among them.

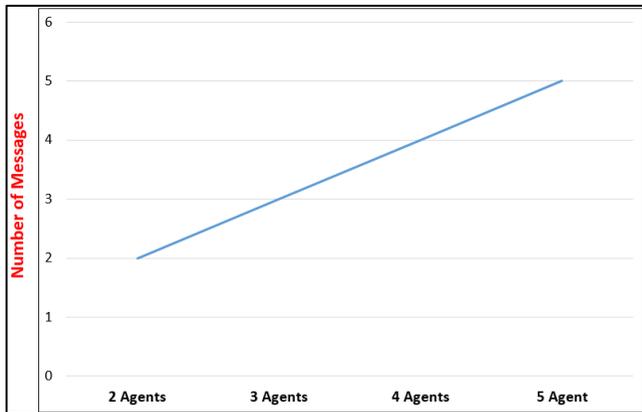


Fig. 8. The number of messages for conflict resolution using arbitration strategy.

To determine how much time is used by the Negotiation and Arbitration strategies, the outcomes on the time taken (in CPU milliseconds) for both strategies, and for each conflict states were plotted. Fig. 9 shows that the number of time ticks in three iterations increases gradually from five ticks in iteration 1 to ten ticks in iteration 2, and fifteen ticks in iteration 3 (Negotiation strategy), and from two ticks in iteration 1, to four ticks in iteration 2, and to six ticks in iteration 3 (Arbitration strategy).

Discussion: Test 1 and Test 2 use the same strategy for all conflicts in all conflict states. Note that the system is unable to detect unimportant conflicts that can be ignored (or can be resolved by other strategies). If there is more than one proposal in Negotiation strategy, the number of messages increases rapidly. Clearly, as the number of proposals increase, the faster the number of messages increases as the number of agent increases.

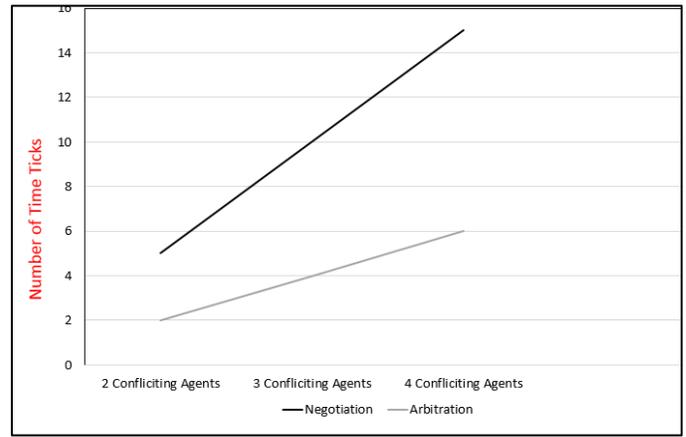


Fig. 9. The number of time ticks for resolving conflicts by using negotiation and arbitration strategies.

In Negotiation strategy, the best case is where the first proposal is accepted. On the other hand, more involved agents require more messages. It is clear that the number of messages and time ticks needed for conflict resolution among agents by using a Negotiation for more one proposal are considered high in comparison with the number of messages needed in Arbitration strategy.

In Arbitration strategy, because the same agent plays both roles, this is the only strategy that does not require inter messages, the number of messages is linear to the number of agents involved.

2) Test Cases Equipped with ConfrSSM

Experiment 2: The experiment presents an analytical model for conflict resolution that is equipped with ConfrSSM, which takes into account conflict strength and confidence level of conflicting agents. Two main factors were considered, conflict strength and confidence level of conflicting agents. Selected cases with simulation results are presented in the following sections:

Case No. 1: Weak Conflict, with five conflicting states:

- When all agents have a low-level confidence.
- When all agents have high-level confidence.
- When 50% of agents have low level confidence and 50% of agents have high-level confidence.
- When 25% of agents have low level confidence and 75% of agents have high-level confidence.
- When 25% of agents have high level confidence and 75% of agents have low-level confidence.

Case No. 2: Strong Conflict, with five conflicting states:

- When all agents have low-level confidence.
- When all agents have high-level confidence.
- When 50% of agents have low level confidence and 50% of agents have high-level confidence.
- When 25% of agents have low level confidence and 75% of agents have high-level confidence.

- When 25% of agents have high level confidence and 55% of agents have low-level confidence.

Test 3: When all conflicting agents have a high confidence level and the conflicts among them are strong. Setting used: CAS: a_S, a_P, a_T, a_F ; CS: Strong; Conf a_S : HCL; Conf a_P : HCL; Conf a_T : HCL; Conf a_F : HCL. Table 1 shows the number of messages and time ticks for resolving the conflicts among four agents.

Test 4: When all conflicting agents have a low confidence level and the conflicts among them are strong. Setting used: CAS: a_S, a_P, a_T, a_F ; CS: Strong; Conf a_S : LCL; Conf a_P : LCL; Conf a_T : LCL; Conf a_F : LCL. Results are tabulated in Table 2.

Test 5: This test detects suitable strategies for resolving strong conflicts when 50% of conflicting agents (a_S, a_P) have a high level of confidence and 50% of agents (a_T, a_F) have a low level of confidence.

TABLE I. NUMBER OF MESSAGES AND TIME TICKS NEEDED FOR RESOLVING THE CONFLICTS SET IN TEST 3

Iteration No.	Iteration 1	Iteration 2	Iteration 3
No. of Conflicting Agents	2	2	2
CR Strategy	Arbitration	Arbitration	Arbitration
No. of Messages	2	2	2
No. of Time Ticks	2	2	2

TABLE II. NUMBER OF MESSAGES AND TIME TICKS THAT ARE NEEDED FOR RESOLVING THE CONFLICTS SET IN TEST 4

Iteration No.	Iteration 1	Iteration 2	Iteration 3
No. of Conflicting Agents	2	2	2
CR Strategy	Negotiation	Negotiation	Negotiation
No. of Messages	11	11	11
No. of Time Ticks	5	5	5

3) Other Tests

This subsection provides other test descriptions:

Test 6: This test detects suitable strategies for resolving strong conflicts when 50% of conflicting agents (a_S, a_F) have a high level of confidence and 50% of agents (a_P, a_T) have a low level of confidence in a sequence of conflict as: a_S, a_P, a_T, a_F .

Test 7: This test detects suitable strategies for resolving strong conflicts when 50% of conflicting agents (a_P, a_T) have a high level of confidence and 50% of agents (a_S, a_F) have a low level of confidence in a sequence of conflict as: a_S, a_P, a_T, a_F .

Test 8: This test detects suitable strategies for resolving strong conflicts when 50% of conflicting agents (a_T, a_F) have a high level of confidence and other two 50% of agents (a_S, a_P) have a low level of confidence in a sequence of conflict: a_S, a_P, a_T, a_F .

Test 9: This test detects suitable strategies for resolving conflicts when 25% of conflicting agents (a_S) have a high level

of confidence and 75% of conflicting agents (a_P, a_T, a_F) have a low level of confidence in a sequence of conflicts: a_S, a_P, a_T, a_F .

Test 10: This test detects suitable strategies for resolving conflicts when 25% of conflicting agents (a_S) have a high level of confidence and 75% of conflicting agents (a_P, a_T, a_F) have a low level of confidence in a sequence of conflict as: a_P, a_T, a_S, a_F .

Test 11: This test detects suitable strategies for resolving conflicts when 25% of conflicting agents (a_S) have a low level of confidence and 75% of conflicting agents (a_P, a_T, a_F) have a high level of confidence in a sequence of conflict as: a_S, a_P, a_T, a_F .

Fig. 10 shows the number of messages and time ticks needed for resolving strong conflicts when 50% of agents have high level of confidence and 50% of agents have low level of confidence. The number of messages are around 6 to 15 and the number of time ticks around 6 to 9. It is considered low when compare to the number of messages and time ticks generated by Test 3.

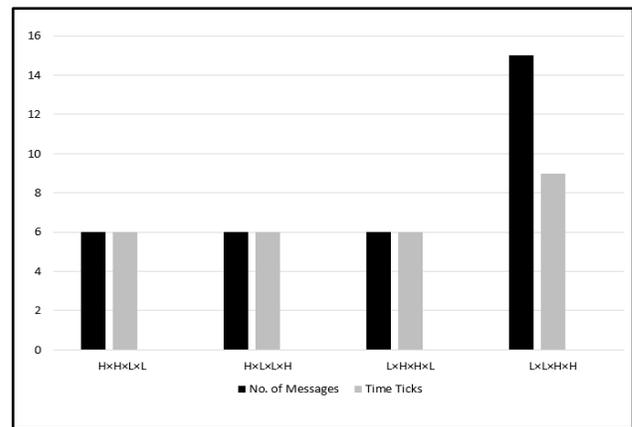


Fig. 10. The number of messages and time ticks for resolving strong conflicts among different sequence of agents when 50% of conflicting agents have high confidence and 50% of conflicting agents have low confidence.

Fig. 11 shows the number of messages and time ticks for resolving weak conflicts among four agents that have an equal confidence level. Note that ConfrSSM ignores weak conflicts among low confidence agents, and the number of messages and time ticks equals zero.

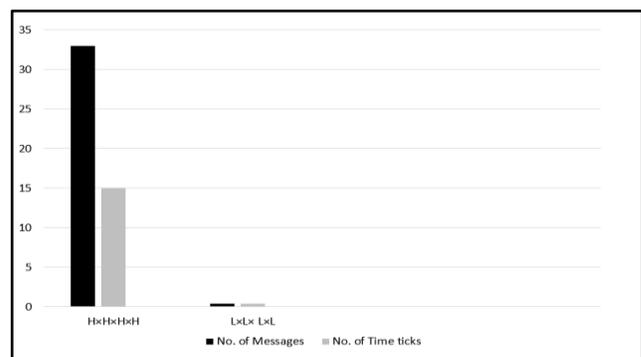


Fig. 11. The number of messages and time ticks needed for resolving conflicts among agents that have the same confidence level when the conflicts are weak.

Fig. 12 shows the number of messages and time ticks needed for resolving conflicts when a 75% of conflicting agents have a low level confidence, and the other 25% of conflicting agents have high confidence level, taking into consideration a multiple conflicts sequences. The result clearly shows that there is a decrease in the number of messages and time ticks in all conflict states.

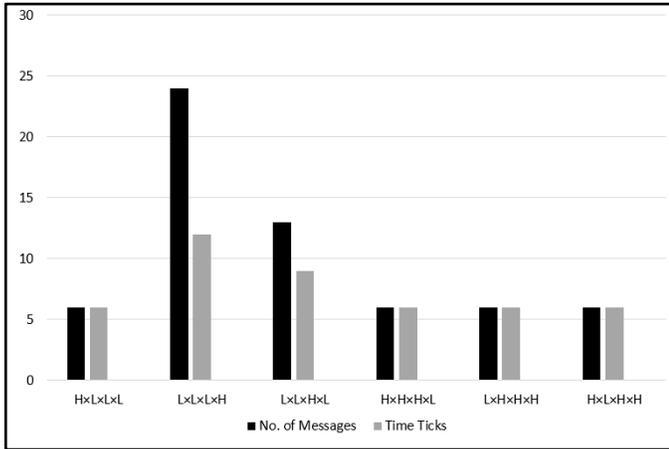


Fig. 12. The number of messages and time ticks for resolving strong conflicts among agents when 75% of conflicting agents have an equal confidence level and 25% have an opposite confidence level.

Test 12: This test detects the strategies for resolving conflicts when 25% of conflicting agents (a_S) have a low level of confidence and 75% of conflicting agents (a_P, a_T, a_F) have a high level of confidence in a sequence of conflict: a_P, a_S, a_T, a_F .

Test 13: This test determines the conflict resolution strategies when all conflicting agents have a high confidence level and the conflict among them is weak.

Test 14: This test determines the conflict resolution strategies when all conflicting agents have a low confidence level and the conflict among them is weak.

Test 15: This test detects suitable strategies for resolving conflicts when two of the conflicting agents (a_S, a_P) have a high level of confidence and other two agents (a_T, a_F) have a low level of confidence and the conflicts are weak in a sequence of conflicts: a_S, a_P, a_T, a_F .

Test 16: This test detects suitable strategies for resolving conflicts when one of conflicting agents (a_S) have a high level of confidence and other three agents (a_P, a_T, a_F) have a low level of confidence and conflicts are weak in a sequence of conflicts: a_S, a_P, a_T, a_F .

Fig. 13 shows the number of messages and time ticks for resolving the weak conflicts when 50% agents have high level of confidence and other 50% of agents have low level of confidence. The number of messages and time ticks are lower because conflict resolution is equipped with ConfrSSM. The number of messages are around 13 to 15 and the number of time ticks around 7 to 9.

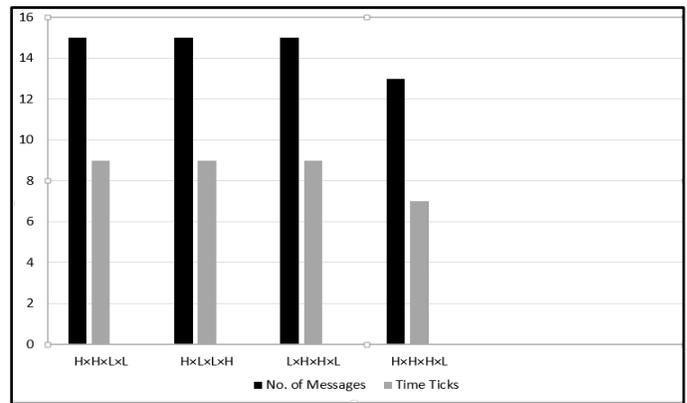


Fig. 13. The number of messages and time ticks for resolving strong conflicts among different sequence of agents when 50% of conflicting agents have high confidence and 50% of conflicting agents have low confidence.

Results of simulation are summarized in Table 3.

TABLE III. SIMULATION RESULTS FOR SELECTED TEST CASES

Confidence Level of Conflicting Agents	Conflict State	Conflict Resolution Strategy	No. of Messages	No. of Time Ticks
HLC, LLC, LLC, HLC	Strong	Forcing, Forcing, Arbitration	6	6
LLC, HLC, HLC, LLC	Strong	Forcing, Arbitration, Forcing	6	6
LLC, LLC, HLC, HLC	Strong	Negotiation, Forcing, Arbitration	15	6
HLC, LLC, LLC, LLC	Strong	Forcing, Forcing, Forcing,	6	6
LLC, LLC, LLC, HLC	Strong	Negotiation, Negotiation, Forcing	24	6
LLC, LLC, HLC, HLC	Strong	Negotiation, Negotiation, Forcing	15	9
LLC, HLC, HLC, HLC	Strong	Forcing, Arbitration, Arbitration	6	9
HLC, HLC, HLC, LLC	Strong	Arbitration, Arbitration, Forcing	6	6
HLC, LLC, HLC, HLC	Strong	Forcing, Arbitration, Arbitration	6	6
HLC, HLC, LLC, LLC	Weak	Negotiation, Negotiation, Negotiation	33	6
LLC, LLC, LLC, LLC	Weak	Ignoring, Ignoring, Ignoring	0	15
HLC, HLC, LLC, LLC	Weak	Negotiation, Submitting, Submitting	15	0

LLC, HLC, HLC, LLC	Weak	Submitting, Negotiation, Submitting	15	9
LLC, LLC, HLC, HLC	Weak	Ignoring, Submitting, Negotiation	13	9
HLC, LLC, LLC, LLC	Weak	Submitting, Submitting, Submitting	6	7
LLC, LLC, LLC, HLC	Weak	Ignoring, Ignoring, Submitting	2	6
LLC, LLC, HLC, HLC	Weak	Ignoring, Submitting, Submitting	4	2
LLC, HLC, HLC, HLC	Weak	Submitting, Negotiation, Negotiation	24	4
HLC, HLC, HLC, LLC	Weak	Negotiation, Negotiation, Submitting	24	12
HLC, LLC, HLC, HLC	Weak	Submitting, Negotiation, Negotiation	24	12

V. RESULT DISCUSSION

The messages and time required in weak conflict cases are low (may reduce to zero) as the result of using Ignoring strategy. This means that weak conflicts are totally ignored when the confidence level of conflicting agents is low (e.g. Test 10). There is a clear decreasing in the number of messages and time ticks for resolving a strong and weak conflicts among multiple conflicts sequence of agents that have a different confidence levels. Also, results show that the number of messages and time ticks for resolving conflicts using Negotiation considered high when agents are using more one proposals. This is critical for multi-agent systems. The high messages in Test 3; Tests 5-6; Tests 9-10; and Tests 15-16 are due to the application of Negotiation strategy. The strategy needs high message number to process as compared to other conflict resolution strategies. There is an obvious decrease in the number of messages and time ticks for resolving a strong and weak conflicts among multiple conflict sequence among agents that have a different confidence level. The number of messages are high (around 9 to 24) in all conflict states that contains conflict between two high confidence agents. While in conflict states that includes two low confidence conflicting agents, the number of messages low (around 2 to 6).

VI. CONCLUSION

Conflicts are likely to be the most critical parameter manifested through agent communication in a distributed multi-agent system. One of the most difficult aspects of the current interest in agent system is selecting an appropriate conflict resolution strategy. Classifying conflict states facilitate the selection of an optimal strategy to resolve conflicts in every conflict situation. Since there is no better strategy suitable for all conflict situations, agent-based systems would benefit from the multiple resolution strategies to resolve unanticipated conflicts. This research attempts to prove the significance of giving software agents the ability to select an appropriate strategy in different conflict states depending on the conflict strengths and confidence levels of the conflicting agents. We presented a novel method to guide strategic decision-making for conflict resolution, and adopted four basic strategies (i.e. Negotiation, Arbitration, Ignoring, and Submitting). In the simulation part, various scenarios were tested with different conflicts among four agents running with the proposed ConFRSSM framework. As expected, using Ignoring, Forcing and Submitting strategies enhanced the conflict resolution performance by decreasing the number of messages and time ticks. Results show ConFRSSM reduces the number of messages and time ticks and thus improving the conflict

resolution process. Further analysis shows that some unimportant conflict states can be ignored, which increases the efficiency of the entire conflict resolution process.

REFERENCES

- [1] Nguyen, N. T. (2002). Consensus System for Solving Conflicts in Distributed Systems. *Journal of Information Sciences*, 147(1-4), 91-122.
- [2] Tessier, C., Chaudron, L., Muller, H.J. (2000). *Conflict Agents, Conflict Management in Multi Agent System*, 1. Springer, Heidelberg.
- [3] Liu, T. H., Goel, A., Martin, C. E. & Barber, K. S. (1998). Classification and Representation of Conflict in Multi-agents Systems. Technical Report TR98-UT-LIPSAGENTS-01, The Laboratory for Intelligent Processes and Systems, University of Texas at Austin.
- [4] M'uller, H. J. & Dieng, R. (2000). *Computational Conflicts- Conflict Modeling for Distributed Artificial Intelligent Systems*. Springer Verlag Publishers.
- [5] Barber K. S., Kim J., (2001). Belief Revision Process Based on Trust: Simulation Experiments, In *Proceedings of Autonomous Agents Workshop on Deception, Fraud, and Trust in Agent Societies*.
- [6] Decker, K. S., & Lesser, V. R. (1995). *Environment Centered Analysis and Design of Coordination Mechanisms* (Doctoral dissertation, University of Massachusetts at Amherst).
- [7] Wang, Y., Mellon, C. & Singh, M. P. (2010). Evidence-Based Trust A Mathematical Model Geared for Multiagent Systems, *ACM Transaction on Autonomous and Adaptive Systems (TAAS)*, 5(4), (pp.1-28).
- [8] Yu, B. & Singh, M. P. (2002). Distributed Reputation Management for Electronic Commerce. *Computational Intelligence*, 18(4), (pp. 5-549).
- [9] Yu, Han, Zhiqi Shen, C. Y. R. I. L. Leung, Chunyan Miao & VICTOR R. Lesser. (2013). *A Survey of Multi-agent Trust Management Systems*. Access, IEEE 1, (pp. 35-50).
- [10] Sen, S. & Sajja, N. (2002). Robustness of Reputation-based Trust: Boolean Case. *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems*, (pp. 288-293), ACM Press.
- [11] Wagner T., Shapiro, J., Xuan, P., Lesser, V., (2007). Multi-level Conflict in Multi-Agent Systems. *Proceeding of AAAI Workshop on Negotiation in Multi-Agent Systems*.
- [12] Alshabi, W., Ramaswamy, S., Itmi, M., & Abdulrab, H. (2007). Coordination, Cooperation and Conflict Resolution in Multi-agent Systems. In *Innovations and Advanced Techniques in Computer and Information Sciences and Engineering* (pp. 495-500). Springer Netherlands.
- [13] Adler, M., Durfee, E., Huhns, M., Punch, W., & Simoudis, E. (1992). AAAI workshop on cooperation among heterogeneous intelligent agents. *AI magazine*, 13(2), 39.
- [14] Alonso, E., d'Inverno, M., Kudenko, D., Luck, M. & Noble, J. (2001). Learning in Multi-agent Systems. Result of a Panel Discussion, In: *Third Workshop of the UK's Special Interest Group on Multi-agent Systems*.
- [15] Boff, E., Vicari, R. M., & Fagundes, M. S. (2008). Using a Probabilistic Agent to Support Learning in Small Groups. In *The 22nd European Conference on Modeling and Simulation, ECMS*.
- [16] Ghosoon Salim Basheer, Mohd Sharifuddin Ahmad, Alicia Y.C. Tang, Sabine Graf, "Certainty, Trust and Evidence: Towards an Integrative

- Model of Confidence in Multi-agent Systems”, *Computers in Human Behavior*, Volume 45, Elsevier, April 2015, Pages 307–315, ISSN: 0747-5632, (Impact Factor 2.273), DOI: 10.1016/j.chb.2014.12.030.
- [17] Ghusoon Salim Basheer, Mohd Sharifuddin Ahmad, Alicia Y.C. Tang, Azhana Ahmad, and Mohd. Zaliman Yussof, “A Novel Conflict Resolution Strategy in Multi-agent Systems: Concept and Model”, *Advanced Approaches to Intelligent Information and Database Systems, study in computational intelligent, Lecture Note in Artificial Intelligence*, Volume 551, pp. 35-45, 2014.
- [18] Ghusoon Salim Basheer, Mohd Sharifuddin Ahmad, Alicia Y. C. Tang, “A Conflict Classification and Resolution Modeling in Multi-agent Systems”, *Encyclopedia of Information Science and Technology* (3rd Ed.), DOI: 10.4018/978-1-4666-5888-2.ch685.
- [19] Ghusoon Salim Basheer, Mohd Sharifuddin Ahmad, and Alicia Y.C. Tang, A Framework for Conflict Resolution in Multi-agent Systems, 5th International Conference on Computational Collective Intelligence Technologies and Applications (ICCCI 2013), 11-13 September 2013, Craiova, Romania. *Lecture Note in Computer Science*, Volume 8083, pp. 195-204, 2013.
- [20] Ghusoon Salim Basheer, Mohd Sharifuddin Ahmad, Alicia Y. C. Tang, A Conceptual Multi-agent Framework using Ant Colony Optimization and Fuzzy Algorithms for Learning Style Detection, 5th Asian Conference, ACIIDS, Kuala Lumpur, Malaysia, March 18 – 20, 2013, *Proceedings, Part II: Intelligent Information and Database Systems, Lecture Notes in Artificial Intelligence (LNAI)*, Vol. 7803, pp. 549-558, 2013.
- [21] Ghusoon Salim Basheer, Alicia Y.C. Tang, Mohd Sharifuddin Ahmad, Designing Teachers’ Observation Questionnaire based on Curry’s Onion Model for Students’ Learning Styles Detection, *TEM Journal*. Volume 5, Issue 4, Pages 515-521, ISSN 2217-8309, DOI: 10.18421/TEM54-16.
- [22] Fleming, N., (2006). I’m different; Not Dumb, Modes of Presentation [V.A.R.K.] in the Tertiary Classroom.
- [23] Fleming, N. (2001). VARK a Guide to Learning Style.
- [24] Alicia Y.C. Tang and Ghusoon Salim Basheer, A Conflict Resolution Strategy Selection Method (ConfRSSM) in Multi-Agent Systems, *International Journal of Advanced Computer Science and Applications (IJACSA)*, Vol 8, Issue 5, June 2017, pp. 398-404.

Modeling and Verification of Payment System in E-Banking

Iqra Obaid

Department of Computer Sciences
COMSATS Institute of Information
Technology, (COMSATS) Lahore,
Pakistan

Syed Asad Raza Kazmi

Department of Computer Science
Government College University
(GCU), Lahore,
Pakistan

Awais Qasim

Department of Computer Science
Government College University
(GCU), Lahore,
Pakistan

Abstract—Formal modeling and verification techniques have been used to ensure the reliability and accuracy of multiple systems to be verified. In contrast to ordinary testing techniques which exhibit the presence of flaws and errors in a system, formal methods prove their absence. Electronic banking (e-banking) services have become very popular with the escalating development in the information and communication technology. Due to the presence of complexity, an e-banking system requires an efficient security model. One important approach to ensure the reliability and security of the e-banking system is through the use of formal methodologies. This study explores the opportunity of modeling interbank payment system through a case study of 1-link Automated Teller Machine (ATM). A generic verification system SPIN (Simple Promela Interpreter) is, therefore, employed to model and then to verify the integrity and security of payment system in e-banking. Linear temporal logic formulas are further summarized to assure the security of the e-banking system. The principal conclusion of the work includes a complete procedure of verification and modeling of the payment system in 1-link ATMs.

Keywords—E-banking; model checking; Simple Promela Interpreter (SPIN); formal methods; Linear Temporal Logic (LTL) formula; Promela introduction

I. INTRODUCTION

Software usage is increasing rapidly in all aspects of life and the reliability of these software has become a prime challenge, especially when the safety-critical software [17] are involved where failures often lead to a sudden loss of life, money or valuables. While using an e-banking payment system to make a transaction, it is vital that the software handling the complete process must guarantee the secure end to end transaction as well as the privacy of data to avoid its misuse. Such software is critical and not easy to be handled and developed.

During the earlier few decades, a number of languages [16] have been suggested for the specification and modeling of software oriented problems. The main aim of these languages is to render the behavior of software at the highest level of abstraction than merely as a code. Model checking verifies the correctness properties of finite-state space, where the properties of the current system are often expressed as formulas of temporal logic (TL). Later, efficient algorithms are adopted that traverse the whole model of the system and identify whether the system holds those properties or not.

Similarly, testing of payment system over the internet is being conducted from a past few years. A number of models [9] are proposed and various formulas are expressed to verify the integrity and security of the payment system, but there is no considerable and definite work to verify the payment system between multiple banks i.e. 1-Link e-banking. The most promising approach to ensure the security of an e-banking system is based on formal methods and model checking [18]. This model checking approach usually involves following steps: firstly, the payment system is modeled including all the main features, secondly, property oriented language is used to specify the reliability properties, and finally, a reachability graph with all the execution paths is drawn to verify that these paths verify the properties.

The immense challenges are: provision of authentic secure services to the banking customers as well as assurance of veracity and confidentiality of all the information that is exchanged during the process. Therefore, an efficient security model is required which should provide the banking customers with a sense of security in data usage and transactions. It should, also, be responsible for ensuring the security of the overall information or data exchanged/used in the end to end transaction. For this purpose, Simple Promela Interpreter (SPIN) [8], a standard verification tool, is employed in this study to model the system. The language used as input by the SPIN allows creating a high-level system model of many distributed systems using three components: processes, objects, message channels.

In this paper, model verification in e-banking is presented through a case study of verifying 1-Link ATMs using SPIN model checker. The results of verification clearly show that method of model checking is feasible to verify the 1-link ATMs. Furthermore, this paper explains how to employ a model checker to verify and analyze the integrity and security of payment system in e-banking system.

This paper is further structured as follows: Section II explicates the previous studies. The preliminaries required in the rest of paper are described in Section III. Then, in Section IV, 1-link ATM system model, and system properties are presented using EFSM. Section V presents the experiments of verification and discusses the results. The paper finishes with some conclusion and future work.

II. BACKGROUND

Current researches are directed towards the identification of malevolent activities and attacks in e-banking systems. These researchers have introduced the attack techniques in which, currently, only vulnerabilities are focused. In [1] an e-learning model has been implemented for secure exchange of e-content over the network. Later, the model has been formally verified using SPIN which shows that no unreachability state exists, thus, the system is viable. In [2] a protocol is proposed to identify the legitimate user. But it lacks a technique to authenticate already built e-banking systems.

In [3] PIN based ATM authentication method is evaluated and shows how contextual factors like distraction, trust, memorability influence the ATM use. Later on the basis of the findings, several implications are drawn to design an alternative secure ATM authentication system.

In [4], DHCP is presented according to modeling and verification concepts. In [5], a formal method of e-commerce system based on ebXML for the verification is presented which highlights some weaknesses of that protocol due to lack of any complete and clear specifications. In [6], a model-checking approach is applied to examine the features of ad-hoc networks. Therefore, it demonstrates, how model checking and SPIN are appropriate to study the ad-hoc networks system properties. In [7] a case study of web services is presented, and a verification technique based on model checking and SPIN [8] is proposed. The problem in adopting this checking approach is a state-space explosion. On the other hand, multiple approaches are available to combat the problem, which could be categorized as either simplifying the investigating model of the system under consideration by a higher level of abstraction, or reduction of resources consumption in the model-checking process.

In [9] an approach is proposed to verify retail banking system, which was verified in SPIN model checker. It later verifies that the model checking and SPIN are applicable for inspecting a banking system. In [10]-[15] multiple systems are presented which are verified using model checking approaches.

III. PRELIMINARIES

The model checking technique principally depends on modeling a finite state model (FSM) of a current system and then finally checking whether the desired property holds in the system or not. This approach is primarily used in the verification of protocol and hardware verification, but currently, the technique is also used in software systems. For model checking two approaches are often used, firstly, temporal model checking where the finite transition system is used to model the system and temporal logic is used to express the specifications. On the other hand, in the later approach, the automaton is used to present specifications as well as the system. Further, this system is compared to specifications in order to determine whether the behavior confirms specifications or not.

A wide-ranging model checking tools are available and in use, such as, NuSMV2 [16], SPIN, FDR, JAVA Pathfinder

and Maria. Among all, the model checker SPIN provides a user-friendly interface and it groups multiple process executions in respective equivalence classes using the theory of partial order reduction and accepts PROMELA for model specification. PROMELA language models the verification models which represent a system abstract, where only those properties are presented that are needed to be verified. PROMELA language consists of three types of objects: asynchronous message channels, processes, and data objects. Variables and message channels can be declared both globally as well as locally in a process whereas processes can only be defined globally. Processes specify the system behavior while variables and channels define the environment where processes occur.

Two basic ways can be used to verify the system using model checker SPIN. The foremost method is to take any current system and on the basis of that system, verification models are built that includes all the behaviors of the system which needs to be verified. The next approach is to construct a verification model that shows all the necessary specifications of the system. Such system models serve as a high-level description of the system under consideration.

The temporal logics used in model checking can be classified into two types: Computational Tree Logic and Linear Temporal Logic. Computational Tree Logic (CTL) is also identified as branching-time logic i.e. its model is a tree structure which is suitable mostly for hardware verification applications; while Linear Temporal Logic (LTL) is known as linear time logic basically used for software verification applications. The model checker SPIN supports LTL for the specification of system properties, which have natural language like statements. Linear temporal logic consists of a few operators, such as “O” (next state), “U” (until), “<” (eventually), “W” (weak until or unless) and “[]” (always/square). By combining these with Boolean operators, Linear Temporal Logic can be used to define many important properties of a software system under consideration.

IV. FORMAL MODELING OF INTERNET PAYMENT SYSTEM

A. Extended Finite State Machine

ATMs, nowadays, are the most rapidly emerging sensation of the internet banking technology. With the passage of time, the ordinary ATM systems are replaced by 1-link ATM that is linking multiple banks across the countries. Thus, it is not only helping the banks to handle their clients but also the clients to access their bank accounts from anywhere in the world. The main operations of these ATMs, similar to ordinary ATMs, include transaction inquiries, cash withdrawal, cash deposits, account transfers, bills payment and many others. In this section, a model will be presented (including both Promela and EFSM model) of 1-link ATM system. A simple model is designed so that a reduced number of states can be acquired which could be easily managed during formal verification. Particularly, how the PIN or ATM card number or other related details are encrypted or decrypted at various stages during the process, have been ignored.

An ATM is used to login into the bank account using the ATM card and the PIN, to perform the desired operation

against an account (withdraw cash, deposit cash, bill payment, inquire balance, etc.), and finally log off after performing the desired operation. A user always gets three chances to login into the account using some valid PIN, afterwards, if the client/user fails the ATM card is locked by the ATM till it is reset by the bank. But this function is also performed by a regular or an ordinary ATM. The basic difference between an ordinary and a 1-link ATM lies in the use of any ATM card in the ATM i.e. in 1-link ATMs one could perform the transaction from his account using any 1-link ATM whether it belongs to the respective bank or not.

1-link ATM involves four parties, the client or the cardholder, the cash dispenser (terminal), the ATM network (consortium) and the host ATM server. The client interacts with the bank through the cash dispenser or terminal to perform any kind of operation. The cash dispenser first receives a request from the client to perform a specific operation, and the cash dispenser, then, forwards the request to the ATM network or consortium. Consortium after checking the bank details forwards the request to the respective host bank server of which the client holds the account. The host bank server after receiving the request from the ATM Network again gives the response to the ATM network in the form of approval or rejection. So after the response from the bank server, ATM network then forwards this response to the cash dispenser to make an appropriate response to the client.

Some specifications in PROMELA language need to be described of 1-link ATM in order to use SPIN model checker. But before going into the description of properties in PROMELA, we need to model the specifications of 1-link ATM in EFSM (Extended Finite State Machine). Fig. 1 shows the basic specifications of 1-link ATM using EFSM. This model will be further expressed in PROMELA. The variable *loginAttempts* in the EFSM describes the total number of unsuccessful attempts by the client to enter a valid PIN of the

ATM card and when the variable is greater or equal to 3 the card is locked by the ATM.

In Fig. 1 the EFSM of the 1-link ATM is presented, which consists of nine states. The label of a transition *RequestWithdrawal/PINOk/LoginAttempts<3* shows that when the client requests for the cash withdrawal then two conditions should be satisfied, i.e. the PIN should be valid and the login attempts should be less than or equal to 3. In this case, only, the state will be changed from “card valid” to “withdraw balance”. Similarly, the state transition *Logon/PINInvalid/LoginAttempts>=3* shows that even if the client again tries to login and the Login Attempts become greater or equal to 3 the ATM card will be locked by the ATM, therefore, the state will change from “Re-Logon” to “Card Locked”.

On the other hand, the label of transition *WithdrawAmount<AccountBalance/PINOk* shows that when the cardholder/client requests to withdraw amount and the PIN is verified, than the ATM server will check whether the amount to be withdrawn is smaller than the account balance of the client. If so, then the state will be changed to “Transaction Ok” otherwise if *WithdrawAmount>AccountBalance/PINOk* then the state will be changed to “Transaction Invalid”.

B. PROMELA Model

A PROMELA program comprises of 3 basic types of objects: asynchronous message channels, processes, and data types. Processes define the behavior of processes, variables are used to store information of the system being modeled and message channels are basically used for modeling the communication between the processes. The syntax of PROMELA allows the creation of multiple processes dynamically which could be synchronized through message channels as PROMELA language is similar to that of C language.

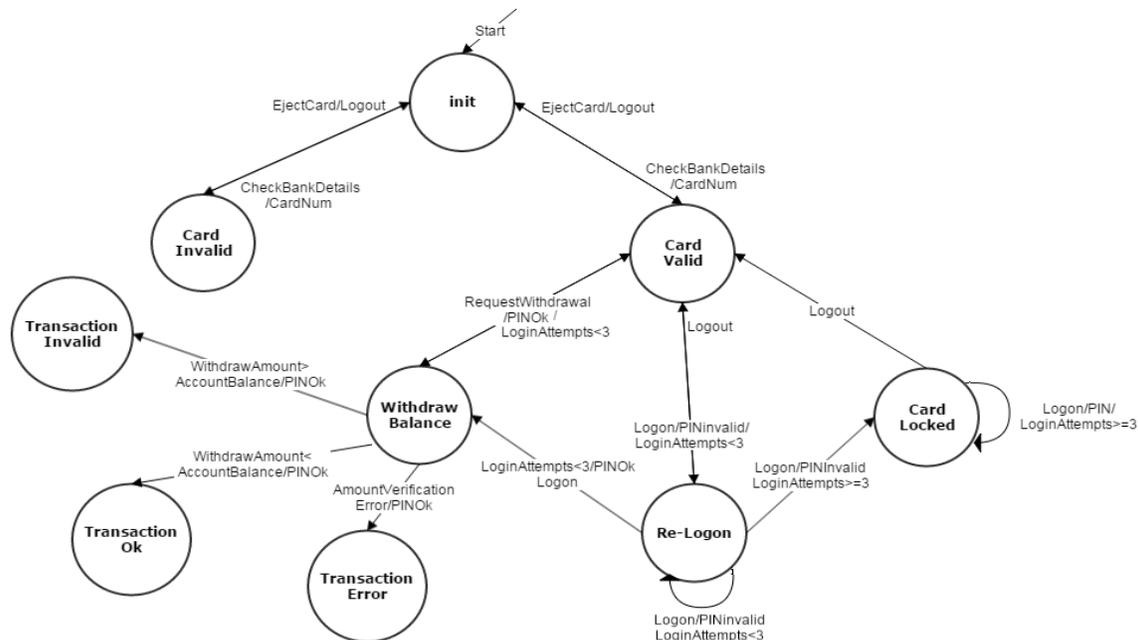


Fig. 1. EFSM of internet payment system.

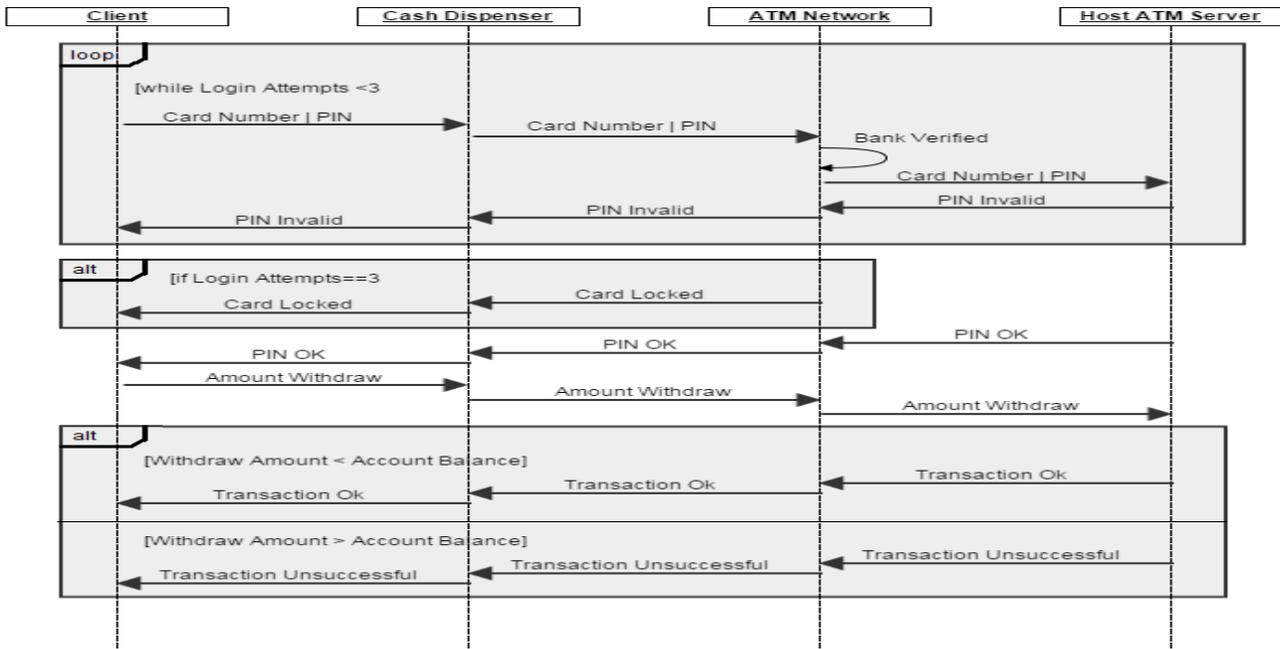


Fig. 2. Sequence diagram of an ATM system.

As EFSM of the system is modeled, the next stage involves the modeling of the system using PROMELA. We need to choose how the system communicates with the four parties. The simple message flow is presented using sequence diagram in Fig. 2. Messages which need to be transferred between the channels are defined as `mtype={card,PIN,insertCard, requestLogon, PINinvalid, PINininvalid, cardLocked, requestWithdrawal, transactionOk, checkBankDetails, amountInvalid, transactionUnsuccessful, verifyPIN, endTransaction, removeCard, cardInvalid, cardValid, transactionError}`. Message channels are used to represent the inter-process communication. The channels are declared using buffer size and data types as shown in Fig. 3.

```

chan cashDispenser_client = [MAX] of {byte,mtype,mtype,byte};
//Message from ATM to card holder
chan client_cashDispenser = [MAX] of {byte,mtype,mtype,byte};
//Message from card holder to ATM
chan cashDispenser_atmNetwork = [MAX] of {byte,mtype,mtype,byte};
//Message from ATM to consortium
chan atmNetwork_cashDispenser = [MAX] of {byte,mtype,mtype,byte};
//Message from consortium to ATM
chan atmNetwork_hostAtmServer = [MAX] of {byte,mtype,mtype,byte};
//Message from consortium to ATM server
chan hostAtmServer_atmNetwork = [MAX] of {byte,mtype,mtype,byte};
//Message from ATM server to consortium
    
```

Fig. 3. Definitions of channels in PROMELA.

Here the keyword MAX is used to represent the terminal numbers or simply the number of clients or card holders. The channel `client_cashDispenser` sends the messages from the card holder to the cash dispenser or ATM and each of the message have four parts: the first part is of type byte that shows the card number of the card inserted by the client in the ATM, the second field is of type mtype that represents the operations carried out by the card holder to the cash dispenser,

the third field represents the amount that relates to the operation and is also of type byte and the last field represents the data send or the response received. Similarly, `cashDispenser_client` sends messages from the cash dispenser/ATM to the client. In this way `cashDispenser_atmNetwork` send messages from the ATM to the ATM Network or the consortium, `atmNetwork_hostAtmServer` send messages from the consortium to the host ATM network, i.e. the network of that bank to which the card belongs and similarly vice versa.

As the model checker SPIN does not allow the user to participate when the process is running, i.e. SPIN runs in closed conditions. Therefore, to traverse all the cases, i.e. in case the PIN is entered incorrect, PIN is entered correct or if the PIN is entered wrong three times and many others, the variables for the amount in the account, the amount withdrawn, PIN is whether OK or not, need to be designed carefully.

For example, if the balance of the account is 4000 and the withdrawal amount is 1000, the SPIN will only traverse the path of `transactionOk`, while rest of the states, `transactionError` and `transactionUnsuccessful`, are not reachable. So just on the base of this case, the result can't be assumed that the system model under consideration does not hold the properties. In addition, if the user enters the wrong PIN 3 times than the path labeled `cardLocked` will only be traversed by the model checker while the rest of the paths `PINOK` and `PINinvalid` are again not reachable. So to model a realistic system, these variables need to be changed. So during the verification of the system, the model is verified against multiple different sets of values. The processes of the card holder, cash dispenser (ATM), consortium and host ATM server respectively are modeled in PROMELA using SPIN. Major functions of each process are represented below. A process of client gets ATM card, check its validity, gets PIN form user.

```
Request_Logon:
if
::atomic{ cashDispenser_client?eval(cardNum),requestLogon,
0,PINinvalid->
    PINOK=true;
    goto selectOperation}

::atomic{ cashDispenser_client?eval(cardNum),requestLogon,
0,PINinvalid->
    PINOK=false;
    printf("You entered an invalid PIN");
    goto Request_Logon}

::atomic{ cashDispenser_client?eval(cardNum),requestLogon,
0,cardLocked->
    PINOK=false;
    printf("Card is Locked");
    cardLock=true;
    goto CardLocked}

fi;
    A process of dispenser dispense cash only when the PIN
and the amount entered is valid

    cashDispenser_atmNetwork!cardNum,requestLogon,0,PIN;
if
:: atomic{
atmNetwork_cashDispenser?eval(cardNum),requestLogon,0,P
INinvalid->
cashDispenser_client!cardNum,requestLogon,0,PINinvalid;
    goto Cash_Withdrawal;}
:: atomic{
atmNetwork_cashDispenser?eval(cardNum),requestLogon,0,P
INinvalid->
cashDispenser_client!cardNum,requestLogon,0,PINinvalid;
    goto start;}
:: atomic{
atmNetwork_cashDispenser?eval(cardNum),requestLogon,0,c
ardLocked->
cashDispenser_client!cardNum,requestLogon,0,cardLocked;
    cardLock=true;
    goto start;
}
fi;

    The process of ATM network refers to the network of the
ATM owner bank which checks bank details of the client and
then sends the details to the host bank network for PIN
verification and other account details verification.

Check_Bank_Details:
if
::atmNetwork_cashDispenser?eval(cardNum),requestLogon,0,
PIN->
if
::(loginAttempts<3)->
    atmNetwork_hostAtmServer!cardNum,verifyPIN,0,PIN->
if
::atomic{hostAtmServer_atmNetwork?eval(cardNum
),verifyPIN,0,PINinvalid->
```

```
atmNetwork_cashDispenser!cardNum,requestLogon,
0,PINinvalid;
    goto CashWithdrawal;}
::atomic{hostAtmServer_atmNetwork?eval(cardNum
),verifyPIN,0,PINinvalid->
    atmNetwork_cashDispenser!cardNum,requestLogon,
0,PINinvalid;
    loginAttempts=loginAttempts+1;}
fi;
::atomic{atmNetwork_cashDispenser!cardNum,requestLogon,
cardLocked->
    loginAttempts=loginAttempts+1;
    goto Check_Bank_Details}
fi;

    The process of host ATM network refers to the network of
client host bank which verifies the PIN and other account
details.

server_start:
atmNetwork_hostAtmServer?eval(cardNum),verifyPIN,0,PIN
valid->
if
::atomic{hostAtmServer_atmNetwork!cardNum,verifyPIN,0,P
INinvalid->
    PINOK=true;
    goto Cash_Withdrawal;}
::atomic{hostAtmServer_atmNetwork!cardNum,verifyPIN,0,P
INinvalid->
    PINOK=false;
    goto server_start;}
fi;
Cash_Withdrawal:
    atmNetwork_hostAtmServer?cardNum,requestWithd
rawal,withdrawAmount,PIN->
if
::atomic{(withdrawAmount<=accountBalance)->
    if
::atomic{hostAtmServer_atmNetwork!cardNum,request
Withdrawal,withdrawAmount,transactionOk->
    accountBalance=accountBalance-withdrawAmount;
    transactionOK=true;
    goto server_start;}
::atomic{hostAtmServer_atmNetwork!cardNum,request
Withdrawal,withdrawAmount,transactionError->
    transactionOK=false;
    goto server_start;}
fi;
}
::atomic{(withdrawAmount>accountBalance)->
    hostAtmServer_atmNetwork!cardNum,requestWithd
rawal,withdrawAmount,transactionUnsuccessful->
    transactionOK=false;
    goto server_start;}
fi;
```

V. RESULTS

For the model specifications given in PROMELA, SPIN helps the users to identify the deadlocks or unreachable code

in the model. In addition, SPIN can verify multiple claims on the execution of model by verifying the LTL properties inserted in SPIN.

TABLE I. LTL FORMULAS OF 1-LINK ATM MODEL

	LTL formulas
1	$\square(\text{PINOK} \ \&\& \ \text{transactionOK} \ \rightarrow \ \diamond \text{cashDispensed})$
2	$\square(\text{ejectCard} \ \rightarrow \ \diamond \text{printReciept})$
3	$\square((\text{cashDispensed} \ \&\& \ \text{!continueTransaction}) \ \rightarrow \ \diamond(\text{printReciept} \ \&\& \ \text{ejectCard}))$
4	$\square!(\text{cardLock} \ \&\& \ \text{!ejectCard})$

In this work a 1-link ATM system is modeled in PROMELA and its various properties are analyzed in the above sections. In this section, several properties of the 1-link ATM system presented above are verified using SPIN. For example, the size of the model and the time for the verification of the model is measured.

First of all, we have run SPIN for three cardholders to check for the errors, elapsed time and memory usage. Fig. 4 represents the results of verification in SPIN, the first line of the results represent the version of the SPIN verifier used in the verification of the model. In the results, the "+" sign in the second line indicates that the default algorithm is adopted. In Line 3 the search type is represented. The "-" sign in the next line represents that it is not using LTL formulas. In Line 5 it is shown that the process doesn't violate any of user defined conditions. Line 6 represents that acceptable cycles are also detected by the process. Later the next line represents invalid end states which indicate the absence of any deadlock. All the later results represent the information about the model about the states, memory usage, etc.

```
(Spin Version 6.1.0 -- 4 May 2011)
+ Partial Order Reduction

Full statespace search for:
  never claim      - (not selected)
  assertion violations +
  acceptance cycles + (fairness disabled)
  invalid end states +

State-vector 444 byte, depth reached 21, errors: 0
 342 states, stored
 571 states, matched
 913 transitions (= stored+matched)
 0 atomic steps
hash conflicts: 0 (resolved)

Stats on memory usage (in Megabytes):
 0.150 equivalent memory usage for states (stored*(State-vector + overhead))
 0.363 actual memory usage for states (unsuccessful compression: 242.23%)
state-vector as stored = 1098 byte + 16 byte overhead
 2.000 memory used for hash table (-w19)
 3.433 memory used for DFS stack (-m100000)
 5.726 total actual memory usage

pan: elapsed time 0.002 seconds
```

Fig. 4. Verification results of 3 card holders.

In addition to this SPIN also helps the user to verify the model using LTL properties. LTL allows the user to express the behavior of the system using temporal properties that system must conform. Table 1 represents LTL formulas applied on the system, the first and most important property about the ATMs is that only when the cardholder enters the correct PIN and he has enough balance in his account, i.e. account balance should be greater than the amount to be withdrawn, then cash could be dispensed from the ATM. This property is expressed as LTL formula as: $\square(\text{PINOK} \ \&\& \ \text{transactionOK} \ \rightarrow \ \diamond \text{cashDispensed})$.

The next LTL formula to be verified confirms that the ATM prints the receipt whenever the ATM ejects the card after the cash is dispensed. The property can be stated as: $\square(\text{ejectCard} \ \rightarrow \ \diamond \text{printReciept})$.

The next property which needs to be verified is that only when the cash is dispensed by the ATM and the user doesn't wish to continue transaction then the ATM will eject the ATM card and print the receipt. This property of the ATM can be verified by the LTL formula stated as: $\square((\text{cashDispensed} \ \&\& \ \text{!continueTransaction}) \ \rightarrow \ \diamond(\text{printReciept} \ \&\& \ \text{ejectCard}))$

```
(Spin Version 6.1.0 -- 4 May 2011)
+ Partial Order Reduction

Full statespace search for:
  never claim      + (l0)
  assertion violations + (if within scope of claim)
  non-progress cycles + (fairness disabled)
  invalid end states - (disabled by never claim)

State-vector 440 byte, depth reached 57, errors: 0
 181 states, stored
 267 states, matched
 448 transitions (= stored+matched)
 0 atomic steps
hash conflicts: 0 (resolved)

2.539 total actual memory usage

unreached in proctype Client
(12 of 52 states)
unreached in proctype Cash_Dispenser
(18 of 53 states)
unreached in proctype ATM_Network
(13 of 49 states)
unreached in proctype Host_ATM_Server
(5 of 33 states)
unreached in init
(0 of 23 states)
unreached in claim l0
 _spin_nvr.tmp:10, state 11, "-end-"
(1 of 11 states)

pan: elapsed time 0.001 seconds
```

Fig. 5. Result of Verification of LTL formula 3.

Above stated LTL properties should be verified by our system, i.e. 1-link ATM system. Now we will present certain results after applying those LTL formulas on the PROMELA model. After performing multiple experiments we came to the conclusion that the LTL formulas verify our PROMELA model. Fig. 5 presents the result, when SPIN performs a full state space search using the LTL formula $\square((\text{cashDispensed} \ \&\& \ \text{!continueTransaction}) \ \rightarrow \ \diamond(\text{printReciept} \ \&\& \ \text{ejectCard}))$ on the PROMELA model. The first few lines of the result are already explained before but as compared to the last results, there are some differences like unreached in proctype. The term represents that there are few unreachable states in this

case because to verify the case we have defined different values to the variables like cashDispensed, continueTransaction, etc. As for the above case, the ATM should have dispensed the cash and the user didn't ask to continue transaction so in that case the ATM will now eject the card and print the receipt. This means the conditions like the user wishes to continue the transaction, or the user enters some invalid PIN, etc. will never arise. So in this way, the result that there are some states that could never be reached in this case is acceptable.

VI. CONCLUSION

In this paper model checking approach is introduced to verify 1-link ATM Systems. Firstly we consider 1-link ATM as an extended finite machine that is further presented in PROMELA. Further different properties of the system are expressed using LTL formulas and then the properties are verified using SPIN model checker. Finally, it proves that the model checking technology and SPIN model checker are both appropriate for verifying the business flows of 1-link ATM systems.

For our future research, we will try to modify the model by verifying more security related properties that include cash deposit, bill payment, and cash transfer using SPIN. Moreover the research can be expanded in other related domains of banking, as mobile banking, also.

REFERENCES

- [1] Al Obisat, Farhan M., and Hazim S. AlRawashdeh. "Formal Verification of a Secure Model for Building E-Learning Systems." *INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS* 7.6 (2016): 377-380.
- [2] Osama Dandash, Phu Dung Le, and Bala Srinivasan. Security analysis for internet banking models. In *Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, 2007. SNPD 2007. Eighth ACIS International Conference on*, volume 3, pages 1141-1146. IEEE, 2007.
- [3] Alexander De Luca, Marc Langheinrich, and Heinrich Hussmann. Towards understanding atm security: a field study of real world atm use. In *Proceedings of the sixth symposium on usable privacy and security*, page 16. ACM, 2010.
- [4] Syed MS Islam, Mohammed H Sqalli, and Sohel Khan. Modeling and formal verification of dhcp using spin. *IJCSA*, 3(2):145-159, 2006.
- [5] Marina Mongiello. Finite-state verification of the ebxml protocol. *Electronic Commerce Research and Applications*, 5(2):147-169, 2006.
- [6] Vladimir A Oleshchuk. Modeling, specification and verification of ad-hoc sensor networks using spin. *Computer Standards & Interfaces*, 28(2):159-165, 2005.
- [7] Raman Kazhamiakin, Marco Pistore, and Marco Roveri. Formal verification of requirements using spin: A case study on web services. In *Software Engineering and Formal Methods, 2004. SEFM 2004. Proceedings of the Second International Conference on*, pages 406-415. IEEE, 2004.
- [8] Gerard J Holzmann. *The SPIN model checker: Primer and reference manual*, volume 1003. Addison-Wesley Reading, 2004.
- [9] Huiling Shi, Wenke Ma, Meihong Yang, and Xinchang Zhang. A case study of model checking retail banking system with spin. *Journal of computers*, 7(10):2503-2510, 2012.
- [10] Wei Zhang. Model checking and verification of the internet payment system with spin. *Journal of Software*, pages 235-257, 2012.
- [11] Tarek MI El-Sakka and M Zaki. Using predicate-based model checker for verifying e-commerce protocols. *IJ Network Security*, 16(2), 2014.
- [12] Both, Andreas, Wolf Zimmermann, and René Franke. "Model checking of component protocol conformance—optimizations by reducing false negatives." *Electronic Notes in Theoretical Computer Science* 263 (2010): 67-94.
- [13] Bernardo M. David Flavio G. Deus Rafael Timoteo de Sousa Jr. Laerte Peotta, Marcelo D. Holtz. A formal classification of internet banking attacks and vulnerabilities. *International Journal of Computer Science & Information Technology (IJCSIT)*, 3, Feb 2011.
- [14] Haiping Xu and Yi-Tsung Cheng. Model checking bidding behaviors in internet concurrent auctions. *International Journal of Computer Systems Science & Engineering*, 22(4):179-191, 2007.
- [15] Tuan, Luu Anh, Man Chun Zheng, and Quan Thanh Tho. "Modeling and verification of safety critical systems: A case study on pacemaker." *Secure Software Integration and Reliability Improvement (SSIRI), 2010 Fourth International Conference on*. IEEE, 2010.
- [16] Cimatti, Alessandro, et al. "Nusmv 2: An opensource tool for symbolic model checking." *International Conference on Computer Aided Verification*. Springer, Berlin, Heidelberg, 2002.
- [17] Lockhart, Jonathan, Carla Purdy, and Philip Wilsey. "Formal methods for safety critical system specification." *Circuits and Systems (MWSCAS), 2014 IEEE 57th International Midwest Symposium on*. IEEE, 2014.
- [18] Bozzano, Marco, and Adolfo Villafiorita. "Improving system reliability via model checking: The FSAP/NuSMV-SA safety analysis platform." *SAFECOMP*. Vol. 2788. 2003.

ReCSDN: Resilient Controller for Software Defined Networks

Soomaiya Hamid, Narmeen Zakaria Bawany, Jawwad Ahmed Shamsi
Systems Research Laboratory, Department of Computer Science
FAST National University of Computer and Emerging Sciences
Karachi, Pakistan

Abstract—Software Defined Networking (SDN) is an emerging network paradigm that provides central control over the network. Although, this simplifies the network management and makes efficient use of network resources, it introduces new threats to network reliability and scalability. In fact, a single centralized controller is a single point of failure. Moreover, a single controller may become a performance bottleneck as processing overhead increases. Distributed SDN controller platforms improve the reliability and scalability to some extent, however they remain vulnerable to Distributed Denial of Service (DDoS) attacks, specifically on control plane. We believe that there is a need for a distributed controller framework that is capable of providing service continuity without performance degradation in case of excessive network traffic or DDoS attacks on controller. In this paper, we aim to address the vulnerabilities of SDN control plane. We propose and implement an efficient and Resilient Controller for Software Defined Network (ReCSDN). This framework is capable of detecting and mitigating DDoS attacks timely and ensures the continuity of services without performance degradation. We created an experimental test bed using Mininet to conduct extensive experiments. We deployed ReCSDN on top of Open Network Operating System (ONOS) cluster to confirm the viability of our approach. The experiment results show that with ReCSDN, control plane is not only able to withstand excessive network load but will also continue to provide services in case of any controller failure.

Keywords—Software Defined Networking (SDN); SDN Controller security; Distributed Denial of Service (DDoS) attack; load balancing; SDN controller cluster; Open Network Operating System (ONOS)

I. INTRODUCTION

Software Defined Networking (SDN) paradigm has revolutionized the traditional networking by separating the control plane and data plane of the network. With this separation of the control plane and data plane, control logic is implemented in logically centralized controller and network switches becomes simple forwarding devices [1]. This decoupling provides several benefits which includes easier network management, increased visibility into the network, programmability, efficient use of network resources, dynamic updating of network policies [2], [3]. The centralized control plane leads to global knowledge of the network thereby providing effective resource management. Moreover, network policies can be easily configured and modified via software applications running on top of the controller. Customized

network applications can be developed and deployed directly without any vendor dependency [4], [5].

Nevertheless, these core benefits that are the hype of SDN are also the main causes of concern. The centralized control plane that provides critical advantages over the traditional networking has introduced new threat vectors. First and foremost it can become the single point of failure [6]. The controller becomes the core of network and any attack, such as, DDoS attack can bring down the whole network. This vulnerability introduces new threat vector in SDN. Many approaches, such as primary backup replication mechanism and distributed controller platforms [7] exists that addresses this critical reliability issue. However, there are numerous issues with these approaches which makes it an open research problem [8], [9].

Second, the controller may turn out to be a performance bottleneck as the network size increases [8]. Whenever a new flow is initiated in the network, the OpenFlow switch forwards it to controller for deciding the suitable forwarding path. Similarly, all the unknown flows that are not recognized by the switch are sent to controller for processing. The performance of the controller is largely affected as the network grows thereby increasing the number of traffic flows. Various schemes for controller load-balancing [10] has been proposed to improve the performance of centralized controller platforms. However, due to their limited capabilities the problem remains an open research area.

Many researchers have explored the new threat vectors introduced by SDN [11], [12]. Several attacks, including DDoS attacks, and their mitigation strategies has been proposed [13]-[15] for SDN networks. However, very limited work has been done to detect and mitigate attacks specifically on SDN controllers[6]. Also, most of this work has been done for centralized controllers such as Floodlight [16]-[18] and POX [19], [20].

Keeping in view the above mentioned limitations we presume that there is a need to explore load balancing and DDoS attack vulnerabilities in distributed SDN controller platforms. We also need a framework that can detect excessive load on controllers and ensure the continuity of services without performance degradation.

To this end, we propose and describe Resilient Controller for Software Defined Networks (ReCSDN) that addresses the above mentioned problems. ReCSDN, is a novel framework

that is built on top of a distributed controller environment. It provides a reliable, efficient and resilient control plane that not only overcomes the single point of failure problem but also ensures the service continuity without performance degradation. ReCSDN is able to detect the excessive network traffic coming to the controller and provides a load-balancing mechanism that ensures that performance of controller is not degraded. Excessive traffic may be generated due to DDoS attack on controller or flash crowds. In either case, the objective of ReCSDN is to provide fault tolerance and service continuity while maintaining the performance quality. ReCSDN also ensures that network latency remains consistent and does not increase as we increase the number of distributed controllers.

The main contributions of this paper are summarized below:

- We proposed and implemented ReCSDN, a reliable, efficient and resilient framework for SDN.
- We performed extensive experiments using Mininet and ONOS [21], a distributed SDN controller platform to test the effectiveness of our framework.
- We were able to detect and mitigate DDoS attack on SDN controller effectively.
- We are able to ensure quality of service performance by providing appropriate load balancing among controllers.
- We are able to provide fault tolerance by using backup controllers timely.

The rest of this paper is organized as follows. Section II comprises three sub-sections. First two sections briefly introduces Software Defined Networking and ONOS followed by a detailed review of existing research on SDN security. The proposed architecture and its implementation is discussed in Section III followed by the threat model which is discussed in Section IV. The experimental setup and results are presented in Section V and Section VI, respectively. Section VII concludes the paper.

II. BACKGROUND AND RELATED WORK

We have divided this section in three sub-sections. Motivation for this research and the benefits of software defined networking over traditional networking are enlightened in the first sub-section. Next sub-section discusses ONOS, followed by the related work.

A. Towards Software Defined Networks

Building and administrating a computer network is an onerous task. Managing networks includes many challenges, such as, heterogeneity of network elements [22], vendor dependency [23], lack of centralized control, no programmability. Moreover configuration of complex networks which are dynamic in nature is more difficult, because of lack of automated mechanism for defining centralized policies. This creates scalability and configuration issues which makes traditional networks less innovative [24]. The network administrator have to configure each network device individually to apply network policies [25]. As the size

of network increases number of devices also increases thereby increasing the administrative overhead.

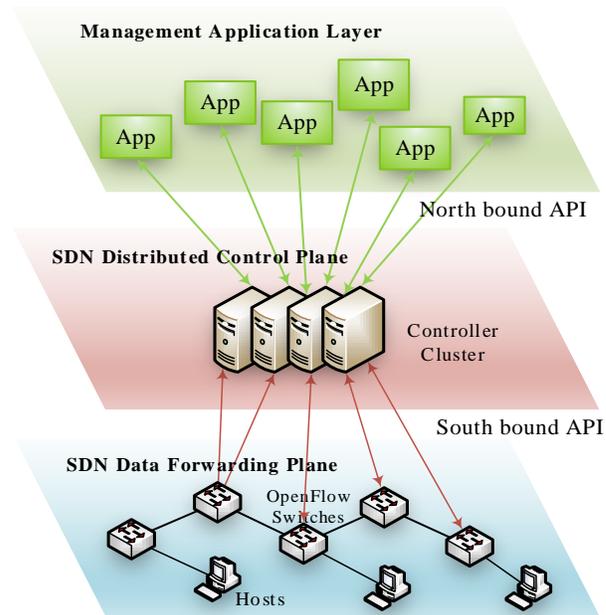


Fig. 1. Software defined network architecture.

SDN addresses the above mentioned issues by separating the control plane and the data plane as shown in Fig. 1. In SDN control plane provides a centralized control of the network. Control plane can manage the entire network centrally [12]. Major objective is to provide a centralized control over the entire network, so that all the control process and services are separated from the data forwarding tasks. Hence, the software that controls the network is decoupled from the devices that implement it [26]. Switches became simple forwarding devices that work according to the policies defined by the controller. Many open source SDN controllers has been developed which includes POX [27], NOX [28], Beacon [29] and Floodlight [30]. More recently distributed SDN controller platforms such as ONOS and OpenDaylight [31], [32] have been developed to cater the needs of large enterprise networks. We briefly discuss ONOS in the next sub-section. Apart from open source controllers, major industry leaders have also developed proprietary SDN controllers such as; HP [33], [34] and brocade [35].

Although, SDN has been gaining immense popularity since its inception, it is no silver bullet. SDN comes with its own set of vulnerabilities that were not present in traditional networks. Subsequently, after the adaptation of SDN in network infrastructures, many researchers have been questioning the security of SDN [36], [37]. The centralized control plane which has been its prominent feature has also become the major point of concern. Adversaries can launch DDoS attack on the control plane of the SDN subsequently leading to service degradation or a complete network shutdown. Similarly, performance, scalability and reliability of SDN have not been thoroughly investigated yet.

B. Open Network Operating System

The ONOS (Open Network Operating System) is an open source project hosted by The Linux Foundation. The software is written in Java and provides support for *distributed SDN* applications atop Apache Karaf OSGi container as shown in Fig. 2. The first version of ONOS was released in 2014. The ONOS is a distributed platform for SDN networks that caters the need of enterprise networks. The key features of ONOS includes scalability, high performance and high availability. ONOS is basically designed to operate as a cluster of nodes such that it can withstand the failure of individual nodes. ONOS overcomes the limitations of centralized SDN controllers like POX, NOX and Floodlight. It provides a high-level abstraction to application programmers by providing a platform for developers to write novel applications that can run on top of ONOS. Its model can be extended by programming variety of applications.

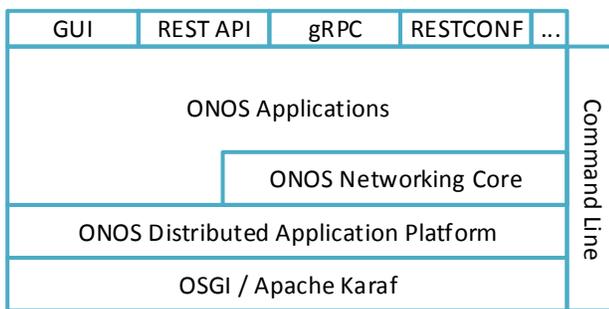


Fig. 2. ONOS software stack.

The ONOS has been used today in variety of applications ranging from multilayer network control to datacenters [38]. Major use cases of ONOS includes CORD (Central Office Re-architected as a Datacenter [39], [40], Multi-Layer Network Control, Migrating MPLS Network, and Global Research Network Development. ONOS also provides its partner driven use cases such as Huawei Agile L3 VP, Huawei Enterprise CPE, DirectTV Multicast and NEC Transport SDN.

To ensure strong consistency ONOS adopted Atomix framework after its v1.4 release. Atomix uses RAFT consensus algorithm [41] to ensure consistency among cluster nodes. Atomix deals with distributed computing problems. In contrast with the Hazelcast [42], Atomix chooses availability over consistency. Due to this Atomix ensures that data is never lost, even in the network partitioning or complete failure.

C. Related Work

Security of SDN has been a point of concern since its adoption [37]. Many researchers have questioned the security of SDN itself [12]. However others have proposed SDN based security solutions [43], [44]. DDoS attack detection in SDN with the entropy variation technique was presented in [6], [18] Niyaz et al. [45] proposed a deep learning multi vector DDoS system. Fonseca et al. [46] designed CPRecovery by component organization. Another technique was AVANT-GUARD [14] which is based on complete TCP handshake mechanism. Hong et al. [47] proposed a TopoGuard

technique. It focused the attack over data plane communication channel. R. Braga et al. [13] classified the flows by self-organizing maps. An inference-relation context based technique was presented by Aleroud et al. [48]. They proposed technique utilizes contextual similarity with existing attack patterns to identify DoS in an OpenFlow infrastructure. Cui et al. [49] performed attack detection by neural network techniques. Botelho et al. [50] has replicated the sheared database of the whole network state to improve reliability.

Majority of the research work discussed above is based on the centralized SDN controller. Few researchers have implemented replication between master and backup controller. When master controller fails, backup controller becomes an active controller. In contrast to existing research, we have developed a resilient framework for distributed controller environment. We emulated our network using ONOS. In our approach, all controllers in a cluster are active. If there is an attack on any of the controllers, load is distributed to other controllers within a network. The controllers share the information of flows and switches consistently. Moreover in previous research works, different SDN controllers [51] were used such as, POX [27], NOX [28], Beacon [29], and Floodlight [30], but ONOS [21] controller was not explored for the attack detection. In this paper we are creating a distributed environment using ONOS controller with Mininet emulation to detect DDoS flooding attack on the controller.

III. PROPOSED APPROACH

This section presents the design of Resilient Controller for Software Defined Network – ReCSDN. The ReCSDN is a proficient solution that efficiently detects and mitigates DDoS attack on the control plane. It is capable of providing fault tolerant and consistent services to the network without performance degradation. ReCSDN detects excessive traffic network coming to the controller and uses load balancing mechanism that ensures the reliability and performance of the control plane. The ReCSDN module runs on top of distributed controller platform. It monitors the processing load of the controller and ensures that the load is distributed to other controllers in the cluster before any controller reaches its full capacity.

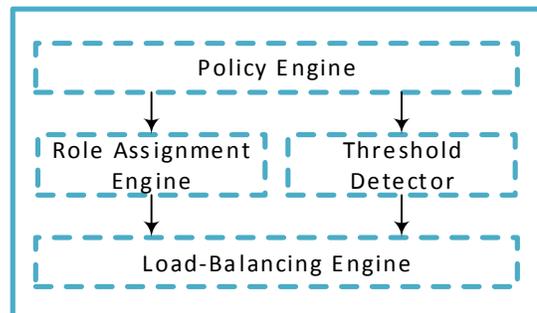


Fig. 3. ReCSDN application workflow.

ReCSDN consists of four modules as depicted in Fig. 3. The Policy Engine is used to configure the number of active and backup controllers within a cluster. Also, threshold for

each controller is setup using the Policy Engine. The threshold value indicates the tolerance level of controller after which the performance of controller may be degraded. Therefore, the threshold Detector module monitors the state of controller to ensure that load of the active controller is distributed by the Load Balancing Engine before crossing the threshold. The Role Assignment Engine is used to assign the master/backup status to controllers within a cluster.

IV. THREAT MODEL

SDN Controller is the most critical element of SDN. It serves as a centralized control of the whole network. The attack on SDN controller will result in complete shutdown of network.

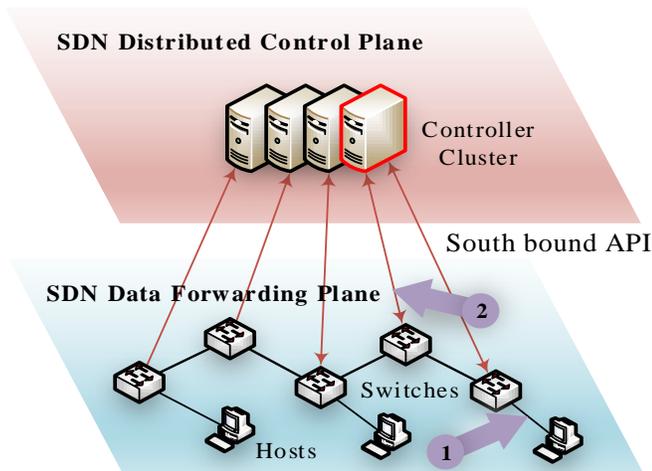


Fig. 4. Threat Model for SDN control plane.

The scope of this work is focused on DDoS attack on SDN controller. In such an attack adversaries may use compromised nodes to send unknown flows to the OpenFlow switches. These unknown flows are not recognized by the OpenFlow switches and are sent to controller for further processing. Thus, the controller is overwhelmed by the huge number of illegitimate packets and is either completely halted or results in its performance degradation.

We have considered two threat vectors that targets SDN control plane in our threat model. The two threat vector are based on generating flows that are not recognized by the switches thereby targeting the SDN controller and the communication channel between SDN control plane and data plane. Fig. 4 depicts the threat model. During a DDoS attack multiple hosts generate fake or forged traffic. Such traffic flows are not recognized by OpenFlow switches and are forwarded to controller for deciding the suitable forwarding path. This scenario not only depletes controller resources but also results in exhaustion of the communication channel between controller and the network.

V. EXPERIMENTAL SETUP

To determine the viability of our approach, we have setup a test bed on a server with an Intel Core i7, 3.67 GHz

processor and 16GB RAM running Ubuntu 14.04.5. We conducted our experiments to emulate the DDoS attack scenario on a controller using Mininet and ONOS cluster. We deployed ReCSDN module on top of ONOS cluster. We included different types of legitimate traffic to build a realistic scenario. The legitimate traffic included TCP, UDP and ICMP. The D-ITG tool [52] was used to generate the traffic and to collect performance metrics. The metrics include delay, jitter and number of packet loss.

To create DDoS attack scenario on a controller huge number of new flow requests were generated. When a new flow is received by the OpenFlow switches, it is not recognized and is forwarded to the SDN controller for deciding the transmission path. The increase in the number of new flow requests, increases the processing overhead of controller leading to performance degradation or completed denial of service. The ReCSDN module monitors the network and controllers state and ensures that load of the controller is distributed to other controllers in the network before the threshold is reached. The ReCSDN provides fault tolerance mechanism by using back controllers. These back controllers are active controllers that can also be used for load balancing in case of DDoS attack or flash crowds.

We conducted extensive experiments discussed in next section to evaluate the performance and reliability of ReCSDN.

VI. RESULTS AND ANALYSIS

One of the key characteristics of the ReCSDN is achieving resiliency. We exploited the distributed architecture of ONOS to build a fault tolerant environment. We created a cluster of ONOS controllers that provided multiple backups for each active controller. Multiple backup controllers lead to more fault tolerance. As ReCSDN is specifically developed to work with distributed controller cluster a key aspect of characterizing the performance of ReCSDN is to analyze and compare performance at various scales. We created several scenarios to measure the response time as number of controllers in a cluster scales from 1 node to 3, 5, and 7 nodes. We observed that increasing the number of controllers within a cluster has no overhead and response time remains below 0.1ms. Fig. 5 depicts the result of experiment.

To evaluate the effect of increasing number of controllers in a ReCSDN cluster on latency we conducted multiple experiments. For each experiment we increased the number of controllers from 1 to 3, 5 and 7. We generated constant amount of TCP traffic for each experiment and noted delay and jitter. The network traffic comprises huge number of unknown flows. The ReCSDN ensured that load is distributed among the other controllers before the master controller is overwhelmed. As we increase the number of controllers in the cluster the delay decreases as shown in Fig. 6.

The latency decreased due to the consistent load distribution among the controllers. The overall performance of network improved as ReCSDN enabled load balancing before the maximum capacity of a controller is reached.

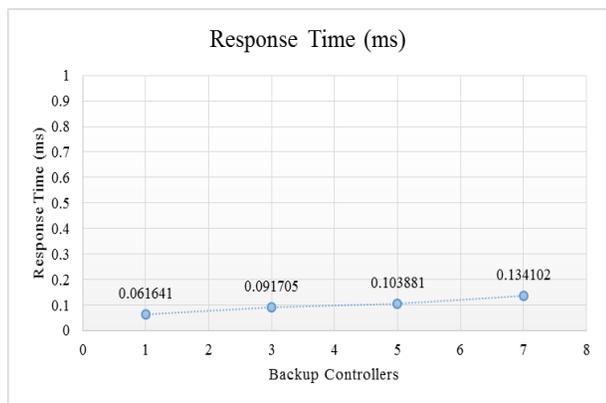


Fig. 5. Effect of adding backup controllers by calculating response time. This test was performed with variant number of backup controllers, 1, 3, 5, and 7 respectively.

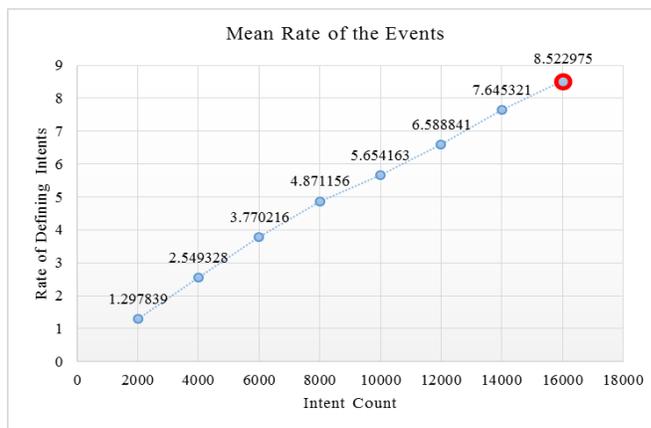


Fig. 7. Stress test for checking controller processing capability. Red indicator shows controller resources saturation point.

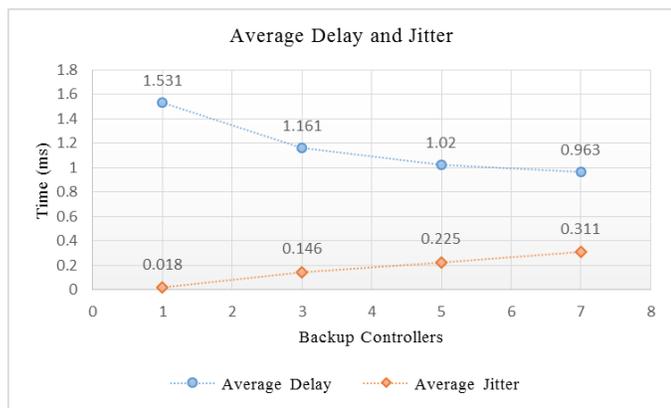


Fig. 6. Delay decreases as number of controllers increased in ReCSDN cluster.

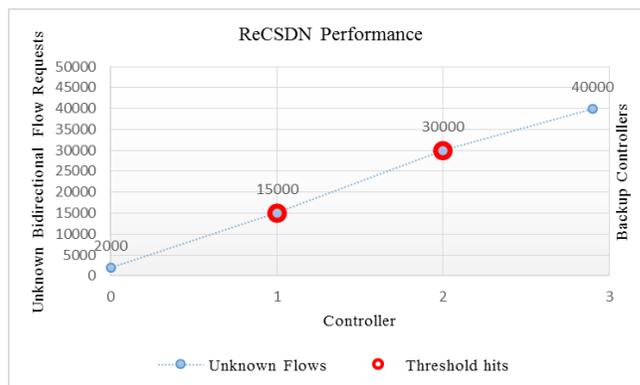


Fig. 8. ReCSDN performance evaluation.

To determine the single controller’s capacity of processing maximum number of flows we performed a stress test. We flooded the controller with new flow requests, generated by pushing random intents. Intents are high-level policies that are translated by ONOS Intent Framework into installable forwarding rules. We repeatedly pushed 2000 intents till the controller halted. Fig. 7 illustrates the capacity of single controller. For our experiment, as the intent count reached 1600, the controller stopped responding. However, the capacity and performance of controller is dependent upon the configuration of physical machine on which the controller is running. After repeating the experiments number of times we choose 15000 as a threshold value for next ReCSDN experiment on this configuration. Nonetheless, the threshold value can be configured using the Policy Engine of ReCSDN whenever required.

After determining the threshold value, we launched a DDoS attack on SDN controller by pushing unknown flows in the network. We created a three controller ReCSDN cluster and started pushing intents gradually. As we moved from 1000 intents to 40,000 the ReCSDN control plane remained active without performance degradation as shown in Fig. 8. The master ReCSDN controller distributed the load to ensure the continuity of service. We also generated the legitimate traffic on the network during the attack. There were no packet losses and the response time remained consistent throughout.

ReCSDN is capable of provided resiliency not only in case of DDoS attack but also in case of controller failure. It improves the network performance by timely load distribution among the controllers.

VII. CONCLUSION AND FUTURE WORK

A Software Defined Network (SDN) is an emerging network paradigm that provides central control over the network. Although the centralized control is one of the major advantages of SDN, it also brings about many critical concerns including a single point of failure in case of attacks. The central control can also become a bottleneck affecting the network’s overall quality of service.

In this paper we highlighted the security threats specific to centralized control, that is, SDN control plane. We addressed the SDN’s control plane issues of performance bottle neck and single point of failure.

In order to improve the performance and fault tolerance of SDN, we proposed and implemented a resilient framework-ReCSDN. Our proposed solution is not only capable of detecting excessive network traffic coming towards an SDN controller but also provides a mechanism to ensure the continuity of services in case of DDoS attack. ReCSDN uses load balancing strategy to invoke backup controllers in ReCSDN cluster to distribute and manage the load without performance degradation. We performed extensive

experiments by emulating the network using Mininet and implementing ReCSDN on top of ONOS. The experiments prove that the proposed framework provides resiliency and improved performance consistency. Even though, our results are specific to the ONOS controller but the methodology we presented is general and can be applied to any distributed controller platform. In future, we intend to experiment with larger number controllers.

ACKNOWLEDGMENT

This research is supported by Higher Education Commission, Pakistan grants HEC NRPU 5946 – 2017.

REFERENCES

- [1] D. Kreutz, F. M. V. Ramos, P. E. Verissimo, C. E. Rothenberg, S. Azodolmolky, and S. Uhlig, "Software-Defined Networking: A Comprehensive Survey," *Proc. IEEE*, vol. 103, no. 1, pp. 14–76, 2015.
- [2] N. Bawany, J. Shamsi, and K. Saleh, "DDoS Attack Detection and Mitigation using SDN: Methods, Practices, and Solutions," *Arab. J. Sci. Eng. Springer*, Feb. 2017.
- [3] B. N. Astuto, M. Mendon, X. N. Nguyen, and K. Obraczka, "A Survey of Software-Defined Networking: Past, Present, and Future of Programmable Networks To cite this version.," 2014.
- [4] L. Tancevski, "SDN concept: from theory to network implementation," in *Optical Fiber Communication Conference*, 2014, p. W1E.3.
- [5] C. Sieber et al., "Network configuration with quality of service abstractions for SDN and legacy networks," in *2015 IFIP/IEEE International Symposium on Integrated Network Management (IM)*, 2015, pp. 1135–1136.
- [6] S. M. Mousavi and M. St-Hilaire, "Early detection of DDoS attacks against SDN controllers," in *2015 International Conference on Computing, Networking and Communications (ICNC)*, 2015, pp. 77–81.
- [7] D. Kreutz, F. M. V. Ramos, P. Verissimo, C. E. Rothenberg, S. Azodolmolky, and S. Uhlig, "Software-Defined Networking: A Comprehensive Survey," *Proc. IEEE*, pp. 1–62, 2014.
- [8] I. F. Akyildiz, A. Lee, P. Wang, M. Luo, and W. Chou, "A roadmap for traffic engineering in software defined networks," *Comput. Networks*, vol. 71, pp. 1–30, 2014.
- [9] M. Jammal, T. Singh, A. Shami, R. Asal, and Y. Li, "Software defined networking: State of the art and research challenges," *Comput. Networks*, vol. 72, pp. 74–98, Oct. 2014.
- [10] I. Akyildiz, A. Lee, P. Wang, M. Luo, and W. Chou, "Research challenges for traffic engineering in software defined networks," *IEEE Netw.*, no. June, pp. 52–58, 2016.
- [11] X. Wen, Y. Chen, C. Hu, and Y. Wang, "Towards a Secure Controller Platform for OpenFlow Applications," pp. 171–172.
- [12] D. Kreutz, F. M. V. Ramos, and P. Verissimo, "Towards secure and dependable software-defined networks," in *Proceedings of the second ACM SIGCOMM workshop on Hot topics in software defined networking - HotSDN '13*, 2013, p. 55.
- [13] R. Braga, E. Mota, and A. Passito, "Lightweight DDoS flooding attack detection using NOX/OpenFlow," in *IEEE Local Computer Network Conference*, 2010, pp. 408–415.
- [14] S. Shin, V. Yegneswaran, P. Porras, and G. Gu, "AVANT-GUARD: scalable and vigilant switch flow management in software-defined networks," in *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security - CCS '13*, 2013, pp. 413–424.
- [15] B. Wang, Y. Zheng, W. Lou, and Y. T. Hou, "DDoS Attack Protection in the Era of Cloud Computing and Software-Defined Networking," *2014 IEEE 22nd Int. Conf. Netw. Protoc.*, pp. 624–629, Oct. 2014.
- [16] J. M. Dover, "A denial of service attack against the Open Floodlight SDN controller," vol. 21037, no. December 2013.
- [17] Y. Xie and S. Z. Yu, "Monitoring the application-layer DDoS attacks for popular websites," *IEEE/ACM Trans. Netw.*, vol. 17, no. 1, pp. 15–25, 2009.
- [18] R. Wang, Z. Jia, and L. Ju, "An Entropy-Based Distributed DDoS Detection Mechanism in Software-Defined," in *Trustcom/BigDataSE/ISPA*, *IEEE*, Vol. 1, 2015, pp. 310–317.
- [19] T. Chin and X. Mountridou, "Selective Packet Inspection to Detect DoS Flooding Using Software Defined Networking (SDN)," 2015.
- [20] K. Giotis, "Leveraging SDN for Efficient Anomaly Detection and Mitigation on Legacy Networks Malicious traffic growth □ Companies, governments and institutions are," 2014.
- [21] P. Berde et al., "ONOS: Towards an Open, Distributed SDN OS."
- [22] Z. Sanaei, S. Abolfazli, A. Gani, and R. Buyya, "Heterogeneity in Mobile Cloud Computing: Taxonomy and Open Challenges," *IEEE Commun. Surv. Tutorials*, vol. 16, no. 1, pp. 369–392, 2014.
- [23] H. Kim and N. Feamster, "Improving network management with software defined networking," *IEEE Commun. Mag.*, vol. 51, no. 2, pp. 114–119, Feb. 2013.
- [24] A. Martinez et al., "Network Management Challenges and Trends in Multi-Layer and Multi-Vendor Settings for Carrier-Grade Networks," *IEEE Commun. Surv. Tutorials*, vol. 16, no. 4, pp. 2207–2230, 2014.
- [25] T. Benson and A. Akella, "Unraveling the Complexity of Network Management," in *e 6th USENIX Symposium on Networked Systems Design and Implementation*, ser. NSDI'09, Berkeley, CA, USA, 2009, pp. 335–348.
- [26] M. Casado, N. Foster, and A. Guha, "Abstractions for software-defined networks," *Commun. ACM*, vol. 57, no. 10, pp. 86–95, Sep. 2014.
- [27] Ligia Rodrigues Prete, A. A. Shinoda, C. M. Schweitzer, and R. L. S. de Oliveira, "Simulation in an SDN network scenario using the POX Controller," in *2014 IEEE Colombian Conference on Communications and Computing (COLCOM)*, 2014, pp. 1–6.
- [28] N. Gude et al., "NOX: towards an operating system for networks," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 38, no. 3, p. 105, Jul. 2008.
- [29] D. Erickson and David, "The beacon openflow controller," in *Proceedings of the second ACM SIGCOMM workshop on Hot topics in software defined networking - HotSDN '13*, 2013, p. 13.
- [30] A. Shalimov, D. Zuikov, D. Zimarina, V. Pashkov, and R. Smeliansky, "Advanced study of SDN/OpenFlow controllers," in *Proceedings of the 9th Central & Eastern European Software Engineering Conference in Russia on - CEE-SECR '13*, 2013, pp. 1–6.
- [31] J. Medved, R. Varga, A. Tkacik, and K. Gray, "OpenDaylight: Towards a Model-Driven SDN Controller architecture," in *Proceeding of IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks 2014*, 2014, pp. 1–6.
- [32] Z. K. Khattak, M. Awais, and A. Iqbal, "Performance evaluation of OpenDaylight SDN controller," in *2014 20th IEEE International Conference on Parallel and Distributed Systems (ICPADS)*, 2014, pp. 671–676.
- [33] J. Tourrilhes, P. Sharma, S. Banerjee, and J. Pettit, "SDN and OpenFlow Evolution: A Standards Perspective," *Computer (Long Beach Calif.)*, vol. 47, no. 11, pp. 22–29, Nov. 2014.
- [34] A. Auyoung et al., "Corybantic: Towards the Modular Composition of SDN Control Programs."
- [35] "Brocade SDN Controller - Brocade," 2015. [Online]. Available: <http://www.brocade.com/en/products-services/software-networking/sdn-controllers-applications/sdn-controller.html>. [Accessed: 23-Jul-2017].
- [36] S. Shin and G. Gu, "Attacking software-defined networks," in *Proceedings of the second ACM SIGCOMM workshop on Hot topics in software defined networking - HotSDN '13*, 2013, p. 165.
- [37] S. Scott-Hayward, G. O'Callaghan, and S. Sezer, "Sdn Security: A Survey," in *2013 IEEE SDN for Future Networks and Services (SDN4FNS)*, 2013, pp. 1–7.
- [38] ONOS organization, "Use Cases - ONOS," 2015. [Online]. Available: <http://onosproject.org/use-cases/>. [Accessed: 23-Jul-2017].
- [39] A.-S. Ali and L. Peterson, "CORD: Central Office Re-architected as a Datacenter," in *OpenStack Summit*, 2015.
- [40] L. Peterson and A. Bavier, "CORD: CENTRAL OFFICE RE-ARCHITECTED AS A DATACENTER."

- [41] D. Ongaro and J. Ousterhout, "In Search of an Understandable Consensus Algorithm," in Proceedings of USENIX Annual Technical Conference, 2014.
- [42] M. Johns, *Getting Started with Hazelcast*. Packt Publishing Ltd, 2015.
- [43] N. Z. Bawany and J. A. Shamsi, "Application Layer DDoS Attack Defense Framework for Smart City using SDN," *Comput. Sci. Comput. Eng. Soc. Media (CSCESM)*, 2016, pp. 1–9, 2016.
- [44] S. Mousavi, "Early Detection of DDoS Attacks in Software Defined Networks Controller," 2014, 2014.
- [45] Q. Niyaz, W. Sun, and A. Y. Javaid, "A Deep Learning Based DDoS Detection System in Software-Defined Networking (SDN)," Nov. 2016.
- [46] P. Fonseca, R. Bennesby, E. Mota, and A. Passito, "A replication component for resilient OpenFlow-based networking," in 2012 IEEE Network Operations and Management Symposium, 2012, pp. 933–939.
- [47] S. Hong, L. Xu, H. Wang, and G. Gu, "Poisoning Network Visibility in Software-Defined Networks: New Attacks and Countermeasures."
- [48] A. Aleroud and I. Alsmadi, "Identifying DoS attacks on software defined networks: A relation context approach," in NOMS 2016 - 2016 IEEE/IFIP Network Operations and Management Symposium, 2016, pp. 853–857.
- [49] Y. Cui et al., "SD-Anti-DDoS: Fast and efficient DDoS defense in software-defined networks," *J. Netw. Comput. Appl.*, vol. 68, pp. 65–79, 2016.
- [50] F. Botelho, A. Bessani, F. M. V. Ramos, and P. Ferreira, "On the Design of Practical Fault-Tolerant SDN Controllers," in 2014 Third European Workshop on Software Defined Networks, 2014, pp. 73–78.
- [51] M. P. Fernandez, "Comparing OpenFlow Controller Paradigms Scalability: Reactive and Proactive," in 2013 IEEE 27th International Conference on Advanced Information Networking and Applications (AINA), 2013, pp. 1009–1016.
- [52] D. Manual, A. Botta, W. De Donato, A. Dainotti, S. Avallone, and A. Pescap, "D-ITG 2.8.1 Manual," pp. 1–35, 2013.

Detection and Prevention of SQL Injection Attack by Dynamic Analyzer and Testing Model

Rana Muhammad Nadeem¹
Computer Science Department
Govt. Post Graduate College
Burewala, Pakistan

Rana Muhammad Saleem²
Computer Science Department
UAF Sub Campus Burewala
Burewala, Pakistan

Rabnawaz Bashir³
Computer Science Department
Comsats Institute of Information Technology
Vehari, Pakistan

Sidra Habib⁴
Computer Science Department
UAF Sub Campus Burewala
Burewala, Pakistan

Abstract—With the emergence and popularity of web application, threats related to web applications has increased to large extent. Among many other web applications threats Structured Query Language Injection Attack (SQLIA) is the dominant in its use due to its ability to access the data. Many solutions are proposed in this regard that has success in specific conditions. The proposed model is based on the dynamic analyzer model. The proposed model also has certain advantages like wide applicability, fast response time, coverage to large number of techniques of SQL Injections (SQLI) and efficient in term of resource usage.

Keywords—Structured Query Language (SQL); injection attack; request receiver; analyzer and tester

I. INTRODUCTION

It is the information age and information is critical for business process. Web applications are major source of information for business process critical for the survival for any organizations [1]. With the popularity of web applications there is also increase in web application vulnerabilities. Across many types of web vulnerabilities SQL Injection (SQLI) has become the predominant method due to its rewarding nature to have access to the data and due to advances in its techniques. It is observed that SQLI is the most widely used techniques for the web applications [2]. According to Open Web Application Security Project (OWASP) (Organization that ranked the web Applications risks) in SQLIA is the dominant web application security risk as shown in Fig. 1.



Fig. 1. OWASP SQLIA ranking over the years [3].

Due to huge rewarding of having access to the database the SQLIA has become the predominant web application security risks and their technique has become more sophisticated over time [4].

Due to emergence of different sophisticated techniques SQLIA has shown a tremendous increase in its spread to web applications of finance banks, educational institutes, global market and many more [5]. The following Fig. 2 shows the relative spread of SQLIA as compared to other types of Web Vulnerabilities. SQLIA is also among the top when compare the spread or choice of web vulnerability among the intruders.

Percentage of web sites vulnerability by class

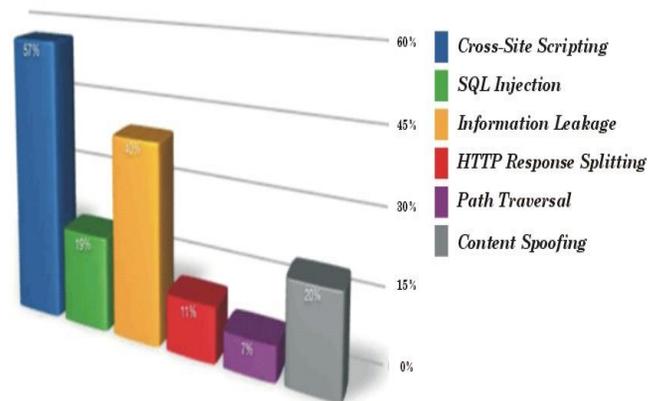


Fig. 2. Volume percentage of web application security risks [6].

For smooth operations of the organization that utilize web applications it is necessary that web applications operate at reasonable level of security. Due to complexities of web technologies and varieties of risks it is not an easy task to save the web applications from intruders and threats [7]. Even sometimes it is very difficult to detect that some serious threat has been occurred [8]. In Fig. 3 to 5 statistics shows the relative difficulty in detection of web application threats.

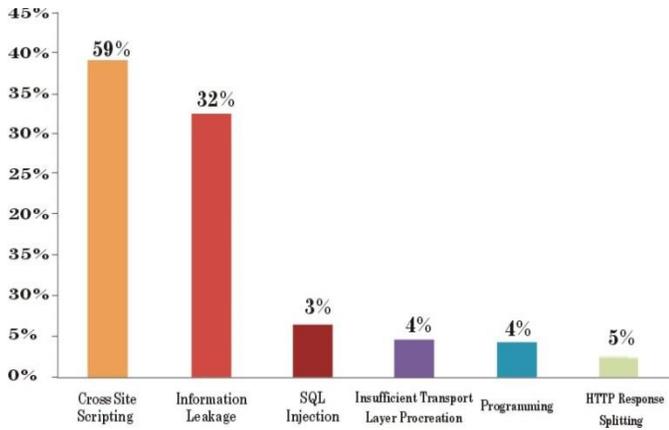


Fig. 3. Percentage of web security risks in web applications [6].

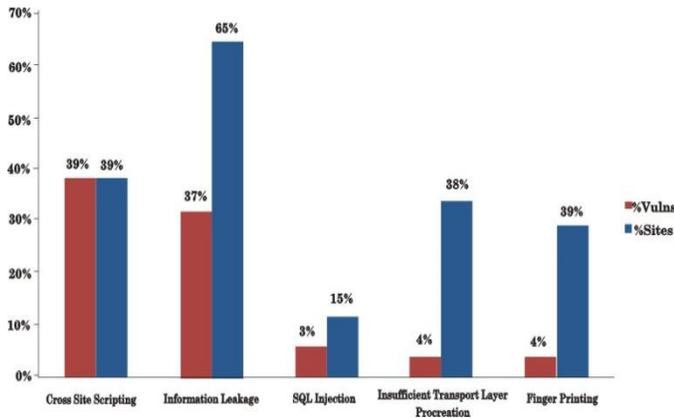


Fig. 4. Probability to detect Vulnerabilities in web application [6].

According to the Web Application Security Consortium (WASC) 78% of web applications are susceptible of security risks and 49% of web applications are susceptible to risks of high level [6].

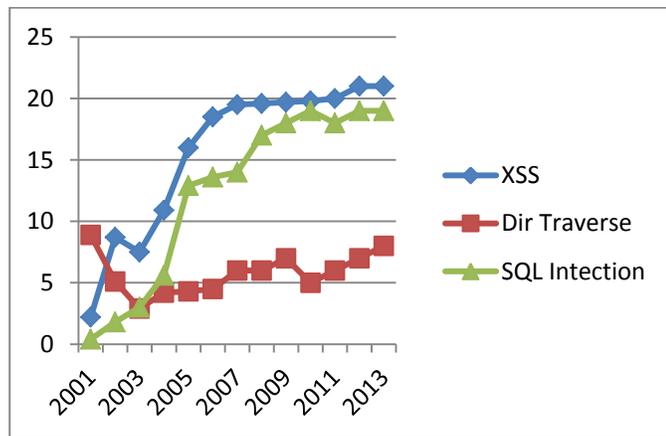


Fig. 5. Growth of web Application vulnerabilities from 2001 to 2010 [9].

SQL Injection Attacks (SQLIA) are among the top of the Input validation attacks and in top five among all the web application security risks [10].

It is important to observe that SQLIA injection Attacks are 30% of total web application security risks due to potential

advantage associated with the SQLIA for the intruders to gain access to the data and much useful information [11]. Due to emergence of needs of more secure web applications it is strongly required that research should focus on the SQLIA and come with a solution that can overcome the problems associated with previous proposed solutions like performance issues, code change and inefficient use of the resources [9].

A. SQL Injection Attack (SQLIA) Process

Data driven web sites are vulnerable to SQL Injection attack where database is a black box in three tier architectures. In this architecture SQL statements are generated in response to HTTP requests [12]. These HTTP request may contain parameters that are used by attackers to produce a query of their interest to have illegal access to the database as shown in Fig. 6.

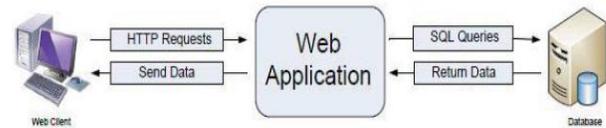


Fig. 6. SQL Injection attack process.

Log In page as shown in Fig. 7 is the most vulnerable for the SQLIA attack and following is the PHP code snippet that produce dynamic query in response to user input [9] as shown in Fig. 8 and 9.

Username:

Password:

Fig. 7. Log In form.

```

// connect to a database
mysql_connect(servername,username,password);
// store user input in the variables collected from the user input login form
$username=$_POST['username'];
$password=$_POST['password'];
// dynamically build the query from the user input
$query="SELECT* FROM tbl_users WHERE username = '$username' AND password = '$password';
// execute a query
$result = mysql_query($query);
if($result)
    return true;
else
    return false;
    
```

Fig. 8. PHP Code snippet to generate dynamic query in response to user input.

SELECT * FROM tbl users WHERE username = ' user_Name AND PASSWORD='pwd' . . ;

Fig. 9. SQL query as a result of code.

In next Fig. 10 at the same form user try to attempt a simple SQLIA to bypass the authentication.

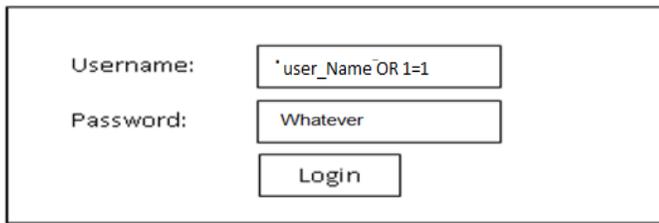


Fig. 10. Simple attempt for SQLIA.

```
SELECT * FROM tbl users WHERE username = 'user_Name OR 1=1 AND password= 'Whatever'
```

Fig. 11. Dynamic generated query in response to above input.

In Fig. 11 attacker try to ignore the password by using the – comment operator as everything would be ignored after the comments operator even the password. In this scenario user name is tried to be true using the OR operator. This, the simple scenario and with different techniques intruders want to add query of their interest to have access to the information of their interest.

B. Techniques of SQL Injection Attack (SQLIA)

1) Tautology Based Attacks

In tautology attack, malicious contents are added using the conditional statement that always evaluate to true. Previous scenario is the perfect example of this attack [13].

Select * from tbl users Where username='rabnawaz' or '1'='1' and password='whatever'

2) Union Attack

In this technique, malicious query is added with the safe query using the UNION keyword [14].

['UNION SELECT pwd FROM user-info WHERE id='abc' and pwd='']

3) Logically Incorrect query Attack

In this type of technique logically incorrect type of query is performed to have information about some structures of the data base to proceeds further [15].

4) Piggybacked Query

Certain delimiters like “;”, “;”, “;” used to join the legitimate query with the illegitimate one [6].

Select * from users where id='rabnawaz' and pwd=''; Drop table users...

5) Alternate Encoding

By changing the coding schema, the illegal query can be bypassed through the filter that tests the legitimacy of the query [16].

6) Inference Attack

Blind and timing techniques are used in inference attacks. In blind attack, a series of the simple queries are performed to have guess about the structure of the data base. In timing attack the query processing time is observed to infer some information presence in the data base.

C. Consequences of SQL Injection Attacks

It has been observed that due to access to the data base SQLIA has become the dominant web application security risks over the last ten years. Database is the very critical for successful operations of any organization. Sensitive information in the database can be used in many ways to serve the attacker purpose [17]. Followings could be the intentions of the attackers to use SQLIA.

To gain information about data base finger prints like type of data base, SQL language used, etc. This information helps the attacker to proceeds or use more sophisticated attacks [18].

- 1) To gain information about user credentials [19].
- 2) To get the database schema [20].
- 3) To extract and modify the data base [1].
- 4) To perform Denial of Services like shutting down the data base, dropping tables, etc. [21].
- 5) Replacements of files with false or tempered information [19].
- 6) Execution of remote commands.
- 7) Shop lifting, account balance change.
- 8) Interacting with underlying operating system.

II. INTRODUCTION TO EXISTING TOOLS

Following's are the major tools available for detection and prevention of SQLIA vulnerabilities:

A. Acunetix

Acunetix web application vulnerability detection scanners that use the XSS black box and Advance SQL injection techniques. It crawls and scans sites and with help of black box and grey box hacking techniques for identification of serious vulnerabilities. Acute nix claimed to detect more than three thousand web application vulnerability including SQL injection Attacks, XSS and host header injections [22].

B. SQLmap

SQLmap is open source analysis tool that automatically detect SQL injection vulnerabilities. It is a powerful tool that has powerful detection engine many niche database penetration features [23].

C. SQLiX

SQLiX is a scanner that crawls and detects SQL Injections. This tool can detect normal and blind SQL injections and there is no need to change the original SQL request [24].

D. Wapiti

Wapiti is web vulnerability scanner for the web application that helps to audit or assess the security of a web application. It uses black box scan that do not scans the code instead use the script and forms where actually injection took place. Wapiti can detect the various techniques of SQLIA [25].

E. Paros

Paros is also a scanner for detection of web vulnerabilities that is java based HTTP/HTTPS proxy. It allows to analysis of the HTTP request with support of spiders, proxy-chaining, XSS, SQL Injection Client certification, etc. [26].

F. Pixy

Pixy is open source tool to detect web application security risks [27].

III. PROPOSED SOLUTION

In this article, a solution is proposed that is based on dynamic analyzer.

A. Proposed Solution Architecture

Proposed model based on the dynamic analyzer that work as user would request the page and that request is received and analyzed to check that request is for pages without vulnerabilities (P') and with vulnerabilities (P), with help of knowledge base. If the user request is for P pages then request is served and if the request is for the P' pages then tester would handle the situation by testing the user request. Tester would generate the possible expected response from the user and user request would be served. On response from the user the response is compared with the expected result and any discrepancy is observed. If the user response is normal then the request is served otherwise user request is rejected and knowledge base is updated for page vulnerabilities and possible rule addition. The complete flow of proposed solution is shown in Fig. 12.

$$P = \sum_{i=1}^n P_i = P_1 + P_2 + P_3 \dots \dots P_n$$

Equation 1: Set of Pages without possible vulnerabilities and P' is the pages where no serve side scripting or not vulnerable.

$$P' = \sum_{i=1}^n P'_i = P'_1 + P'_2 + P'_3 \dots \dots P'_n$$

Equation 2: Pages with possible vulnerabilities

$$R = \sum_{i=1}^n r'_i = r_1 + r_2 + r_3 \dots \dots r_n$$

Equation 3: Set of knowledge base rule

B. Algorithm of Proposed Solution

```

Function Analyzer (Requested_Page)
{
  Mark_Page=Mark(Requested_Page)
  If (Mark_Page is Vulnerable)
  {
    Tester(Requested_Page);
  }
  else
  {
    Serve_Request(Requested_Page)
  }
  Tester(Requested_Page)
  {
    Generate_Expected_Response(Page_Request);
    Serve_Request(Requested_Page);
    Response= Test_Reponse();
    If (Response is Expected)
    {
    }
    else
    {
      Block_Request ();
      Update_Knowledge_Base ();
    }
  }
}
    
```

C. Flow Chart of Proposed Solution

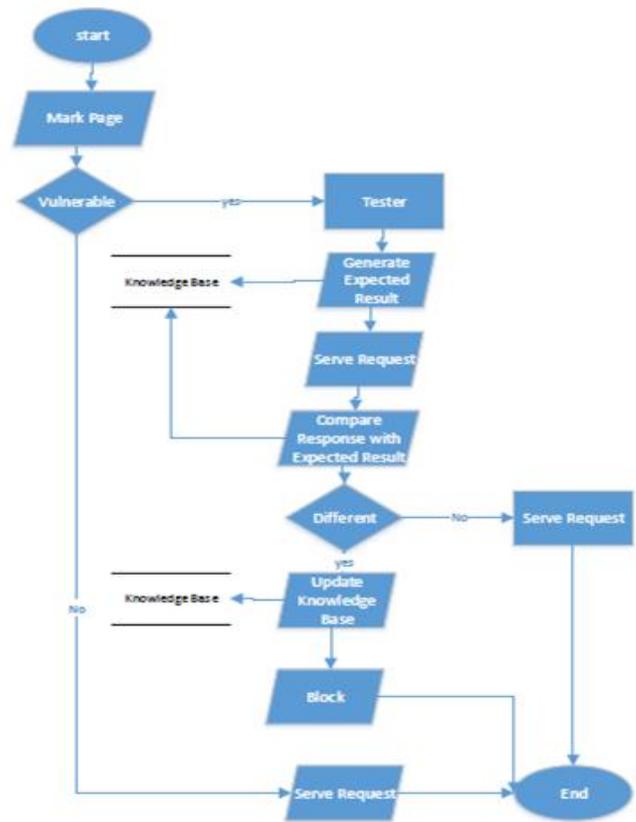


Fig. 12. Flow chart of the proposed solution.

D. Implementation and Evaluation of the Proposed Solution

To evaluate the proposed solution its performance is compared with existing tools described in previous sections. These tools and proposed solutions are applied to detect the SQLIA and block the SQLIA in different types of web application specified in Table 1. These tools are evaluated against different criteria mentioned below.

1) Implementation

Using ASP.Net different classes of web Application are used to evaluate the different tools against the different SQL Injection Attack.

2) Test Scenarios

Following criteria are used to judge the performance of different tools.

- a) No of SQLIA attacks detected
- b) No of SQLIA attacks blocked
- c) Time taken to prevent SQL Injection Attack
- d) Time taken to block SQL Injection Attack
- e) No of types of SQLIA detected
- f) No of types of SQLIA blocked
- g) No of Database supported.

Following dataset is used to evaluate the above-mentioned conditions.

TABLE. I. DATA SET FOR EVALUATION OF DIFFERENT TOOLS AND TECHNIQUES

Applications	No. of Inputs
Portals	100
Classifieds	100
Online Shopping	100
University Database	100
Financial Database	100

E. Evaluation Results

The different evaluation results have been achieved by using above mentioned test scenario as shown in Fig. 13 to 19.

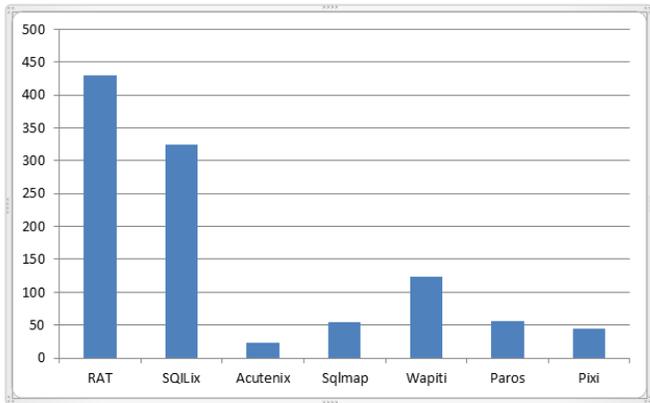


Fig. 13. Number of SQL injection attacks detected.

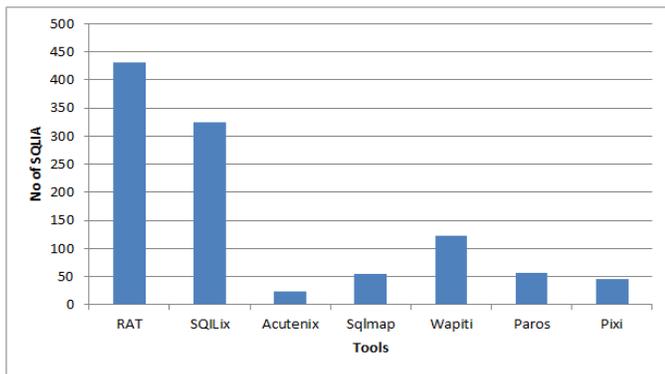


Fig. 14. Number of SQL injection attacks blocked.

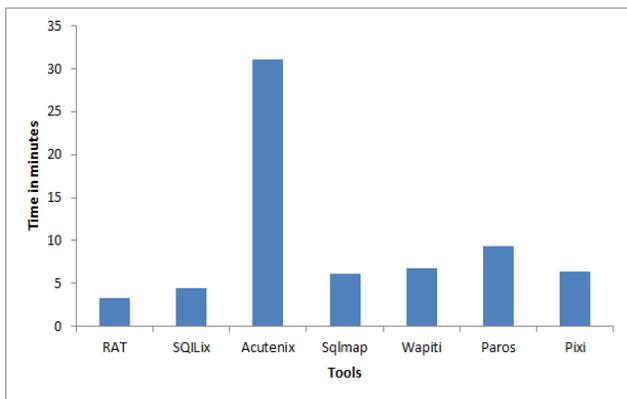


Fig. 15. Average time taken to detect SQLIA.

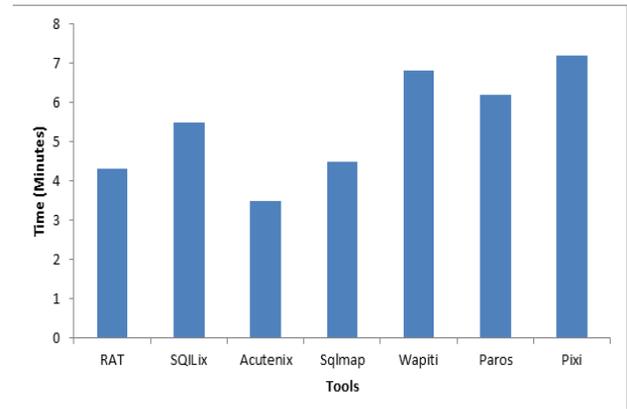


Fig. 16. Average time taken to block SQLIA.

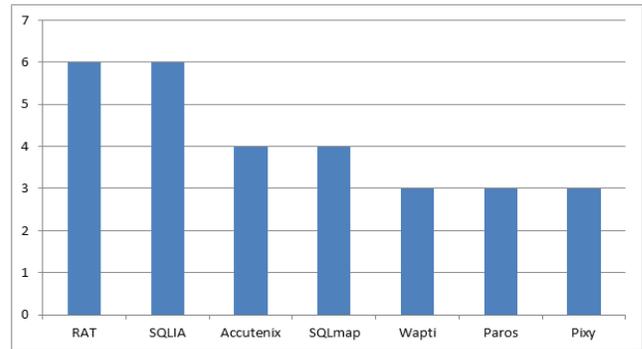


Fig. 17. Number of types of SQL injections techniques detected.

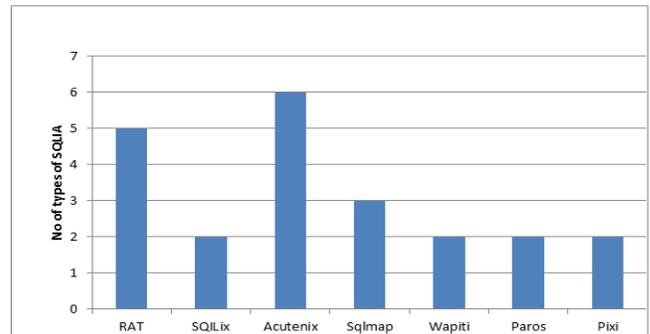


Fig. 18. Number of types of SQL injection techniques blocked.

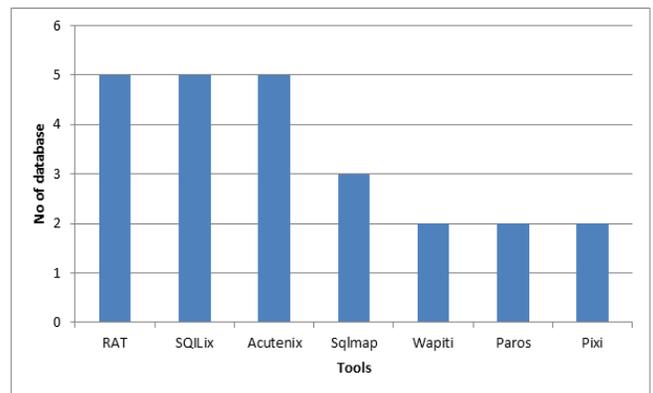


Fig. 19. Number of database supported.

IV. CONCLUSION

SQL Injection Attack has emerged as major threats to web applications. Many solutions were proposed to detect the SQLIA vulnerabilities in web application. Proposed solution based on dynamic Analyzer and tester performed well to detect and block the SQLIA and response time is also excellent as compared to another tool. The proposed solution also needs not to change the source code of the web application and use minimum resources of the system. One major advantage of the proposed solution is that it can handle the advanced SQLIA techniques as knowledge base is updated to handle modern types of threats.

V. FUTURE WORK

The proposed solution use MS SQL analyzer for possible vulnerabilities detections and page marking. The tools need to improve in such a way that any sort of analyzer can be configured for analysis. Knowledge base maintains the techniques and knowledge about different attacks. Knowledge base should be updated using different machine learning approaches.

REFERENCES

- [1] A. Anchlia and S. Jain, "A novel injection aware approach for the testing of database applications," in Proceedings of the 2010 International Conference on recent trends in information, telecommunication and computing ITC, Wasington DC, 2010.
- [2] A. Ciampa, C. A. Visaggio and M. D. Penta, "A heuristic-based approach for detecting sql-injection vulnerabilities in web applications," in In Proceedings of the 2010 ICSE Workshop on Software Engineering for Secure Systems, SESS '10, New York, NY, USA, 2010.
- [3] "https://www.owasp.org," 01 June 2017. [Online]. Available: https://www.owasp.org/index.php/Top_10_2013-Main. [Accessed 12 June 2017].
- [4] A. Kieyzun, P. J. Guo and K. Jayaraman, "Ernst. Automatic creation of sql injection and cross-site scripting attacks," in 31st International Conference on Software Engineering, ICSE '09,, Washington, 2009.
- [5] A. Liu, Y. Yuan, D. Wijesekera and A. Stavrou, "Sqlprob: a proxy-based architecture towards preventing sql injection attacks," in 2009 ACM symposium on Applied Computing, SAC '09, New York, 2009.
- [6] S. Gordeychik, 15 December 2013. [Online]. [Accessed December 2013].
- [7] A. Razaq, A. Hur, N. Haider and F. Ahmad, "Multi-layered defense against web application attacks," in Sixth International Conference on Information Technology: New Generations, Washington, DC, 2009.
- [8] A. Tajpour, M. Massrum and M. Heydari, "Comparison of sql injection detection and prevention techniques," in Education Technology and Computer (ICETC), 2010 2nd International Conference, 2010.
- [9] D. A. Anup Shakya, "A Taxonomy of SQL Injection Defense Techniques," Karlskrona Sweden, 2011.
- [10] A. Tajpour and .. Shooshtari, "Evaluation of sql injection detection and prevention techniques," in Computational Intelligence, Communication Systems and Networks (CICSyN), 2010 Second International Conference, 2010.
- [11] A. Ciampa, C. A. Visaggio and M. D. Penta, "A heuristic-based approach for detecting SQL-injection vulnerabilities in web applications," in Proceeding SESS '10 Proceedings of the 2010 ICSE Workshop on Software Engineering for Secure Systems, New York, 2010.
- [12] A. Tajpour, S. Ibrahim and M. Masrom, "SQL injection Prevnetion and detection Techniques," International Journal of Advancements in Computing Technology, vol. 3, no. 7, pp. 85-91, August 2011.
- [13] B. Indrani and E. Ramaraj., "X-log authentication technique to prevent sql injection attacks," International Journal of Information Technology and Knowledge Management ., vol. 4, pp. 4:323-328,, 2011.
- [14] C. T. M and B. J., "Design considerations for a honeypot for sql injection attacks," in LCN'09, 2009.
- [15] D. Das, U. Sharma and D. Bhattacharyya, "An approach to detection of sql injection attack based on dynamic query matching," International Journal of Computer Applications, vol. 1, no. 25, p. 28-34, February 2010.
- [16] K. Amirathimasebi, S. Jalalinia and S. Khadem, "A Survey of sql injection defence mechanisms," in International Conference Internet Technology and Secured Transactions ICITST 2009, 2009.
- [17] A. Moosa, "Artificial Neural Network based Web Application Firewall for SQL Injection," World Academy of Science, Engineering and Technology, vol. 40, pp. 42-51, April 2010.
- [18] Z. Lijiu, Q. Gu, S. Peng and X. Chen, "D-WAV A Web Application Vulnerabilities Detection Tool Using Characteristics of Web Forms," in Fifth International Conference on Software Engineering Advances (ICSEA), 2010, Nice, 2010.
- [19] Z. Jan, M. Shah, A. Rauf, M. Khan and S. Mahfooz, "Access control mechanism for web databases by using parameterized cursor," in Future Information Technology (FutureTech), 2010 5th International Conference, 2010.
- [20] Xiang Fu and K. Qian, "SAFELI – SQL Injection Scanner Using Symbolic Execution," in Workshop on Testing, Analysis and Verification of Web Software, July 21, 2008.
- [21] M. Cova, D. Balzarotti, V. Felmetzger and G. Vigna, "Swaddler: An Approach for the Anomaly-based Detection of State Violations in Web Applications," 12 December 2013. [Online]. Available: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.127.6909>.
- [22] I. Musacat, "https://www.acunetix.com/blog/docs/blind-sql-injector/," 1 Feburary 2017. [Online]. Available: <https://www.acunetix.com/blog/docs/blind-sql-injector/>. [Accessed 15 June 2017].
- [23] B. Damele A. G. and . S. Miroslav, "http://sqlmap.org/," 12 June 2016. [Online]. Available: <http://sqlmap.org/>. [Accessed 13 June 2017].
- [24] AnirudhAnand, "https://www.owasp.org/index.php/Category:OWASP_SQLiX_Project," 16 March 2014. [Online]. Available: https://www.owasp.org/index.php/Category:OWASP_SQLiX_Project. [Accessed 10 June 2017].
- [25] "http://wapiti.sourceforge.net/," 20 October 2014. [Online]. Available: <http://wapiti.sourceforge.net/>. [Accessed 10 June 2017].
- [26] "http://sectools.org/," 15 December 2015. [Online]. Available: <http://sectools.org/tool/paros/>. [Accessed 05 june 2017].
- [27] J. N., C. Kruegel and E. K. , "Pixy: a static analysis tool for detecting Web application vulnerabilities," Security and Privacy, 2006 IEEE Symposium on, pp. 41-46, 2006.

Normalisation of Technology use in a Developing Country Higher Education Institution

Ibrahim Osman Adam

Department of Accountancy and Commerce
University for Development Studies
Wa, Ghana

Osman Issah

Department of Accountancy and Commerce
University for Development Studies
Wa, Ghana

Abstract—The purpose of this study is to understand how the use of an online course and lecturer evaluation becomes a normalised way of evaluating courses and lecturers in a developing country higher education institution. Extant literature on course and lecturer evaluations has concentrated on the approaches to evaluating courses, lecturers, and its effectiveness and benefits. However, less attention has been paid to how online evaluations become the medium for lecturer and course evaluation. To address this gap, this study used an interpretive case study approach to collect data through semi-structured interviews, documents and participant observation. Data analysis was conducted using hermeneutics and using Normalisation Process Theory as the theoretical lens. The results show that the online evaluation of courses and lecturers is now a normal practice because of participant's investment in the meaning of the online evaluation process, their enrolment in the process and the crucial investment of their actions, feedback during implementation, and use of which ensured the normalization.

Keywords—Course and lecturer evaluation; Higher Education Institution (HEI); Normalisation; Normalisation Process Theory (NPT)

I. INTRODUCTION

The purpose of this paper is to understand how the use of an online course and lecturer evaluation process becomes normalised evaluation process in a higher education institution (HEI). Course and lecturer evaluation is the most commonly used method of assessing a course and lecturer effectiveness because it offers important opportunities for feedback and development [1], and has been routinely used in academic institutions to inform curricular change and assess lecturer's performance [2]. Whilst course and lecturer evaluations may be paper-based or virtual, it has largely been conducted in class at the end of the academic semester through the use of paper evaluation forms [3] in many developing countries. Despite the limitations of high financial cost, waste of time and problems with analysis the paper-based evaluations are widespread [4], [5] in developing countries. As a result, many HEIs in developing countries are migrating to online evaluations where students use online forms [1], [4]. A lot of studies on the course and lecturer evaluation in HEIs [6]-[8] have largely been quantitative with less qualitative studies. Apart from the lack of qualitative research in the area, there exist some knowledge gaps in understanding how the use of online evaluations becomes a normal practice, especially in HEIs in SSA.

Paper-based evaluations have been cited to have some problems such as the vulnerability of lecturers influencing students on the day of the evaluation by their presence or otherwise [9]. This is because the presence of the lecturer when the students are conducting the evaluation may create an intimidating environment which may influence what the students put on the evaluation forms. Also, the security of the evaluation form is a problem. This is because lecturers can pick and choose which forms to take forward as part of the evaluation. Unfortunately, very little attention has been given in the literature to understanding how the new ways of evaluating courses and lecturers become the norm. The research question, therefore, concerns how the implementation and use of such new technology can be normalised in HEIs.

The study is conducted in a University in Ghana (Herein referred to as UNID). Ghana was selected for a number of reasons. First, the researchers are Ghanaians and are faculty staff in Ghanaian Universities and believe that their knowledge about the country and the University set up as well as their social networks there could facilitate gaining research access.

The rest of the paper is aimed at how this question is answered using a coalescing of Normalisation Process Theory (NPT) and empirical evidence derived from an interpretive case study approach. The following sections are organized as follows. Section 2 examines the literature on the course and lecturer evaluations. The theoretical foundation and the methodology are discussed in Sections 3 and 4, respectively. Section 5 presents the case study description and the analysis and discussion of the findings are presented in Section 6. Finally, Section 7 concludes the paper and outlines its contribution, implications and suggestion for further research.

II. BACKGROUND OF COURSE AND LECTURER EVALUATIONS

Student evaluations of courses and lecturers are also one of the most controversial and highly-debated measures [42]. Nonetheless, they are still widely used and many have argued that there is no other option that provides the same sort of quantifiable and comparable data [40].

Largely, course and lecturer evaluations are used to make personnel decisions in terms of hiring tenure, promotion, and so on and this is based in part on a student's evaluation of lecturer's teaching effectiveness. The qualitative responses are also used as a feedback for lecturers and other teaching

support offices to ensure improved teaching and course development. In [43], author cautions against the use of instruments not specifically designed to provide feedback for this purpose, and that separate instruments should be designed to provide summative and formative feedback, respectively.

Much has been written about the problems with the course and lecturer evaluations. Educational scholars have examined issues of bias and concerns regarding the statistical reliability of evaluations of lecturers and have questioned their ability to accurately gauge the teaching effectiveness of staff. In addition, some have argued that the feedback provided by course and lecturer evaluations does not effectively promote change in lecturer's behaviour. However, a significant majority of researchers consider student evaluations to be a useful measure of the instructional behaviour that contributes to teaching effectiveness [40], [41].

Whilst student evaluations have largely been conducted physically using paper-based evaluation forms, many educational institutions are migrating to online evaluations [15], [38]. HEIs are leveraging on the advantages that an online evaluation could bring. This is because student evaluations are seen as a very important yardstick in the retention, promotion and tenure decisions of lecturers in higher educational institutions [43]. With this importance, many academic staff is concerned that a migration to an online evaluation may have effects that can change the whole evaluation process. Lower response rates by students have been cited as one of the effects [6]. Though there is less research on online course and lecturer evaluations and its implementation in the developing world, several institutions in the developed world have successfully implemented online student evaluations [9], [10].

Despite the widespread implementation in the developed world many higher educational institutions and academic staff still question their value [11], [12]. Several advantages have been cited in the literature for the migration of physical evaluation of lecturers to online evaluations. The quick turnaround of student evaluations is one of the mainly cited advantage. This provides academics more rapid feedback to refine the curricula or the overall educational design [11]-[13] cites the ease for students to write their reflections of the learning experiences on a keyboard than by hand.

The research on student course and lecturer evaluation is widely dominated by literature on students' experiences [9], [13]. However, in a recent study by [12] on the migration from paper to online evaluations, it was found that most lecturers still preferred traditional paper-based evaluations. The lecturer's perception was that the paper-based methods resulted in higher response rates. Others have mentioned lower response rates in online evaluations because it involves out-of-class time and students can be distracted and not remember to fill the form or they may simply choose not to do it [14]. Technical glitches in accessing the online forms are discussed in [15], and the issue of anonymity of online responses are discussed in [9], [16].

III. NORMALISATION PROCESS THEORY

NPT provides a framework for understanding how a new intervention becomes or not becomes part of normal practice [17] by examining how social processes affect the new ways of working [18]. NPT seeks to understand the dynamics of embedding a practice in an institution as part of implementing, integrating and using this practice to influence business processes [19]. NPT provides a set of tools that explains the processes through which new or modified practices of thinking, enacting, and organizing work is operationalized in institutional settings' [20].

Normalisation is the work that actors do as they engage with some ensemble of activities (that may include new or changed ways of thinking, acting, and organizing) and by which means it becomes routinely part of already existing, socially patterned, knowledge and practices [20, p. 540]. The basic tenet of the theory is that when organisations are confronted with a change they must find ways of accommodating that change. The theory, therefore, aims to develop an understanding of the process by which an information system is implemented, accepted and used.

In particular, the theory is concerned with three issues:

1) *Implementation*: These are the processes of bringing a practice or practices into action.

2) *Embedding*: The processes through which practices become or do not become are routinely part of the everyday work of individuals and groups.

3) *Integration*: The processes by which practices are not only reproduced but are sustained in organisational processes.

This means that first, work practices are normalized when people work either individually or collectively to endorse them. Secondly, the processes involved in enacting a practice is enhanced or inhibited through the operation of some social processes through which human action is expressed. These processes are called generative mechanisms and are coherence, cognitive participation, collective action, reflexive monitoring [18]. Third, the production and reproduction of a practice require continuous investment to ensure its sustainability [20].

The four generative processes underpin the three core issues and are discussed below:

- *Coherence*: This is the process of understanding that allows or prevents the use of a practice by participants [17]. Coherence involves four sub-components which are *differentiation*, *communal specification*, *individual specification* and *internalisation*.
- *Cognitive participation*: This involves anything that allows or prevents users' involvement in a practice [17]. It involves the work undertaken to engage the participants who are part of the new intervention. It is this engagement that will position the actors for collective action [21]. Cognitive participation covers four sub-components. These are *initiation*, *enrollment*, *legitimation* and *activation*.

- *Collective action*: This involves the work performed by individual or groups [17]. Achieving this goal may include resistance, subversion or reinvention from the users [21]. The components of this mechanism are *interactional workability, relational integration, skill-set workability and contextual integration*.
- *Reflexive monitoring*: This promotes or inhibits users' understanding of the effects of a practice [17]. The collective action and the outcomes should be continuously evaluated, both formally and informally, by participants engaged in the implementation processes [20]. The components of reflexive action are; *systemisation, communal appraisal, individual appraisal and reconfiguration*.

According to [20], NPT is a theory of action and is different from other theories because it seeks to explain how

innovations are becoming routine in an organisation by focusing on individual and collective learning. In the literature robust social science theories already explain individual differences in attitudes to new technologies and practices (e.g. Theory of Planned Behaviour) [22], the flow of innovations through social networks (e.g. Diffusion of Innovations Theory) [23], reciprocal interactions between people and artefacts (e.g. Actor Network Theory) [24]. NPT, therefore, explains phenomena not well covered by existing theories. NPT may shed light on why some IS normalise while others do not [18] and as [25] puts it, NPT offers a coherent framework of propositions that may provide useful insights in the way systems become normalised within organisations. The sub-components mentioned above and they mean in this study is expatiated in Table 1 below.

TABLE I. NPT ANALYTICAL FRAME FOR THE IMPLEMENTATION OF THE ONLINE COURSE AND LECTURER EVALUATION. ADAPTED FROM [26].

Coherence (Sense-Making Work)	Cognitive Participation (Relationship Work)	Collective Action (Enacting Work)	Reflexive Monitoring (Appraisal Work)
Differentiation: Participants understood the difference between the manual and the online course and lecturer evaluation	Initiation: The participants are working to drive the change forward	Interactional workability: The work that participants did with each other to operationalize the online course and lecturer evaluation	Systemisation: When students attempt to determine how effective and useful the online course and lecturer was for them and others
Communal: specification: Respondent's had a shared understanding of why the online course evaluation was introduced and the expected benefits	Enrolments: Participants (re)organise themselves and others in order to contribute to the online course and lecturer evaluation collectively	Relational Integration: The knowledge work that participants did to build accountability and maintain confidence in a set of practices and in each other as they use them	Communal appraisal: When staff attempt to appraise the worth of the online course and lecturer evaluation
Individual Specification: Actions that help students, lecturers and administrative staff understand their specific tasks/responsibilities	Legitimation: Participants believing it is right to be involved and that they can make a valid contribution	Skill set workability: Describes the distribution and conduct of the practices as they were operationalized in the real world	Individual procedures: When staff attempt to appraise the effects of them and the context in which they were set
Internalisation: Participants understand the importance of the online course and lecturer evaluation	Activation: Participants collectively define the actions and procedures needed to sustain the online course and lecturer evaluation	Contextual integration: Refers to the incorporation of the online course and lecturer evaluations within the context of the university	Reconfiguration: Appraisal work that may lead to attempts to redefine procedures

IV. RESEARCH METHODOLOGY

The study uses an interpretive case study method [27], [28]. Following the interpretive tradition [29] means that the philosophical assumptions underlying this study are a subjective epistemology and the ontological belief that reality is socially constructed. These assumptions supported the researchers to understand the behaviour of students and staff in social and organisational contexts by assuming that as they interact they create subjective meaning through their interactions [30].

Multiple data collection methods [27] through documents interviews and participant observation were used. Interviews were the primary data source because it is through this that the researchers' best accessed the interpretations of participants' actions and the events taking place [31]. Valuable insight was also being gained from the analysis of research conducted by the AQAU. These secondary data supported the preparation for interviews and helped the researcher to learn about the key

stakeholders, technical details and other organisational issues. However, the access to documents and staff of the AQAU as well as students and lecturers was duly guided by the appropriate procedures for gaining access [32] such as endorsements and familiarity with some interviewees [39]. Purposive sampling was used to identify interviewees [33]. The number of interviewees was not limited to a particular number but continued until a number was arrived at heuristically. This meant the researcher only stopped interviewing when it was realised that nothing new was being gathered from the interviews. Semi-structured interviews were used because of its flexibility to explore emerging themes during the interview. Each interview lasted between 20 to 25 minutes. In all 19 participants were interviewed initially but 2 follow-up interviews was conducted. However, two key participants were contacted so many times over the period of the research. The interviews were digitally recorded and transcribed using NVivo 10 as the data management tool.

There is a thin line separating the data collection and the data analysis. This is because the two belong to an iterative process and the results can help guide the other. The data collected was analysed using hermeneutics. This is because hermeneutics is consistent with the interpretive qualitative study. This analysis technique was used because it is consistent with the type of data that was collected. Hermeneutics is primarily concerned with making meaning of textual data by providing a set of concepts to help a researcher interpret and understand the meaning of the text or multiple texts. Hermeneutics is the view that the understanding of a research phenomenon is derived through an iterative process between the understanding of the interdependent meaning of the parts and the whole [27]. The process of data analysis involves a number of stages involving the stages of familiarisation, identification of a thematic framework, indexing and interpretation [34], [35]. The first stage involved the familiarisation with the data. This was done by going through the interview transcripts several times. This enabled me to fill in the gaps I had missed either through the transcription or during the interview. Some facts were also cross-checked with my interview notes.

This was followed by identifying the themes and concepts from the transcripts and putting this into a framework. The framework relied on the four main constructs of the NPT (Coherence, Cognitive participation, Collective Action and Reflexive monitoring) as the main codes and the related sub-constructs to guide the sub-themes.

V. CASE STUDY DESCRIPTION

The fieldwork for this study was conducted at UNID in close collaboration with its Academic Quality Assurance Unit. The AQAU oversees the standards of academic work in the university by supporting developing world-class human resources and capabilities to meet national development needs and global challenges through quality teaching, learning, research and knowledge dissemination. The AQAU has several mandates one of which is to conduct student evaluation of courses and lecturers. The evaluations are conducted on every course and teaching staff every semester. The first researcher was attached to the unit for a period of six months as part of a PhD experiential learning. The majority of the first researcher's time was spent on the campus interacting with students, lecturers and administrative staff of the AQAU.

VI. ONLINE EVALUATION OF LECTURERS

The University conducted a paper-based evaluation of courses and lecturers for a long time until 2014 when it was stopped. During this time the University ensured that all departments had a procedure in place for dealing with student evaluation of courses and that this was clearly communicated to students. All students taking a course completed a questionnaire that was prepared by the AQAU and administered by the department through the lecturer. The questionnaire had two main sections; an objective portion where students selected the most suitable option and a subjective or written portion for comments from the students. Students were required to complete both sections of the evaluation form.

The online evaluation was developed by the AQAU in conjunction with a UNID IT Department (ITD). Whilst AQAU handled the administrative aspect of determining the content of the evaluation form and how the data will be analysed, ITD was involved in the technical aspect of developing the web page and making sure that this was up and running during the period of the evaluation. When evaluations are completed, ITD extracts the data and hands it over to AQAU for analysis. However, any feedback received by AQAU from the use of the system is communicated to ITD for improvement in subsequent evaluations.

The online evaluation of lecturers was provided through the University's website. An active link is provided about three weeks to the end of the semester at the homepage of the University website. A click on the link directs students to a login page where a student number and pin is required. After logging in the student is presented with options to choose his/her college first and then department. After this, the courses the student has registered for the semester, the name of the lecturer, the academic year and the semester are populated in a drop down list. After choosing these, the student then proceeded to start the evaluation which was in three main parts; course evaluation, lecturer evaluation and comments and suggestions for improvement.

The systems have evolved from the previously scannable forms. When the online system was first implemented the students were granted access to log into the systems using a security token in order to enable them to conduct the evaluation. When this was implemented, the response rate was quite high but in the subsequent evaluation, it dropped drastically. When the AQAU interacted with some students it was realised that students were sceptical about conducting the evaluation because of fear of getting their identification (IDs) tied to the evaluation.

In the following semester, the feedback of the students was taken into consideration and the token and log in approach were abandoned. An open link was then provided at the homepage of the University website where the students could just visit and start filling out the form without having to log in with the IDs. However, this approach was saddled with issues such as multiple evaluations by students without being noticed. Even a lecturer who feared that he may be evaluated negatively could visit the page and evaluate himself multiple times in order to raise his/her score. To ensure that students did not feel that their identification is tied to the evaluation, AQAU and ITD organised a demonstration session with a cross section of students who were very conversant in the way this type of technology works. This was to allay the fears of the students. Other problems were student complaints that they could not find their courses in the online system. Some students complained of missing course codes, course names and lecturer names. Also, it was reported that the system did not provide avenues for lecturers who had co-taught a course to be evaluated individually.

VII. ANALYSIS OF FINDINGS AND DISCUSSION

In terms of coherence, the participants had a shared understanding of why the online course evaluation was introduced and what are the expected benefits (communal

specification). This was much clearer with the staff at the academic quality unit. However, this shared understanding is not being translated into use. The unit responsible for this exercise is investing efforts to make the system better but this is coupled with declining use by students. There are no organised fora to discuss declining student use with students or section of them. There is a clear gap in these shared understanding being translated into shared involvement and participation by students and use by the majority of the students.

The three categories of participants demonstrated different levels of understanding of the aim, objectives and expected benefits of course and lecturer evaluation generally and the online evaluation in particular (differentiation). The participants understood that the online evaluation differed from the paper-based course and lecturer evaluation. This understanding was reflected in the attitude of students towards conducting the evaluation online as opposed to the former paper-based one; an attitude that reflected both scepticism and interest. This is probably because the online evaluation is new and this may be the semblance of the acceptance and use of a newly introduced technology [36], [37].

Overall, several initiatives have been taking place to ensure that students, lecturers and administrative staff understand their specific tasks/responsibilities in the online evaluation process (individual specification). These are evidenced by e-mails sent to academic staff to remind them to alert students to evaluate courses and flyers posted at student's hostels and lecture halls. Despite these efforts, some difficulties are still being encountered to ensure a full uptake of the online evaluation by students. Some students still claim they have not heard of the online evaluation before whilst others have exhibited the lack of seriousness to conducting this exercise.

All the participants understood the importance of the online course and lecturer evaluation exercise (internalisation). The unit emphasised the need for the University have migrated from the paper-based to the manual, citing the cost cutting reasons, faster processing times and the need to ensure that students get the level of quality of teaching they expect when they come to the University. The reason that the new online evaluation systems would result in improved evaluation of course and lecturers was seen as an important reason to go online. However, the online evaluation required a new approach to making it work. One student participant felt unsure of how the University wanted to achieve the level of quality they needed if student participation is getting lower since and the evaluation processes goes on unnoticed by many students. However, some students think this exercise is a platform they can use to get back at their lecturers who they feel have delivered poorly. Even though the benefits and the importance of the online course and lecturer evaluation are popular among staff, low student patronage could still occur in future evaluations. Students may have to use the new systems as if the paper-based evaluation method never existed. This is because the continuous comparison of the old and the new continues to draw the line indicating how the current method is not being patronised.

In terms of cognitive participation, the determination of the University to build and sustain the new online evaluation process is high (initiation). Equally, the other participants are aware of how the process can be driven forward despite the barriers that are being encountered. The University has been using several approaches to ensure that students are properly (re)organised to engage in the evaluation at the end of every semester (enrolment). This is directed at ensuring that courses/lecturers are collectively evaluated by the students. Though the use of the online evaluations has not reached the level compared to the paper-based one, there is the need to improve communication between all participants in order to ensure that there is common footing with regards to the idea behind the migration from the paper-based to the online evaluation. Student's participants were fully knowledgeable of the need to evaluate their courses and lecturers though majority are still not doing it. The university understands this and is cognisance of the fact that the evaluation process cannot be mandatory for students. A lecturer indicated that:

Through my lecturers, When I start my lectures for the semester I tell the students that they will have to evaluate the course and the lecturer at the end of the semester. During my last lecture, I remind the students to go online and do it if I don't forget.

This is evidence of how lecturers are trying to get students to be involved in the exercise, however, a comment by a lecturer that:

I am not sure lecturers were involved in any way in this new process, at least I never heard of this movement until the end of one semester when the Quality Assurance sent an email to the staff list about how the semester evaluation was being done.

This shows how lecturers were not involved in the migration to the online evaluation. Both lecturers and students indicated that they were not involved in the migration from the paper-based evaluation to the manual one (legitimation). They believed it would have been proper to be involved and that they could have made some critical contributions. Lecturer 1 indicated that:

I am not involved much. When it was paper-based we used to support the process by taking the evaluations forms to the last lecturer for student to go through the exercise but now we just have to tell the students in the last lecturer to go online and do it and that's it.

The lecturers and students did not play any part in defining the actions and procedures needed for the online course and lecturer evaluation to work (activation). This was solely decided by the University. The online evaluation has evolved to its current form because of the University's commitment to ensuring that it succeeds. In terms of collective action, the main issue was concerned with all what the participants did in order to involve each other to ensure that the online course and lecturer evaluation is operationalized (interactional workability). To ensure that the new practices of evaluating courses/lecturers online is fully enacted, the University is met with difficulties in fully getting student participation, the lecturers do their best to let students participate in the exercise

by giving reminders during their lectures and some of the students want to their colleagues to participate by informing them. However, there is still resistance on the part of students because of the fear of being victimised if they leave negative comments for a lecturer who is under performing. Apart from the lack of fear, involvement by students is sometimes deterred by problems in the online evaluation systems itself as reported by one student that:

Well I can't say I completed the process because I needed to evaluate 6 lecturers and I ended evaluating only 1, I didn't even go halfway. This is because either I couldn't find the course code or the lecturers name was not there.

In terms of relational integration, it was clear that the conduct of the online evaluation process is distributed among the participants in the University (skill-set workability). The University does this through flyers to students, and emails to the academic staff. Though the University is aiming at incorporating the online course and lecturer evaluations within the context of the university there is much they can do to let students fully embrace it (contextual integration) though there is a limit to what can be done to ensure that students must do it.

In terms of reflexive monitoring, the participants were able to determine how effective and useful the online course and lecturer was for them (systemisation); along the way, the University to appraise the worth of the online course and lecturer evaluation (communal appraisal). Collectively this can be done by all participants. From the lecturers and students, the general feedback is for the University to intensify its awareness among students. The individual participant's made attempts to appraise the effects of the evaluation process (individual procedures). The general feedback from lecturers and students were that the university needs to improve the awareness of this exercise especially in the student community. Other complaints were that of missing course codes, course names and lecturer names. Also, it was reported that system did not provide avenues for lecturers who had co-taught a course to be evaluated individually.

The online evaluation process is a constantly evolving one and continuously needs to be appraised so that the procedure that can make it a success are redefined (reconfiguration). Among all the participants interviewed the general call is for the university to intensify awareness campaign especially among students to ensure that there is high patronage of the system. The University admits it is working to ensure this. Apart from these, other issues such as the difficulty in locating courses have been identified as a key problem. There were complaints of the possibility of students or lecturers going online to evaluate a course and lecturer multiple times since no log or security is required before the exercise can be done.

The research findings show that the University virtualised its course and lecturer evaluations because of several reasons. The cost of printing, administering and processing the survey results were the most compelling reasons. The time savings in terms of administration and processing of feedback was also key [1]. The time savings related to students too because they would have some time in their own time to reflect on their answers before they submit [9]. Whilst the issue of instilling

objective evaluations in students was cited and that the online evaluations will eliminate the possibility of students feeling intimidated in the presence of their lecturer, some students complained about their anonymity in the online process for fear that their identities can easily be tracked in an online system as compared to the paper-based one they have been used to.

The theoretical foundation (NPT) in this study was partly supported by the data collected. There was the presence of the four generative mechanisms of the NPT suggesting that the implementation may have been completed. This is because it is so in some respects. The virtualisation of the paper-based evaluation is completed. However, feedback at the end of each semester is fed back to improve the next semester evaluation process. This is evidence in another respect that the implementation though completed is an evolving process. In its current form, the reflexive monitoring dimension is currently low an indication that the feedback process needs to be intensified for the current evaluation systems to evolve into a better one and also make it sustainable. The presence of the reflexive monitoring dimension is evidence of the possible sustainability of the new system.

The paper-based evaluation was transformed to the online one through a process of creating an understanding of the system by the students who are the key users. However, this was revealed by the data to be low. Engagement of students and lecturers were absent in major aspects of the migration to online. The collective action of the students to conduct the evaluations was high and the lastly not much feedback is received to improve and make the systems more sustainable. The migration to a virtual course and lecturer evaluation was found to be slow especially the awareness need to ensure a full-scale uptake of the online evaluation when it was first introduced a couple of years ago. This confirms the data that whilst coherence is widespread cognitive participation and collective action among students is low.

The effect of the virtualisation of the course and lecturer evaluations was varied. The low response rate is the key effect of the new systems though the issue of convenience and ease of use for of students to conduct the evaluation was also evidenced in the data.

To improve, sensitisation among the student community is key. However, should be directed at reminders for students to conduct the evaluation instead of sensitisation on the objectives of the exercise or why the systems are being introduced. This is because students are fully aware of the objectives and the reasons why the online systems were introduced. Other efforts to encourage student involvement could be through instituting mechanisms such as vouchers to encourage them to complete the evaluation or a feedback process to understand why students are not fully participating. Mechanism should also be put in place to ensure the full involvement in any changes to the process by deepening stakeholder engagement and activation.

VIII. CONCLUSIONS

The study investigated the normalisation of an online course and lecturer evaluation in an HEI. The implementation

of the online course and lecturer evaluation though complete is an evolving process and this is supported by NPT that transformations such as this do not have a complete end point for the implementations process since the systems continue to evolve through constant feedback and update. The study contributes to both IS and HEI literature as an attempt to offering rich insight into how a newly introduced technology can become the normal way of evaluating courses and lecturers in an HEI in a developing country context. It also offers implications for research and practice. For research, the study enjoins IS scholars to move beyond an examination of migration from physical to virtual platforms per se or the introduction of a new technology as a panacea to its normal use and adoption and explore how new technology become a routine use. This research is, however, limited by its single case study nature in one developing country HEI but the findings provide insight into how NPT can be used to explain the normalisation of a technology use. Another limitation is the small number of participants in the study; however, they represented the whole University's participant in the course and lecturer evaluation process and this small sample provided very rich textual data for the study. Future research can compare the experience of different HEIs in different developing countries in order to account for contextual issues.

REFERENCES

- [1] Riskey, A., Vaughan, E., & Murphy, M. (2015). Online student evaluations of teaching: what are we sacrificing for the affordances of technology? *Assessment & Evaluation in Higher Education*, 40(1), 120-134.
- [2] Hatfield, C. L., & Coyle, E. A. (2013). Factors that influence student completion of course and faculty evaluations. *American Journal of Pharmaceutical Education*, 77(2).
- [3] Capa-Aydin, Y. (2014). Student evaluation of instruction: comparison between in-class and online methods. *Assessment & Evaluation in Higher Education*(ahead-of-print), 1-15.
- [4] Morrison, K. (2013). Online and paper evaluations of courses: a literature review and case study. *Educational Research and Evaluation*, 19(7), 585-604.
- [5] Spooen, P., Brockx, B., & Mortelmans, D. (2013). On the Validity of Student Evaluation of Teaching The State of the Art. *Review of Educational Research*, 83(4), 598-642.
- [6] Rienties, B. (2014). Understanding academics' resistance towards (online) student evaluation. *Assessment & Evaluation in Higher Education*, 39(8), 987-1001. doi: 10.1080/02602938.2014.880777
- [7] Shevlin, M., Banyard, P., Davies, M., & Griffiths, M. (2000). The validity of student evaluation of teaching in higher education: love me, love my lectures? *Assessment & Evaluation in Higher Education*, 25(4), 397-405.
- [8] Struyven, K., Dochy, F., & Janssens, S. (2011). Explaining students' appraisal of lectures and student-activating teaching: perceived context and student characteristics. *Interactive Learning Environments*, 20(5), 391-422. doi: 10.1080/10494820.2010.500084
- [9] Dommeyer, C. J., Baum, P., Hanna, R. W., & Chapman, K. S. (2004). Gathering faculty teaching evaluations by in-class and online surveys: their effects on response rates and evaluations. *Assessment & Evaluation in Higher Education*, 29(5), 611-623. doi: 10.1080/02602930410001689171
- [10] Nulty, D. D. (2008). The adequacy of response rates to online and paper surveys: what can be done? *Assessment & Evaluation in Higher Education*, 33(3), 301-314.
- [11] Bennett, T., & De Bellis, D. (2010). The move to a system of flexible delivery mode (online v. paper) unit of study student evaluations at Flinders University. *Management issues and the study of initial changes in survey, volume, response rate and response level. Journal of Institutional Research*, 15(1), 41-53.
- [12] Crews, T. B., & Curtis, D. F. (2011). Online course evaluations: Faculty perspective and strategies for improved response rates. *Assessment & Evaluation in Higher Education*, 36(7), 865-878.
- [13] Stowell, J. R., Addison, W. E., & Smith, J. L. (2012). Comparison of online and classroom-based student evaluations of instruction. *Assessment & Evaluation in Higher Education*, 37(4), 465-473.
- [14] Laubsch, P. (2006). Online and in-person evaluations: A literature review and exploratory comparison. *Journal of Online Learning and Teaching*, 2(2), 62-73.
- [15] Anderson, H. M., Cain, J., & Bird, E. (2005). Online student course evaluations: Review of literature and a pilot study. *American Journal of Pharmaceutical Education*, 69(1), 34-43.
- [16] Layne, B. H., DeCristoforo, J. R., & McGinty, D. (1999). Electronic versus traditional student ratings of instruction. *Research in Higher Education*, 40(2), 221-232.
- [17] Finch, T. L., Rapley, T., Girling, M., Mair, F. S., Murray, E., Treweek, S., . . . May, C. R. (2013). Improving the normalization of complex interventions: measure development based on normalization process theory (NoMAD): study protocol. *Implementation Science*, 8(1), 43.
- [18] May, C. (2006). A rational model for assessing and evaluating complex interventions in health care. *BMC health services research*, 6(1), 86.
- [19] Sooklal, R., Papadopoulos, T., & Ojiako, U. (2011). Information systems development: a normalisation process theory perspective. *Industrial Management & Data Systems*, 111(8), 1270-1286. doi: doi:10.1108/02635571111170794
- [20] May, C., & Finch, T. (2009). Implementing, embedding, and integrating practices: an outline of normalization process theory. *Sociology*, 43(3), 535-554.
- [21] May, C. R., Mair, F., Finch, T., MacFarlane, A., Dowrick, C., Treweek, S., . . . Rogers, A. (2009). Development of a theory of implementation and integration: Normalization Process Theory. *Implement Sci*, 4(29), 29.
- [22] Ajzen, I. (1991). The theory of planned behavior. *Organizational behavior and human decision processes*, 50(2), 179-211.
- [23] Rogers, E. M. (1995). *Diffusion of innovations*. New York.
- [24] Latour, B. (2005). Reassembling the social-an introduction to actor-network-theory. *Reassembling the Social-An Introduction to Actor-Network-Theory*, by Bruno Latour, pp. 316. Foreword by Bruno Latour. Oxford University Press, Sep 2005. ISBN-10: 0199256047. ISBN-13: 9780199256044, 1.
- [25] Elwyn, G., Légaré, F., van der Weijden, T., Edwards, A., & May, C. (2008). Arduous implementation: does the Normalisation Process Model explain why it's so difficult to embed decision support technologies for patients in routine clinical practice. *Implement Sci*, 3(1), 57.
- [26] Alharbi, T., & Carlström, E. (2014). Implementation of person-centred care: management perspective. *Journal of Hospital*
- [27] Myers, M. D. (2013). *Qualitative research in business and management*: Sage.
- [28] Myers, M. D. (2015). *Commentaries on methodological practice. Formulating Research Methods for Information Systems*, 1.
- [29] Walsham, G. (2006). Doing interpretive research. *European Journal of information systems*, 15(3), 320-330.
- [30] Orlikowski, W. J., & Baroudi, J. J. (1991). Studying information technology in organizations: Research approaches and assumptions. *Information systems research*, 2(1), 1-28.
- [31] Walsham, G. (1995). Interpretive case studies in IS research: nature and method. *European Journal of information systems*, 4(2), 74-81.
- [32] Feldman, M. S., Bell, J., & Berger, M. T. (2004). *Gaining access: A practical and theoretical guide for qualitative researchers*: Rowman Altamira.
- [33] Silverman, D. (2013). *Doing qualitative research: A practical handbook*: SAGE Publications Limited.
- [34] Ritchie, J., Lewis, J., Nicholls, C. M., & Ormston, R. (2013). *Qualitative research practice: A guide for social science students and researchers*: Sage.

- [35] Spencer, L., Ritchie, J., & O'Connor, W. (2003). Carrying out qualitative analysis. *Qualitative research practice: A guide for social science students and researchers*, 219-262.
- [36] Davis, F. D. (1986). A technology acceptance model for empirically testing new end-user information systems: Theory and results. *Massachusetts Institute of Technology*.
- [37] Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, 319-340.
- [38] Davison, E., & Price, J. (2009). How do we rate? An evaluation of online student evaluations. *Assessment & Evaluation in Higher Education*, 34(1), 51-65.
- [39] Shenton, A. K., & Hayter, S. (2004). Strategies for gaining access to organisations and informants in qualitative studies. *Education for Information*, 22(3), 223-231.
- [40] Falchikov, N. (2013). *Improving assessment through student involvement: Practical solutions for aiding learning in higher and further education*: Routledge.
- [41] Gaillard, F. D., Mitchell, S. P., & Kavota, V. (2011). Students, faculty, and administrators' perception of students' evaluations of faculty in higher education business schools. *Journal of College Teaching & Learning (TLC)*, 3(8).
- [42] Hobson, S. M., & Talbot, D. M. (2001). Understanding student evaluations: What all faculty should know. *College teaching*, 49(1), 26-31.
- [43] Kember, D., & Ginns, P. (2012). *Evaluating teaching and learning: A practical handbook for colleges, universities and the scholarship of teaching*: Routledge.
- [44] Wright, R. E. (2006). Student evaluations of faculty: Concerns raised in the literature, and possible solutions. *College Student Journal*, 40(2), 417.

Design and Simulation of a Novel Dual Band Microstrip Antenna for LTE-3 and LTE-7 Bands

Abdullah Al Hasan

Electronic and Telecommunication
Engineering
International Islamic University
Chittagong
Chittagong, Bangladesh

Mohammad Shahriar Siraj

Electronic and Telecommunication
Engineering
International Islamic University
Chittagong
Chittagong, Bangladesh

Muhammad Mostafa Amir Faisal

Electronic and Telecommunication
Engineering
International Islamic University
Chittagong
Chittagong, Bangladesh

Abstract—Long Term Evolution (LTE) is currently being used in many developed countries and hopefully will be implemented in more countries. An antenna operating in LTE-3 band can support global roaming in ITU Regions 1 and 3, Costa Rica, Brazil and partially in some Caribbean countries and antenna operating in LTE-7 band are appropriate for global roaming in ITU regions 1, 2 and 3. An antenna operating at both the bands will make the place taken by the antenna in a device into half and allow roaming in all the regions mentioned above. The geometry of the current available antenna operating in LTE-3 and LTE-7 bands has a considerably large size. A dual band microstrip antenna operating in LTE-3 and LTE-7 bands is proposed in this work with notable size reduction. The proposed antenna simulation shows resonant frequencies at 1.88GHz and 2.55GHz with return loss below -10dB that covers both LTE-3 and LTE-7 bands. Design and simulation of the proposed antenna is done by IE3D Zeland software. This proposed antenna is suitable for global roaming in ITU regions 1, 2 and 3, which cover most of the world telecom network.

Keywords—Long Term Evolution (LTE); microstrip; dual band; u-slot

I. INTRODUCTION

Microstrip Patch Antenna (MPA) is the most popular antenna in today's wireless transmission technology due to its incomparable characteristics. Microstrip antenna is embodiment of a conducting patch normally copper on a dielectric material having a ground plane on the other side. The microstrip antenna has the luxury of being low profile, low-priced, easy to fabricate and modify, flexible in shape with rectangular, circular, triangular, elliptical or any shape needed, and can support dual or multiband frequency operation providing linear or circular polarization [1].

The Dual Band characteristic of microstrip patch antenna with u-slot has been established by many authors [2]. In this research work a microstrip dual band antenna is developed which consists of a rectangular patch on top of the dielectric substrate and over which a u-shaped slot has been cut to create dual band antenna.

Long term evolution (LTE) is the cutting edge technology introduced in the wireless communication system to be the next big thing for at least ten years from now on. LTE is the successor of the third generation technology of 3GPP by

fulfilling the requirement of fourth generation. LTE provides broadband internet in cellular or any devices. LTE data speed is ten times faster than existing third generation technologies. The two main advantages of LTE are that it covers more area and provide faster speed in wireless environment. Users of LTE don't need to be at home or office to experience broadband internet, just needed to be in the LTE coverage area [3].

LTE have many bands; where bands 1 to 31 are Frequency Division Duplex (FDD) and 33 to 48 and above are Time division duplex (TDD). The most used bands of LTE are LTE-3 and LTE-7 [3]. A single element antenna covering these two bands gives coverage to the most of the telecom world and will support roaming and discard the need of using two antennas, hence reducing size of the device.

In recent days LTE supported technologies are being implemented globally and therefore, the need of antennas supporting LTE bands are increasing day by day. From this perspective, designing an antenna that will support two different LTE bands will allow roaming in different parts of the world which is much needed. Networks using the LTE-3 and LTE-7 bands will be available in ITU region 1, 2 & 3 and they are already employed in most of the developed countries [3]. The goal of this proposed work is to design and simulate a single element dual-band micro-strip antenna for LTE-3 and LTE-7 bands which can be used in near future for roaming.

II. LITERATURE REVIEW

Extensive research is ongoing to develop single element microstrip antenna with multiband operation. The antenna proposed in [4], [5] cover only one band either LTE-3 or LTE-7, although some of them are multiband or dual band. The antenna dimensions in [6]-[9] are considerably large. The geometry of the antenna proposed in [10]-[11] is quite complex and can only be used in base stations. The VSWR reference value taken [12] is 2.75 which is not good. The literature review facilitates the knowledge of the current available antenna literatures dealing with LTE-3 and LTE-7 bands. There is room available to improve the performance of the antenna through developing an antenna which covers both LTE-3 and LTE-7 bands with good return loss, small and simple geometry and ease of fabrication.

III. ANTENNA DESIGN METHODOLOGY

Overall methodology in designing the antenna is given below:

- Step 1: Developing a single band antenna.
- Step 2: Simulation of the single band antenna.
- Step 3: Introducing u-slot to obtain dual band operation.
- Step 4: Simulation of the dual band antenna.
- Step 5: Finalizing the dual band antenna.

IV. DESIGN OF THE SINGLE BAND MPA

A microstrip single band antenna consists of a rectangular patch on top of the dielectric substrate. In designing the antenna, Roger Corporations substrate RT Duroid 5880 with dielectric constant 2.2 is used. The length and width of the patch are determined by using (1)-(5) [13].

$$W = \frac{C}{2f_r} \sqrt{\frac{2}{\epsilon_r + 1}} \quad (1)$$

Where, C = Velocity of Light

Effective dielectric constant is given by,

$$\epsilon_{eff} = \frac{\epsilon_r + 1}{2} + \frac{\epsilon_r - 1}{2} \left(1 + \frac{10h}{W}\right) \quad (2)$$

Where,

ϵ_{eff} = Effective dielectric constant,

ϵ_r = Dielectric constant of substrate,

h = Height of dielectric substrate,

W = Width of the patch.

For a given resonance frequency f_r , the effective length is given by,

$$L_{eff} = \frac{C}{2f_r \sqrt{\epsilon_{eff}}} \quad (3)$$

The actual length of the patch is given by,

$$L = L_{eff} - \Delta L \quad (4)$$

Where,

$$\Delta L = 0.412h \frac{(\epsilon_r + 0.3) \left(\frac{W}{h} + 0.264\right)}{(\epsilon_r - 0.258) \left(\frac{W}{h} + 0.8\right)} \quad (5)$$

Using the above formulas and using a dielectric constant of $\epsilon_r=2.2$, with thickness of the substrate of 1.6mm, the geometry of single element single band microstrip antenna is shown in Fig. 1 and the basic parameters are given in Table 1.

The return loss of the single band antenna is given in Fig. 2 and it can be seen that the antenna resonates at 1.75GHz and operating with return loss value less than -10dB.

TABLE I. PROPOSED DUALBAND ANTENNA DIMENSION

Parameters	Value
Frequency (f)	1.8GHz
Substrate Thickness	1.6mm
Dielectric Constant	2.2
Length (L)	65.8808
Width (W)	71.05
Polarisation	Linear
Probe Fed Position (F)	29.9404mm

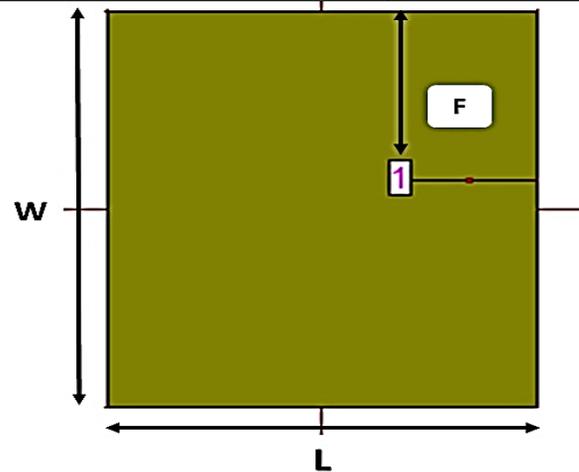


Fig. 1. Antenna geometry.

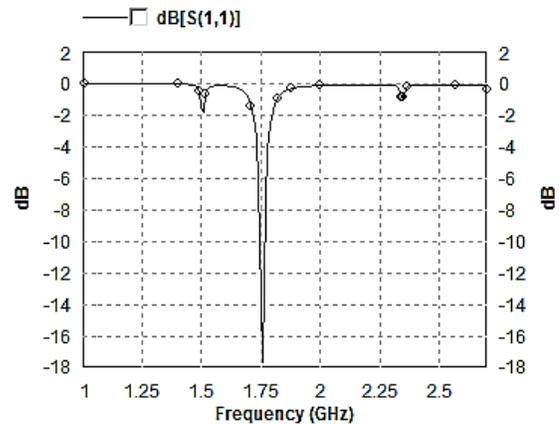


Fig. 2. Return loss Vs. Frequency of single band antenna.

V. ANTENNA CONFIGURATION

The dual band microstrip antenna is developed by cutting the slots of different shapes like U-slot, V-slot, pair of rectangular slots and step slots, etc. Hence the antenna designers need to adjust the dimensions and the position of the slots. The geometry of the dual band rectangular microstrip antenna is shown in Fig. 3. It is constructed on the substrate having dielectric constant (ϵ_r) 2.2 and thickness (h) 1.6 mm. The dimensions of the proposed dual band antenna are given in Table 2 and the geometry of the proposed antenna given in Fig. 3.

TABLE II. PROPOSED DUALBAND ANTENNA DIMENSION

Parameters	Value(mm)
Length (L)	65.8808
Width (W)	71.05
Slot Length Horizontal (L _s)	44
Slot length Vertical (W _s)	40
Slot Thickness (t)	2.5
Probe fed position (F)	27.9404

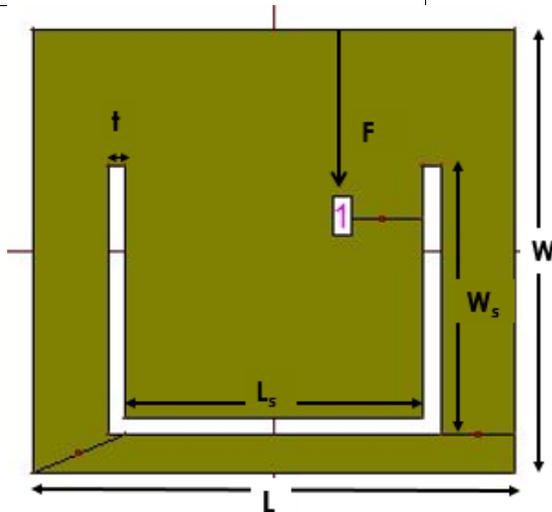


Fig. 3. Antenna geometry.

The proposed structure is simulated using IE3D ZELAND software. After getting the geometry of single band antenna, slots are cut in the geometry to achieve dual band operation. Then, from the simulation, a result is achieved which is not necessarily good enough for resonant frequency. As a consequence, optimization on the dual band geometry is applied. After the optimization process, it can be observed that the antenna work suitably for the desired resonant frequencies.

VI. SIMULATION RESULTS AND DISCUSSION

A. Return Loss Plot

The reflected power of an antenna is determined from the value of the antenna return loss. The return loss, S11 plot of the dual band antenna is given in Fig. 4. The s11 parameter value of dual band antenna is below -10dB at both LTE-3 and LTE-7 bands. The antenna is resonant at 1.88GHz and 2.55GHz. Antenna is observed to be transmitting more than ninety percent of the excited power.

The VSWR value of the proposed antenna is given in Fig. 5 where it can be seen that the value is between 1 and 2 in both bands meaning that the antenna operates efficiently in LTE-3 and LTE-7 bands.

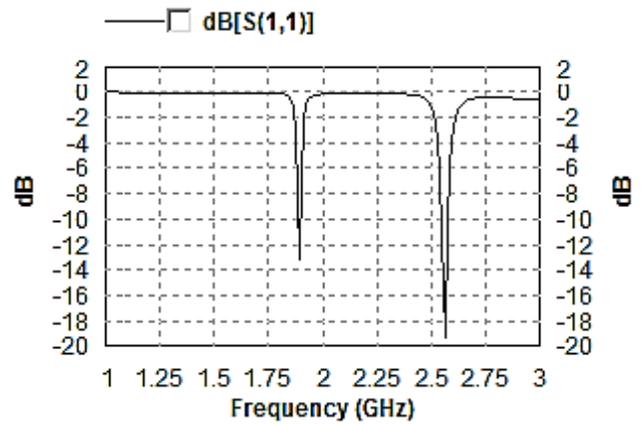


Fig. 4. Return loss Vs. Frequency plot of dual band antenna.

B. VSWR Plot

VSWR shows how much power is reflected back from the antenna towards the source. If the VSWR value is 1 that means all of the given power to the antenna is transmitted. An antenna will be considered a good one if its VSWR value is between 1 and 2. Antenna with VSWR value greater than 2 is not a good antenna.

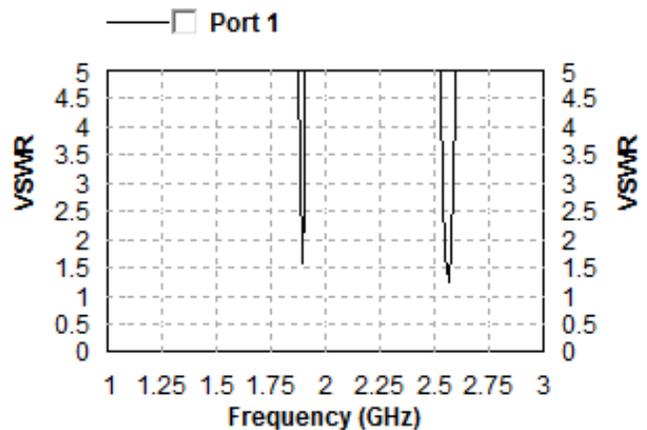


Fig. 5. VSWR vs. Frequency plot of the dual band antenna.

C. Average Current Distribution

The average current distribution provides the information about radiating side and non-radiating side of the antenna. Typically, antenna resonance occurs at half wavelength length. The current distribution of the microstrip antenna is known to be congested in the middle and gradually lesser at edges of the patch. The congested current can be seen by yellowish color in the middle, where the antenna is excited by the probe fed and less current density at the border of the patch indicated by the blue color. The radiating side and non-radiating side are also known to be acting as length and width, respectively.

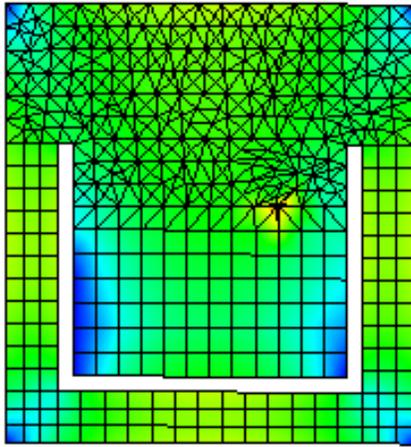


Fig. 6. Average current distribution at 1.88GHz.

The average current distribution in Fig. 6 shows that the antenna is working with maximum congested current at the middle of the vertical side of the patch except at the corners of the patch. This gives the information that the vertical side is the radiating side of the basic single element MPA.

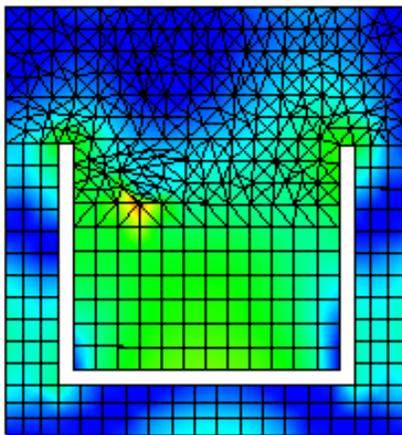


Fig. 7. Average current distribution at 2.55GHz.

Fig. 7 shows that at the second band of 2.55GHz, the slot edges produce the radiation with maximum current distributed inside the two arm of u-slot.

D. Vector Current Distribution

Vector current distribution provides the information about the distribution and flow direction of the current on the conductor patch. The polarization of the antenna can easily be seen from the information provided by the vector current distribution. In Fig. 8, the polarization of the antenna is clearly linear polarization as current follows a linear path on the surface of the conductor. And the maximum current is found to be congested in the middle of the antenna and minimum at the border of the conductor. This figure represents the vector current distribution of the dual band antenna at 1.88GHz.

And in Fig. 9 the vector current distribution shows that the current is mainly distributed around the two arms of the u-slot with current flowing shows that the polarization of the MPA at 2.66GHz is nonlinear polarization.

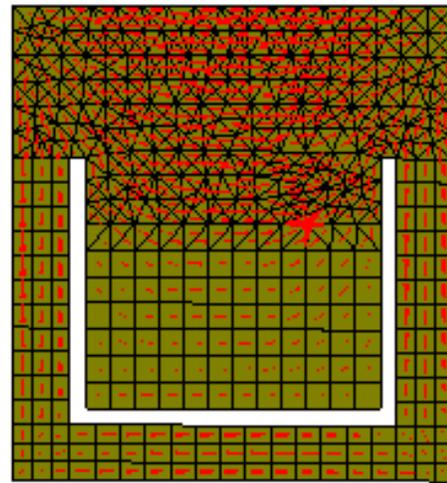


Fig. 8. Vector current distribution at 1.88GHz.

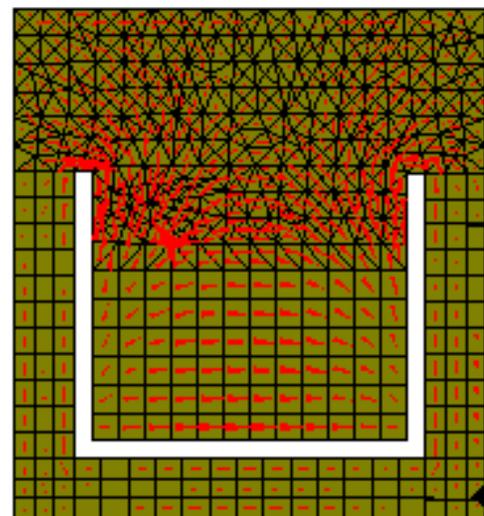


Fig. 9. Vector current distribution at 2.55GHz.

E. 2D Radiation Pattern

The necessity of 2D radiation pattern arises due to its vital role to grasp how the antenna is radiating in 3D. The 3D radiation pattern cannot be properly shown on a 2D surface. So it is very important to show how the antenna is radiating in 2D to perceive how it is actually radiating in 3D surface. In 2D radiation pattern of microstrip antenna, there will be no radiation in lower half and the radiation should be half circular as 2D radiation pattern takes 0-degree and 90-degree to show its radiation beam width.

2D radiation pattern of dual band in Fig. 10 and 11 are almost the same indicating that proposed antenna provides a good radiation pattern. The radiation of the proposed antenna is nearly half circular which is quite good. There is no radiation in lower half as the ground plane attenuates the signal downwards. The 2D radiation pattern of microstrip patch antenna should be half circular and no radiation should exist in underneath, this characteristic gives microstrip antenna advantage when the antenna is incorporated with printed circuit board.

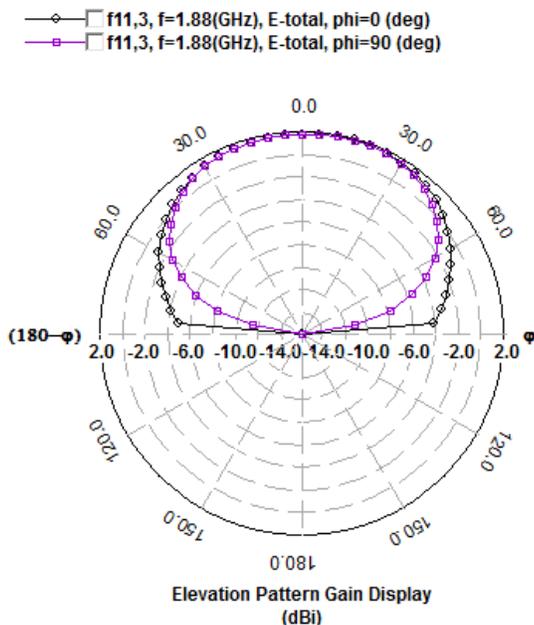


Fig. 10. 2D Radiation Pattern at 1.88 GHz.

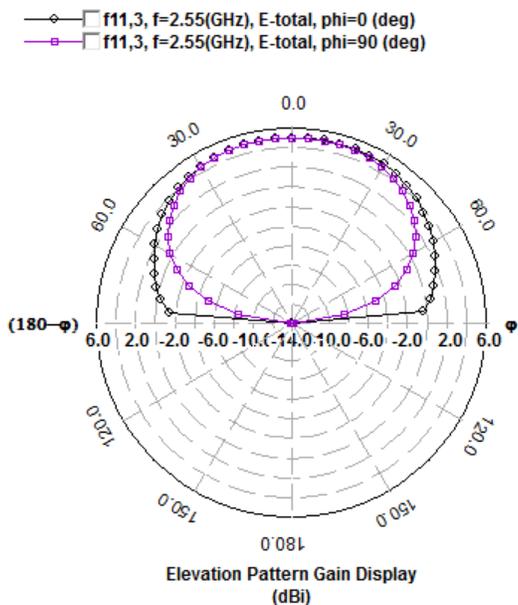


Fig. 11. 2D Radiation Pattern at 2.55 GHz.

F. Final Results

The final results of the propose antenna are given in Table 3. Proposed antenna shows good return loss at both LTE-3 and LTE-7 bands with S11 values less than 10dB. The VSWR of the proposed antenna is less than 1.75 in both bands which describe that the antenna reflects less than 10 percent of the given power while radiating more than 90 percent. The gain of the antenna is in conformity to work perfectly and directivity of the proposed antenna shows that it has a good directional performance. Antenna efficiency is very good at both resonant frequencies having around 74 percent at LTE-3 and 84% in LTE-7. Also the radiation efficiency of the designed antenna is nearly 80 percent in both LTE-3 and LTE-7 bands.

TABLE III. FINAL RESULTS OF THE PROPOSED ANTENNA

Parameters	Value	Value	Standard Value [13]
Resonate frequency	1.88 GHz	2.55GHz	As per need
Returnloss (S11)	-12dB	-15.6dB	Below -10db
VSWR	1.73	1.41	Less than 2
Gain	4.48 dBi	5.17dBi	6-9dBi for MPA
Directivity (dBi)	5.82 dBi	5.93dBi	5-8dBi for MPA
Antenna Efficiency	73.78%	83.97 %	70%
Radiation Efficiency	79.65 %	86.86 %	70%

VII. CONCLUSION

A dual band u-slot microstrip antenna for LTE-3 and LTE-7 bands has been successfully designed in a single patch with 50Ω probe feed. The antenna simulation gives us dual band LTE-3 and LTE-7 operations with return loss below -10dB. The VSWR value of the simulated antenna is less than 1.75 which is in conformity with standard values; which means that the antenna operates efficiently with reflected power of less than ten percent. Antenna performance efficiency is nearly 80% in both bands and the antenna is properly working in the designated bands of LTE-3 and LTE-7 which are 1.8GHz and 2.6GHz. The designed and simulated antenna is unique in feature of its dual band characteristics because it operates in LTE-3 and LTE-7. Although there are some multiband antennas which cover two bands, there is hardly any dual band antenna which covers only these two bands.

REFERENCES

- [1] Anuj Mehta, "Microstrip Antenna", International Journal of Scientific & Technology Research, Volume 4, Issue 03, March 2015.
- [2] R. Bhalla1 and L. Shafai, "Resonance Behavior Of Single U-slot Microstrip Patch Antenna", Microwave and Optical Technology Letters, 2002.
- [3] Christopher Cox, "An Introduction to LTE: LTE, LTE-Advanced, SAE and 4G Mobile Communications", wiley and son ltd, 2012
- [4] Habib M.S., Rafiqul I.M., Abdullah K., Jakpar M.J, "U-Slot Rectangular Patch Antenna for Dual Band Application." In: Sulaiman H., Othman M., Othman M., Rahim Y., Pee N. (eds) Advanced Computer and Communication Engineering Technology. Lecture Notes in Electrical Engineering, vol 315. Springer, Cham(2015),
- [5] M. O. Katie, M. F. Jamlou, H. Lago and S. S. AL-Bawri, "Slots-loaded dual-band elliptical polarized antenna," IEEE International RF and Microwave Conference (RFM), Kuching, pp. 190-193, 2015
- [6] W. S. Chen, Y. Chi, F. S. Chang and C. Y. Hsu, "Coupled-fed LTE antenna design for tablet applications," 2016 IEEE 5th Asia-Pacific Conference on Antennas and Propagation (APCAP), Kaohsiung, pp. 141-142, 2016.
- [7] H. W. Badri, H. Zairi and A. Gharsallah, "Design of a Dual-Band antenna for GSM, UMTS, WLAN, LTE and Wi-MAX applications," IEEE 15th Mediterranean Microwave Symposium (MMS), Lecce, pp. 1-3, 2015.
- [8] M. Madani Fadoul, T. A. Rahman, and A. Moradikordalivand, "Novel Planar Antenna for Long Term Evolution (LTE)," International Journal of Information and Electronics Engineering vol. 4, no. 1, pp. 59-61, 2014.
- [9] Md. H. Haroun, HAYad, J. Jomaa "Design a tri-band Microstrip slot antenna for LTE applications," 978-1-4799-4129-2/15/\$31©2015 IEEE
- [10] J. Zhang, X. Q. Lin, L. Y. Nie, J. W. Yu and Y. Fan, "Wideband Dual-Polarization Patch Antenna Array With Parallel Strip Line Balun

- Feeding,” in IEEE Antennas and Wireless Propagation Letters, vol. 15, no., pp. 1499-1501, 2016.
- [11] Y. Cui, F. Li, Y. Pan, Y. Fan and R. Li, "A novel dual-polarized broadband planar antenna for base stations," 2015 IEEE International Symposium on Antennas and Propagation & USNC/URSI National Radio Science Meeting, Vancouver, BC, 2015, pp. 1964-1965.
- [12] F. Ahmed, Y. Feng, R. Li "A Multiband Multiple-input Multiple-output Antenna System for Long Term Evolution and Wireless Local Area Networks Handsets," IJE TRANSACTIONS B: Applications Vol. 29, No. 8, (August 2016).
- [13] Balanis, Constantine A. Antenna Theory: Analysis and Design. Hoboken, NJ: John Wiley, 2005.

Mobile Learning Application Development for Improvement of English Listening Comprehension

In Rural Primary School Students of Grade 1, 2 and 3 of Pakistan

Zahida Parveen Laghari

Department of computer science
Isra University
Hyderabad, Pakistan

Hameedullah Kazi

Department of computer science
Isra University
Hyderabad, Pakistan

Muhammad Ali Nizamani

Department of computer science
Isra University
Hyderabad, Pakistan

Abstract—Trend towards English language learning has been increased because it is considered as Lingua franca i.e. language of communication for all. However students of Pakistan are behind in this pace, especially rural elementary students. In rural areas there is crucial need to get assistance in their own curriculum after school because mostly they do not find anyone to help at home. M-Learning (Mobile Learning) assists learning anywhere and anytime. This ubiquitous power of M-Learning helps in after school programs and education in rural areas. The aim of this study is to develop M-Learning application for improvement of English listening comprehension in rural primary school students. This study developed English learning application based on Listening Comprehension, which embeds English curriculum of Sindh Textbook board for grade 1, 2 and 3. This study took the form of an after-school program in a village in Pakistan. There were 45 students of grade 3 from rural primary school of Pakistan selected as participants. Since developed application is based on recognition and memorization of information, so that knowledge and comprehension level of cognitive domain from Bloom's taxonomy were selected for choosing the type of evaluation questions. On the basis of those question types, EGRA (Early Grade Reading Assessment) test is used for evaluation. This test was conducted on two experimental groups and one control group and the results of the groups were compared to one another. The results confirm that English M-learning applications can become helpful tool for students who live in rural areas where they face problems in learning of their English curriculum, since their relatives are not capable to teach them as accordingly.

Keywords—Mobile Learning (M-Learning); Early Grade Reading Assessment (EGRA); English as Secondary Language (ESL); Automatic Speech Recognition (ASR); Personal Digital Assistance (PDA)

I. INTRODUCTION

English is a core curriculum which is taught to students from grade 1 in Pakistan. In spite of that it is not getting prosperity in primary level rural schools of Pakistan [1]. Rural students are unable to speak English because they are just paying attention to passing the examination instead of learning the language as a tool for communication. In rural areas of Pakistan [2] the failure ratio in the English language as compared to other subjects is very high and unfavorable. English curriculum is not implemented properly in rural primary level schools due to general challenges of external

factors (family background, economic condition, mother tongue, un-trained teachers, traditional atmosphere of village, exam oriented learning and copying in exams) in education system [3]-[5]. Thus rural students are not competent compared to urban students in English language [6], because their families work in agriculture and they focus on earning instead of learning.

At the same time, M-Learning is considered as inspiring approach of twenty first century as it comes with new trend of learning [7], where learner's Smartphone is embedded with M-learning applications. Learning ESL (English as Secondary Language) [8] through a mobile device can provide such flexibility in the improvement of education and it provides opportunity to educate every one and constraint of time slots and locations are detached. M-learning has great impact on ESL learning [9] particularly for those, who live in rural areas. Learning applications integrated on mobile devices, offers benefits to students of all levels and ages. Students have been promoted by mobile learning applications both in and out of teaching space [10]. These Applications are based on famous theories of language learning to improve ESL [11], so the students have opportunity to learn on demand.

Listening acts a significant role in everyday communication and educational process. It has been investigated that second language acquirement must require listening comprehension, quality of listener and listening comprehension purely influenced by listener variables (vocabulary knowledge, meta cognition, working memory and auditory discrimination). Listening comprehension mode enables learner [12] to gain English language by listening well. For ESL learning there should be huge amount of listening provided to listener and teacher must encourage students to engage in concentrated listening in class. This method enables students [13] to understand the meaning, pronunciation, intonation and the change in language flow. For the improvement of English Curriculum's Comprehension there is need to provide audio input to students for achieving Language learning [14], because listening is basic language skill in language learning. Listening comprehension can influence on listeners' [15] cognitive process and enhance his/her efficiency, utility and success in English learning achievement. English curriculum can be proficient when it could be taught as language and learning of language needs colossal introduction of auditory [16]. Therefore it can be

believed that similar outcomes can be replicated with M-learning application that are based on listening comprehension methodology which targets to rural primary students and the application embeds their own English curriculum.

II. RELATED WORK

There are numerous M-learning applications developed for mobile devices which help in learning ESL, some of them are discussed below.

1) *Speaking English 60 Junoir*: It was based on ASR (Automatic Speech Recognition) and it empowers students to practice English with immediate result from application. It includes 60 key expression extracted from the Korean national curriculum, consists of two parts Lesson and Takes. The students can practice orally whatever content shown to them in Lesson session. They used application for 2 weeks in and out of the classroom. They selected 302 students from five schools. It was found that 54% students agreed upon application convenience and 47% agreed upon application was interesting and learning [17].

2) *Mobile Game-Based Learning Application*: It was designed to enhance students' listening and speaking skills. It was based on PHP, MYSQL and Apache sever. Client platform was run on Android and sever run on windows 2003. Evaluation based on three week study. Four classes of 30 minutes were conducted, two groups of 20 female students were made and named as control-group and experimental group. Results showed that students who learned from mobile-system have more improvement than who have learnt through traditional method [18].

3) *Mobile Learning System*: It was designed for assisting the listening and speaking skills. This system includes six learning initiatives (Vocabulary repetition, Role-play, You speak, then I speak, Brainstorming-photo, words and photo, voice and words) which were embedded on PDA (Personal Digital Assistance). Total 33 Students of 5th grade of elementary school (10 or 11 years old) for one semester were selected. Pre-test and Post-test result showed that students had positive perception towards the system [19].

4) *English Pronunciation App*: It was based on pronunciation of words. This app was designed to improve pronunciation of English in college students of Indonesia. It was based on Android platform; it includes vowels, diaphones and consonant tests. After welcome screen the user has to choose which section he wants to learn (vowels/consonants). Words are shown with options of pronunciation, then learner has to decide which one is best pronunciation. There were 100 respondents who used this application. Majority of students became able to know correct pronunciations of mostly mispronounced words [20].

5) *Voice Recognition Technology*: It embedded on android and based on voice recognition method. Through this feature students can become capable to practice the pronunciation of English words. It was designed for professionals, based on business English. Questions were

presented to learner and learner has to answer in his/her voice then learners' voice goes to Google Voice search. After complete test learner can check his/her mistakes. Questionnaire was distributed to 35 users and they got positive result towards application [21].

6) *Video Caption Modes*: They have investigated impact of various video caption modes (Full-caption, non-caption and target-word mod) on students. It was embedded on mobile phone. Students of 5th grades were experimented for one month. Weekly test conducted to test students' English listening comprehension and vocabulary acquisition. Instructional videos related to lesson were displayed on PDA and after watching the video they immediately took a test to evaluate their listening comprehension and vocabulary proficiency [10].

7) *Grammar Clinic*: It was a web-based mobile application, which was provided as supplementary tool to students. First of all Grammar Clinic displays sentences to user then user has to identify grammatical mistakes from given sentences and make corrections. After user had corrected the sentence, it will show result and short grammar handbook according to that exercise. The types of errors included verb use, article use, noun use, word choice, adverb use, word order, conjunction use, preposition use and unambiguous expression. Intermediate level students were selected as experimental group; they were allowed to use Grammar Clinic three days a week for 16-weeks. Their score was 8.82 out of 10 for higher error correction and 7.29 out of 10 for low error correction. It had positive impact on students' grammar comprehension by identifying and then correcting the errors [22].

8) *Cellphone Game*: It was based on listening comprehension, word recognition, sentence construction, and spelling for various levels. It Included local English needs, and was tested in after school settings. Rural students were selected for study, those having difficulty to access the school. Analyzed low-gain and high-gain students, satisfactory results were found between high-gain and low-gain students. Qualifying test scores of high-gain were 49/50 and scores of low-gain were 46.5/50 [6].

9) *Mobile Learning Technology*: ESL material placed on server and was accessed by mobile device. It was based on Penguin introductory English grammar and exercise books. The course consists of 86 lessons and related exercises for teaching basic English language, ranging from is and are to verb tenses, countable nouns, and other aspects of grammar. Students completed three grammar tests during the study. Pre-test scores were 15/20, Post-test were 17.7/20, and scores in the Retention-test were 18, these scores were found after three tests [23].

Above discussed applications were developed for prosperity of English Language learning, however no existing M-Learning application for English Language have focused on elementary school students plus incorporating their own national curriculum and importantly applying listening comprehension strategy. M-Learning applications provide an

opportunity to students to learn English; however the problem with existing applications is that they are based on general or business English rule. Therefore, these apps are mostly different from their own national English curriculum.

To achieve this goal there is need to develop mobile application that is based on listening comprehension method, incorporates students own English curriculum and specially designed for elementary students, so that they can learn English after school when they do not find anyone to help them.

III. METHODOLOGY

A. Curriculum and Prototype Development

The main focus of this study was to develop an English M-learning application which embeds student's own curriculum. This helps students to learn their English curriculum on their own and increase their learning outcomes particularly in an English curriculum comprehension.

1) *Curriculum Design*: The English curriculum of Sindh Textbook board for grade 1, 2 and 3 was selected for this study. Its was selected in terms of listening comprehension, pronunciation, reading lessons and spellings. It was studied within the classroom theme, which readily related to participants. It also considered participants' competency and performance. Concretely, the curriculum includes:

a) *Common world objects*: Objects which are mostly studied in grade1 and 2.

b) *Reading Text*: Nine lessons were selected from grade 1, 2 and 3 for text reading.

c) *Spellings of words*: Spellings of words were selected alphabetically from grade 1, 2 and 3.

d) *Grammar Exercises*: Exercises, particularly parts of speech were focused from grade 1.

2) *Prototype Development*: This study developed a prototype based on above curriculum and it contained four sections i.e. Learn, Read, Spellings and Play as shown in Fig. 1.

a) *Learn Section*: This section includes different objects' (Alphabets, Numbers, Body Parts, Fruits, Colors, Shapes, Vehicles and Verbs) names with their image and pronunciation, so that students can become capable to pronounce these objects. By clicking play button and listening names of these objects their knowledge base for memory recall will increase. As a result, their cognitive ability will increase [24]. These all objects have been taken from curriculum of grade 1 and 2 of Sindh's Textbooks as shown in Fig. 2.

b) *Read Section*: This section contains lessons of grade 1, 2 and 3 of Sindh's Textbooks so that the students can become capable to remember their lesson by clicking word by word, after school or anywhere as shown in Fig. 3.

c) *Spellings Section*: This section contains spellings from A to Z alphabetical words of grade 1 so that the students can become capable to remember the spellings by clicking letter by letter as shown in Fig. 4.

d) *Play Section*: This section contains prepositions and adjectives exercises from grade 3 books. It contained fill in the blanks exercises. By playing these exercises students become capable to learn about grammar rules as shown in Fig. 5.

B. Participants

The target users were students of rural primary schools of district Sanghar, Taluka Sinjhor, Sindh, Pakistan. Several surveys were conducted to measure students learning abilities. The data was collected from total 45 students of grade 3 with their age ranging from 8-14 years, which were distributed in three groups, where one was Control Group and two were Experimental Groups. Further details of groups are given below:

a) *Group 1 (Control Group)*: This group included 15 students of Nazar Ali Khan Nizamani Boys Primary School Taluka Sinjhor District Sanghar. This school is located in village Nazar Ali Khan Nizamani, which is 26 kilometers away from Sanghar city. There were 4 boys and 11 girls (8-14 years old) and their mother tongue was Sirraiki.

b) *Group 2 (Experimental Group 1)*: This group included 15 students of Abdul Karim Sirewal Boys Primary School Taluka Sinjhor District Sanghar. This school is located in village Abdul Karim Sirewal, which is 10 kilometers away from Sanghar city. There were 3 boys and 12 girls (8-14 years old) and their mother tongue was Balochi and Sirraiki.

c) *Group 3 (Experimental Group 2)*: This group included 15 students of Darul-Elomia-Latifia Sanghar. This school is for religious studies, located in Sanghar City. There were 15 boys who were 8-14 years old. Majority of students never got chance to learn English language before, because they just had focused on Islamic studies and Holy Quran. Their aim of life was to become Hafiz-e Quran (The one who memorizes Holy Quran). Even most of them thought that Islamic studies have superiority, so they just have to concentrate on Islamic studies.

C. Teaching Method

This study took form of after school program and each session lasted for two hours in afternoon. The after-school program took place from January 2016 to March 2016 and spanned sessions on 60 days in total (30 days per each group).

- Control group was taught with two hour regular English class for one month.
- Experimental groups were taught with one hour regular English class and for one hour they were given a tablet PC loaded with developed prototype as a supplementary tool.

D. Research tools

1) *Bloom's Taxonomy*: It is a classification method to classify intellectual ability and behavior essential to learning. It was created by a psychologist Benjamin Bloom and his colleagues in 1956. It was established as a method of categorizing educational goals for student performance

assessment [25]. It consists of three main domains of learning i.e. Cognitive, Affective and Psychomotor.

a) *Cognitive Domain*: This domain covers growth of intellectual abilities and recalling information from long term memory.

b) *Affective Domain*: This domain covers development of emotions, feelings and attitudes.

c) *Psychomotor Domain*: This domain covers development of manipulative behaviour or motor skills.

From above described domains the Cognitive Domain is mostly used to assess learning behavior of students. Since this research evaluates the learning outcome of elementary school students, the Cognitive Domain has been used in this research because this domain focuses on recognition, recall, memorizing and visualizing of learning/educational material. It analysis students initial learning or primary learning behavior.

2) *Cognitive domain*: The Cognitive domain includes knowledge and the growth of intellectual skills. This includes the recall/recognition of particular information, procedural patterns, and ideas that serve in the growth of intellectual skills and abilities. There are six major types of cognitive domain as shown in Fig. 6 and Table. 1.

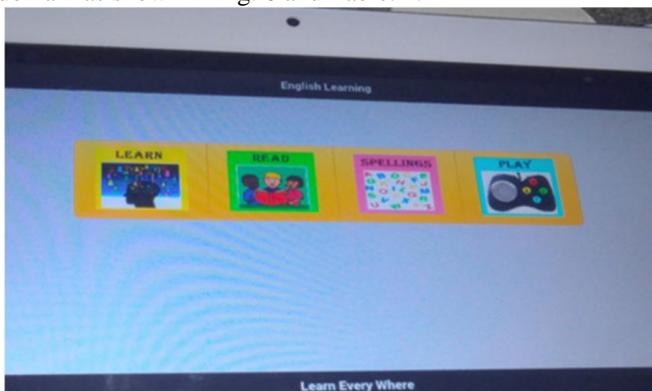
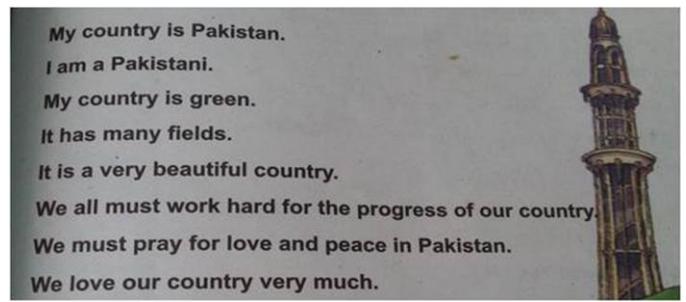


Fig. 1. User interface of prototype, it contained four sections learn, read, spellings and play.



Fig. 2. Learning Section of prototype which contained above objects and clicking play button, students will become able to learn about pronunciation objects names.



(a)



(b)

Fig. 3. Lesson of Sindh's Textbook embedded in reading section.

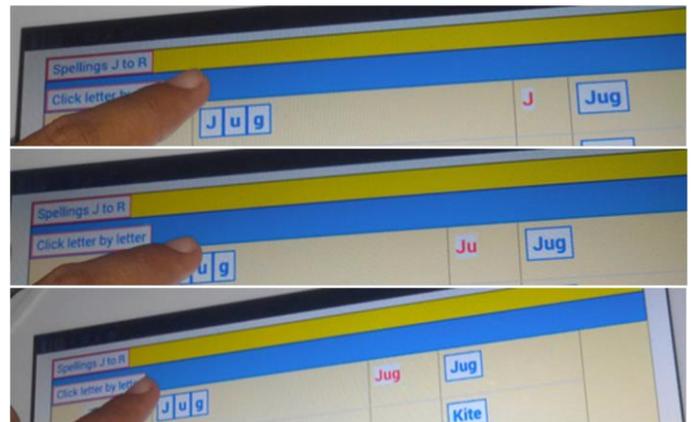


Fig. 4. Spellings section of prototype and its click event.

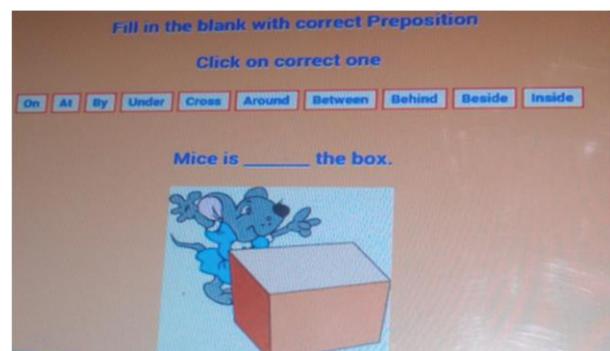


Fig. 5. Play section contains exercise of grade 3, by choosing right option student can fill the preposition into blank.

In this study knowledge and comprehension layers of Cognitive domains were selected to measure learning outcomes of elementary students of grade 1, 2 and 3. On the basis of these layers question types of the EGRA test for their assessment are selected.

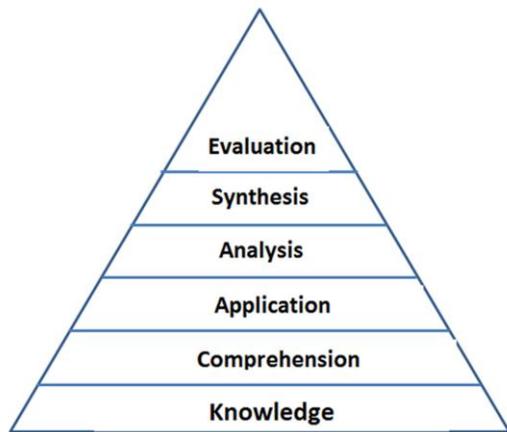


Fig. 6. Layers of cognitive domain of Bloom's Taxonomy.

TABLE I. COGNITIVE DOMAIN LAYERS

Cognitive Domain Layers	Explanation	
	Description	Question Types
Knowledge	Recalling relevant knowledge from long term memory	recall, sort, recite, select, state, tell, find, pick, group, identify, indicate match, name
Comprehension	Making Sense of what you have hear, read and visualized	comprehend, conclude, contrast, illustrate, outline, demonstrate
Application	Use the gained knowledge in new ways	Apply, construct, classify, develop, organize, solve, test, use, utilize
Analysis	Breaking knowledge into parts getting meaning from those parts	analyze, assume, breakdown, divide, deduce, diagram, infer
Synthesis	Making judgment on the basis of guideline	rearrange, write, reconstruct, revise, suggest, what, synthesize
Evaluation	Putting information together into innovative ways	reject/accept, referee, select, settle, support, umpire, weigh, which

3) *EGRA*: It is an individually managed oral assessment of foundation literacy abilities requiring about 15 minutes per child. It was developed by World Bank and USAID (United State Agency for International Development) to develop an instrument for assessing early grade reading. It is used by ministry to identify schools with particular needs and develop instructional approaches for improving foundation abilities. It has a group of subtasks; each with a specific purpose. The test modules are based on references made by an international board of testing experts and reading also include timed, 1-minute assessments of different tasks [26]-[28].

Pre and post-test are designed on the basis of *EGRA*'s seven tasks. These tasks are given below:

a) *Task 1 Phonemic Awareness*: In this Task students have to identify the word that starts with different sound in given set of words.

b) *Task 2 Letter Name Identification*: In this task student has to call the names of set of 100 Upper and Lower case letters.

c) *Task 3 Object Name Translation*: In this task students have to translate given object names into English.

d) *Task 4 Letter to Words Conversion*: In this task students have to convert given letters into word.

e) *Task 5 Word to spellings*: In this task students have to tell spellings of given words.

f) *Task 6 Reading Comprehension*: In this task students have to read the given passage of text and after completion he/she has to answer the given question.

g) *Task 7 Identification of Object names*: In this task student has to first identify the objects then he/she has to tell the name of object in English.

E. Experimental Procedure

At the beginning of the learning activities, the students took the pre-test in order to record their pre-test score before using developed prototype.

Experimental procedure was started by taking pre-test from all students (15 mints per students) from each group (control group N=15, experimental group 1 N=15 and experimental group 2 N=15 where N is the no. of participants), then treatment (developed prototype was given them as supplementary tool with traditional class of English) has been given for one month to experimental groups and control group was taught with traditional teaching method of English in class. Finally the study took post-test from students and calculated pre and post-test mean differences, as well as statistical significance ($p < 0.001$, Mann-Whitney) of results were also considered. Fig. 7 presents a flow chart of the experimental procedure.

F. Results

The tests were made on the basis of *EGRA*'s seven tasks and those tests were taken from all groups i-e one control group and two experimental groups. From results it was found that both experimental groups have higher mean difference values as compared to control group.

1) Comparison of Pre and Post-Test Mean Differences:

First of all pre-test mean of each task in all groups were calculated and after one month post-test mean of each task in all groups was calculated. Then differences of pre and post-test mean were calculated and finally pre and post-test mean differences were compared among all groups. Group wise comparisons are given below.

a) *Pre and Post-test Mean Results of Experimental Group1*: There were 15 students selected of grade 3 and evaluated all 7 tasks of *EGRA* among students and the pre and post evaluation means are shown in Table 2 and Fig. 8.

b) *Pre and Post-test Mean Results of Experimental Group2*: There were 15 selecteds selected of grade 3 and evaluated all 7 tasks of *EGRA* among students and the pre and post evaluation means are shown in Table 3 and Fig. 9.

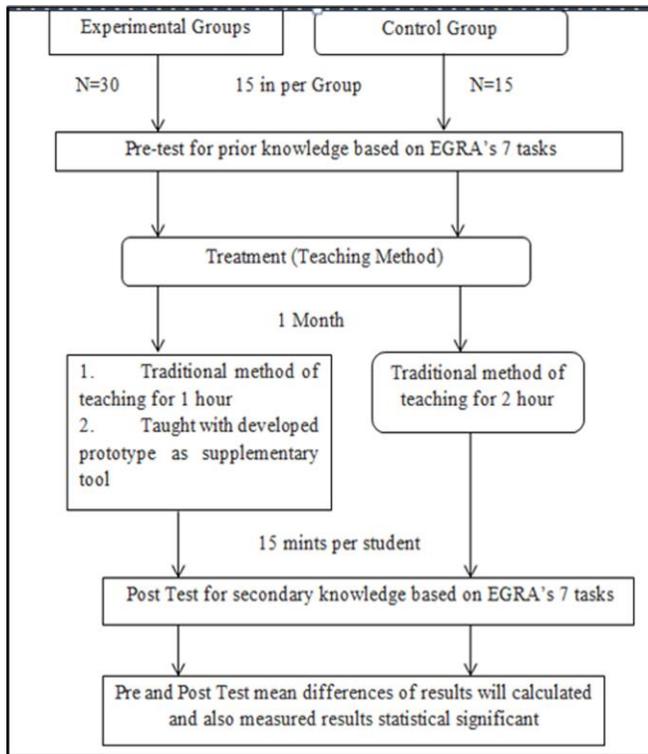


Fig. 7. Flow chart of the experimental procedure.

TABLE II. PRE AND POST EVALUATION MEAN OF EXPERIMENTAL GROUP 1 IN ALL 7 TASKS

Mean of Experimental Group 1	Tasks						
	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7
Pre Evaluation Mean	6.6	37.1	3.7	3.1	0	2.5	14.4
Post Evaluation Mean	9.9	65.9	8.1	8.2	7.4	28.6	20.7

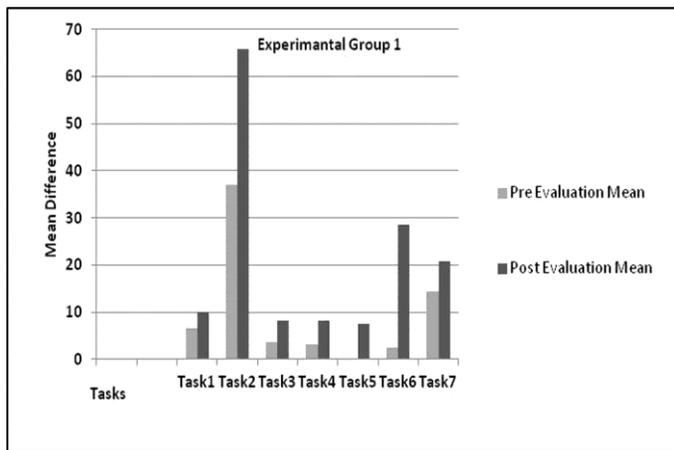


Fig. 8. Pre and post evaluation mean of experimental group 1 in all 7 tasks.

c) *Pre and Post-test Mean Results of Control Group:* There were selected 15 students of grade 3 and evaluating all 7 tasks of EGRA among students and the pre and post evaluation mean are shown in Table 4 and Fig. 10.

d) *Comparison of Mean Differences of Pre and Post-Test Evaluation of All Three Groups:* The results of Mean difference of pre and post-evaluation of all 7 tasks among all three groups, showed that both experimental groups have higher mean dieefrences as compared to control group as shown in Table 5 and Fig. 11.

2) *Measuring Sataistical Significance of Results:* Results showed that both experimental groups had higher post-test means as compared to control group. Then differences of pre and post-test mean of all 7 tasks in all groups and compared among the groups were measured. Finally statistical significance ($p < 0.001$, Mann-Whitney) of pre and post-test mean differences was measured. The term “YES” is used when results were significant ($p < 0.001$) and “NO” was used when results were not significant ($p < 0.001$). Mean differences and their statistical significance are given below.

a) *Comparesion of Mean Diffrence and Statistical Significance of Experimental Group 1 and Control Group:* The result of Mean difference of Pre and Post-Evaluation of all 7 tasks and their Statistically Significant values are shown in Table 6.

b) *Comapresion of Mean Diffrence and Statistical Significance of Experimental Group 2 and Control Group:* The result of Mean difference of Pre and Post-Evaluation of all 7 tasks and their Statistical Significant values are shown in Table 7.

c) *Comapresion of Mean Diffrence and Statistical Significance of Experimental Group 1 and Experimental Group 2:* The result of mean difference of Pre and Post-Evaluation of all 7 tasks and their Statistical Significant values are shown in Table 8.

TABLE III. PRE AND POST EVALUATION MEAN OF EXPERIMENTAL GROUP 2 IN ALL 7 TASKS

Mean of Experimental Group 2	Tasks						
	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7
Pre Evaluation Mean	7.4	21.4	0.7	1.3	0	1.6	3.8
Post Evaluation Mean	10	68.2	9.1	8.8	7.9	31	22.4

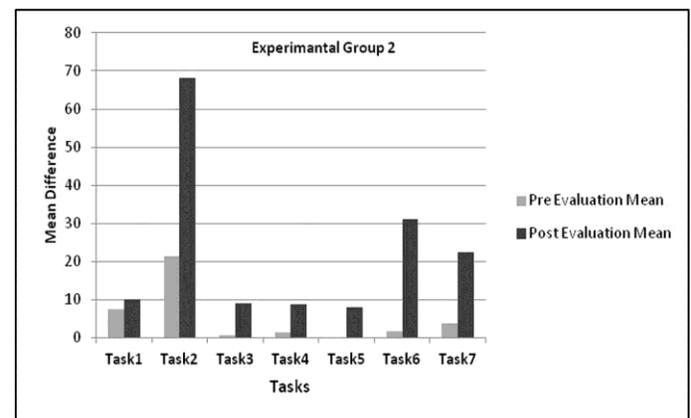


Fig. 9. Pre and post evaluation mean of experimental group 2 in all 7 task.s.

TABLE IV. PRE AND POST EVALUATION MEAN OF CONTROL GROUP IN ALL 7 TASKS

Mean of Control Group	Tasks						
	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7
Pre-evaluation Mean	6.5	34.4	3.7	3.6	0	2.4	11.8
Post-valuation Mean	9.2	58.3	5.1	5.5	4.3	6.1	15.6

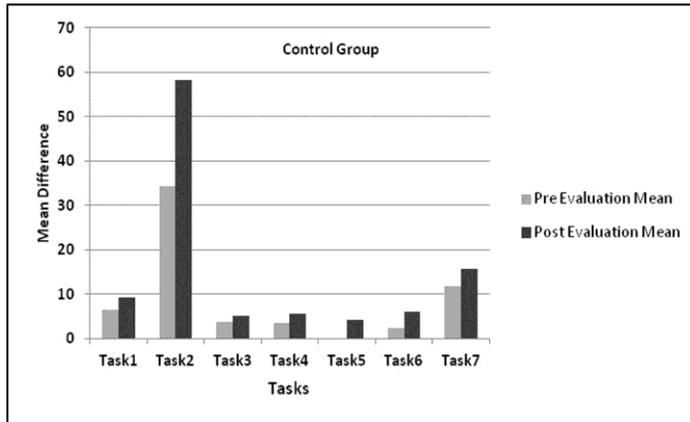


Fig. 10. Pre and post evaluation mean of control group in all 7 tasks.

TABLE V. MEAN DIFFERENCE OF EXPERIMENTAL GROUPS AND CONTROL GROUP IN ALL 7 TASKS

Groups	Tasks						
	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7
Experimental Group 1 Mean Differences	3.3	28.8	4.4	5.1	7.4	26.1	6.3
Experimental Group 2 Mean Differences	2.6	46.8	8.4	7.5	7.9	29.4	18.6
Control Group Mean Differences	2.7	23.9	1.4	1.9	4.3	3.7	3.8

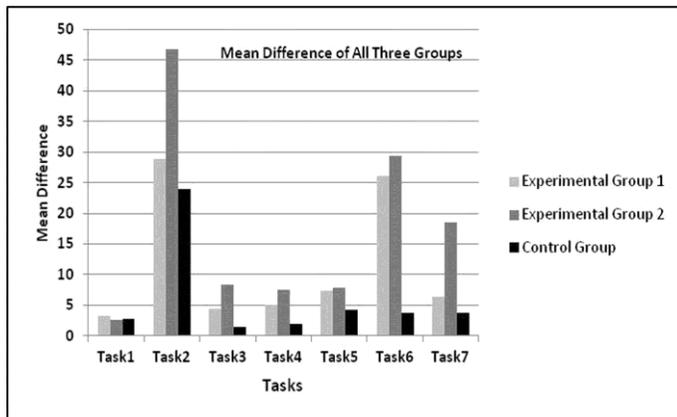


Fig. 11. Pre and post evaluation mean of control and experimental groups in all 7 tasks.

TABLE VI. MEAN DIFFERENCE OF EXPERIMENTAL GROUP 1 AND CONTROL GROUP

Tasks	Groups		
	Experimental: Group 1 Mean Difference	Control Group Mean Difference	Statistical Significant
Task 1	3.3	2.7	No
Task 2	28.8	23.9	No
Task 3	4.4	1.4	Yes
Task 4	5.1	1.9	Yes
Task 5	7.4	4.3	Yes
Task 6	26.1	3.7	Yes
Task 7	6.3	3.8	No
Combined Result of All Tasks	81.4	41.7	Yes

TABLE VII. MEAN DIFFERENCE OF EXPERIMENTAL GROUP 2 AND CONTROL GROUP

Tasks	Groups		
	Experimental: Group 2 Mean Difference	Control Group Mean Difference	Statistical Significant
Task 1	2.6	2.7	No
Task 2	46.8	23.9	Yes
Task 3	8.4	1.4	Yes
Task 4	7.5	1.9	Yes
Task 5	7.9	4.3	Yes
Task 6	29.4	3.7	Yes
Task 7	18.6	3.8	Yes
Combined Result of All Tasks	121.3	41.7	Yes

TABLE VIII. MEAN DIFFERENCE OF EXPERIMENTAL GROUP 1 AND EXPERIMENTAL GROUP 2

Tasks	Groups		
	Experimental: Group 1 Mean Difference	Experimental: Group 2 Mean Difference	Statistical Significant
Task 1	3.3	2.6	No
Task 2	28.8	46.8	Yes
Task 3	4.4	8.4	Yes
Task 4	5.1	7.5	Yes
Task 5	7.4	7.9	No
Task 6	26.1	29.4	No
Task 7	6.3	18.6	Yes
Combined Result of All Tasks	81.4	121.3	Yes

IV. CONCLUSION

The aim of this research was to measure the students learning outcomes through English Learning Mobile Application by improving English Listening Comprehension. For measuring the improvements of English learning

outcomes in students through English learning mobile application, a prototype was developed which was assessed by the students.

From the literature review it was found that English learning mobile application is helpful tool as it not only improve English learning but also measure the learning of the students. The developed prototype was evaluated on the basis of EGRA test which is based on knowledge and comprehension layers of Cognitive domain of Bloom's taxonomy. From the overall evaluations it is concluded that students learning outcomes in terms of remembering object names, reading text/lessons, spellings and knowledge of grammar improved by using the English learning mobile application. The results were compared among three groups which showed that designed prototype has statistically significant ($p < 0.001$, Mann-Whitney) impact on both experimental groups as compared to control group. Further it is concluded that students were motivated, happy and excited to be having learning through M-learning application, so that they can learn anywhere where they want.

The evaluation results showed that developed application was helpful for students who live in rural areas where they face problems doing their homework, since their relatives are not capable to teach them as accordingly.

V. RESEARCH LIMITATIONS

This study is limited to the identification of the problem scenario of English curriculum comprehension in rural students which becomes an obstacle in their learning. This dissertation is confined to English curriculum learning by improving Listening comprehension. The M-Learning application developed in this research is auditory and limited to implementation of English curriculum of grades 1, 2 and 3 of Sindh Textbook Board, which will enable students to hear and comprehend well. For assessment and evaluation, students of grade 3 were selected with age from 10 to 14 years. The EGRA test based on knowledge and comprehension layers of Cognitive domain of Bloom's Taxonomy question types selected for assessment and evaluation.

VI. FUTURE RECOMMENDATION

In this study various possibilities for future work are recommended.

1) *The Prototype can be Enhanced for Secondary Grades:* The prototype has been developed for elementary level students of grade 1, 2 and 3 and it especially embed their own English curriculum. Results of evaluation showed higher mean difference values as well as statically significant for both experimental groups as compared to control group, therefore it can also be extended for secondary grade students which embeds their own curriculum of college.

2) *The Prototype can be Extended for Other Courses:* The prototype has embodied English curriculum of Sindh Textbook Board and results of evaluation showed improvement in English curriculum learning. So it can also be extended for other curricular subjects for improvement of students learning outcomes.

3) *The Prototype can be Evaluated on Other Layers of Cognitive Domain:* The prototype has been developed and evaluated on the basis of Knowledge and Comprehension layers, therefore it can be evaluated on other layers of cognitive domain to identify that how students will response to question types of other layers.

ACKNOWLEDGEMENT

Utmost thanks to the headmasters Mr. Nazar Ali Nizamani and Mr. Abdul Karim Sirewal, and the students of Govt. Boys Primary School, and also to the head and students of Darul-Elomia-Latifia, Dist. Sanghar, Taluka Sinjhor, Sindh, Pakistan for their participation.

REFERENCES

- [1] Aziz. A.A, Umar. M, Dilshad. F, Mustafa. M, "Learning difficulties and strategies of students at higher secondary schools in punjab," Journal of Policy Research, vol. 1, pp. 55-61, 2015.
- [2] Tariq. A. R, Bilal. H. A, Sandhu. M. A, Iqbal. A. & Hayat. U, "Difficulties in learning english as second language in rural areas of pakistan", Academic Research Internationals, vol. 4, pp. 24-34, 2013.
- [3] Salahuddin. A. N. M, Khan. M. M. R & Rahman. M. A, "Challenges of implementing english curriculum at rural primary schools of Bangladesh", The International Journal of Social Sciences, vol. 7, pp. 34-51, 2013.
- [4] Kazmi. S. L, "A study of social factors affecting english learning at elementary level in rural areas of mianwali", M.A Thesis, Allama Iqbal open University.
- [5] Neha D, "An analysis of factors affecting teaching and learning of english language in rural and semi-urban colleges of India", Global Journal For Research Analysis, vol. 4, pp 2277-8160, 2015.
- [6] Kam, M., Kumar, A., Jain, S., Mathur, A. & Canny, J, "Improving literacy in rural india: cellphone games in an after-school program" IEEE, International Conference on Information and Communication Technologies and Development, pp. 139-149, April, 2009.
- [7] Weng. H.T & Chen. J.Y, "Students' perceptions towards the use of smart phone applications for english learning", International Journal of Educational, vol. 5, pp. 1-10, 2015.
- [8] Goundar.S, "What is the potential impact of using mobile devices in education", In Proceedings of SIG GlobDev Fourth Annual Workshop.
- [9] Valk. J. H, Rashid. A. T, & Elder. L. "Using mobile phones to improve educational outcomes: An analysis of evidence from Asia", The International Review of Research in Open and Distributed Learning, vol. 1, pp 117-140, 2010.
- [10] Hsu. C. K., Hwang. G. J, Chang Y. T & Chang. C. K, " Effects of video caption modes on english listening comprehension and vocabulary acquisition using handheld devices", Educational Technology & Society, vol. 16, pp. 403-414, 2013.
- [11] I Andersen "A review of 7 complete english course apps", B.A Thesis in Spanish language Teaching , 2013.
- [12] ANDERGRIFT. L & BAKER. S, " Learner variables in second language listening comprehension: an exploratory path analysis", Language Learning a Journal of Research and Language Learning, vol. 65, pp. 390-416, 2015.
- [13] Gilakjani, A. P., & Ahmadi, M. R. "A study of factors affecting efl learners' english listening comprehension and the strategies for improvement", Journal of Language Teaching and Research, vol. 2, pp. 977-988, 2011.
- [14] Luo. C, "An action research plan for developing and implementing the students' listening comprehension skills", English Language Teaching, vol. 1, pp.25, 2008.
- [15] Vandergrift. L, "Facilitating second language listening comprehension: acquiring successful strategies" ELT journal, vol.53, pp.168-176, 1999.
- [16] Pearson. D. P & Fielding. L, "Instructional implications of listening comprehension research", Reading Education Report No. 39, University of Illinois at Urbana-Champaign, 1983.

- [17] Lee. S. M, "User experience of a mobile speaking application with automatic speech recognition for EFL learning", *British Journal of Educational Technology*, 2015.
- [18] Hwang. W. Y, Shih. T. K., Ma. Z. H, Shadiev. R & Chen. S. Y," Computer assisted language learning, evaluating listening and speaking skills in a mobile game-based learning environment with situational contexts", [ahead-of-print], 1-19, 2015.
- [19] Hwang. W. Y, Shih. T. K., Ma. Z. H. Shadiev. R, & Chen. S. Y, "Effects of using mobile devices on english listening diversity and speaking for efl elementary students", *Australasian Journal of Educational Technology*, vol. 30 ,2014.
- [20] Agusalim. I. D, Assidiqi. M. H, Kom. S & Muhammad. A. F, "Developing mobile application of interactive english pronunciation training to improve efl students' pronunciation skill", *Journal of Education and Practice*, vol. 5, pp.135-139, 2014.
- [21] Suhartono. D, Calvin, Yustina. M, Kurniawati. S, Soeparno. H, Purnomo. F, "Implementation of voice recognition technology on English learning application by self-learning based on android device", 2013.
- [22] Li. Z, & Hegelheimer. V, "Mobile-assisted grammar exercises: effects on self-editing in l2 writing", *Language Learning & Technology*, vol. 17, pp.135-156, 2013.
- [23] Ally. M, Mcgreal. R., Schafer. S, Tin. T., & Cheung. B, " Use of mobile learning technology to train ESL adults", [Digests 6th International Conference on Mobile Learning, Melbourne, 2007].
- [24] Hsu, C. K., Learning motivation and adaptive video caption filtering for EFL learners using handheld devices, vol.27, pp. 84-103, 2015.
- [25] Bloom.B.S, Engelhart. M.D., Furst. E.J, Hill. W.H., Krathwohl. D.R, "Taxonomy of educational objectives, handbook i", *The Cognitive Domain*. New York: David McKay Co Inc, 1956.
- [26] Davidson. M and Hobbs. Jenny, "Delivering reading intervention to the poorest children, The case of Liberia and EGRA-Plus, a primary grade reading assessment and intervision", *International Journals of Educational Development*, vol.33, pp.283-293, 2013.
- [27] Piper. B and korda.M," EGRA Plus: Liberia Program evaluation report", RTI International, 2011.
- [28] Toolkit, *Early Grade Reading Assessment*," 2009.

Toward a New Massively Distributed Virtual Machine based Cloud Micro-Services Team Model for HPC: SPMD Applications

Fatéma Zahra Benchara

Department of Computer Science
Laboratory SSDIA, ENSET Mohammedia, Hassan II
University of Casablanca
Mohammedia, Morocco

Omar Bouattane

Department of Computer Science
Laboratory SSDIA, ENSET Mohammedia, Hassan II
University of Casablanca
Mohammedia, Morocco

Mohamed Youssfi

Department of Computer Science
Laboratory SSDIA, ENSET Mohammedia, Hassan II
University of Casablanca
Mohammedia, Morocco

Ouafae Serrar

Department of Computer Science
CRMEF MARRAKECH-SAFI,
MARRAKECH, Morocco

Hassan Ouajji

Laboratory SSDIA, ENSET Mohammedia, Hassan II
University of Casablanca
Mohammedia, Morocco

Abstract—This paper aims to propose a new massively distributed virtual machine with scalable and efficient parallel computing models for High Performance Computing (HPC). The message passing paradigm of the Processing Units has a significant impact on HPC with high communication cost that penalizes the performance of these models. Accordingly, the proposed micro-services model allows the HPC applications to enhance the processing power with low communication cost. Thus, the model based Micro-services Virtual Processing Units (MsVPUs) cooperate using asynchronous communication mechanism through the Advanced Message Queuing Protocol (AMQP) protocol in order to maintain the scalability of the Single Program Multiple Data (SPMD) applications. Additionally, this mechanism enhances also the efficiency of the model based load balancing service with time optimized load balancing strategy. The proposed virtual machine is tested and validated through an application of fine grained parallel programs for big data classification. Experimental results present reduced execution time compared to the virtual machine based mobile agent's model.

Keywords—Parallel and distributed computing; micro-services; cloud computing; distributed virtual machine; high performance computing

I. INTRODUCTION

Recently, computer science application converges to HPC one. This is due to the new application expectations for Big data analysis [1], and real time information accessibility on multiple devices (Smartphones, Laptops, Tablets...). Thus, the data to be processed and the related complex computations

oriented these applications to new HPC processing environments (clusters, grids and clouds [2], [3]) which provide the required processing power. The HPC systems based cloud computing are constituted by a set of distributed heterogeneous machines connected through an interconnection network and collaborate by their own resources in order to provide the processing power with an optimized computation time; such as in Amazon Elastic Compute Cloud (EC2) [4] that aims to enhance the execution of HPC applications in cloud. The collaboration between the distributed processing units is based on the HPC environment middleware which orchestrates the computation and manages the distribution of data and tasks between them. However, the performance of these environments is related to the one of their based middleware [5]. Normally, this middleware has to manage these two following major HPC challenges: 1) Message passing challenge the intensive communication between the computing units, has a great impact on the global computation time and the scalability of these applications, with the corresponding high communication cost. 2) Heterogeneity of computing nodes challenge the difference of nodes performance influence also the global computation time with an unbalanced computing environment caused by the overloaded workload of the slowest node. Indeed, the middleware based massively distributed computing environment has to deal with the above challenges in order to provide a scalable and efficient massively distributed computing environment. Thus, what are the promising paradigms for managing these challenges? This paper presents a new massively distributed virtual machine model based on cloud micro-services which aims to implement

an asynchronous communication mechanism for computing units message passing. The main contributions of this paper are that 1) the proposed virtual machine considers providing the processing power needed for HPC applications by its integrated micro-services team model which is constituted by virtual computing units, and 2) considers the communication challenge using a lightweight communication mechanism, and also 3) considers the heterogeneity challenge by implementing a load balancing strategies. The paper is organized as follows:

- We present the virtual machine based cloud distributed computing model and its innovative components (Section 3) which are the micro-services.
- We demonstrate that the model based middleware is promising (Section 4); and that by implementing some SPMD applications (Section 5) we ensure a scalable and efficient cooperative parallel and distributed computing environment.

II. BACKGROUND

To highlight the aim of this paper, we present the parallel and distributed computing [6] field and its key techniques for performing intensive computation in a few time. For example, in order to perform a password encryption program on 1000 passwords (Fig. 1) there are two main case study: 1) Sequential case where the program is performed on a single machine with an execution time TE per password, and the global computation time $Tt_{seq} = \sum TE$. Despite, in 2) Parallel and Distributed computing case, the program is encapsulated on 10 machines which cooperate and distribute the data between them and work in parallel so that the global computation time will be reduced significantly with $Tt(p\&d) \ll Tt_{Seq}$. The last case will perform a high performance computing if the computing model integrates some mechanisms for parallel and distributed computing challenges; the communication and the load balancing challenges. So, the scalable computing model will be the one which can optimize significantly the global computation time. This model is implemented on parallel and distributed virtual machine that orchestrates and manages the distribution of data and tasks between the nodes.

There are several inspiring proposed parallel and distributed virtual machines [7]-[11] that used different technologies such as the MCC(Mesh Connected Computer) mesh and the FPGA (Field-Programmable Gate Array). However, the scalability and efficiency of these virtual machines depends on the ability of their corresponding middleware to handle the HPC computing challenges. The Middleware is the main components in the distributed systems that can manage a set of heterogeneous nodes. The Multi Agent System MAS is a promising technology for implementing such middleware. However, the micro-services implements the flexibility with the others technologies trends and the easy integration in cloud to improve HPC.

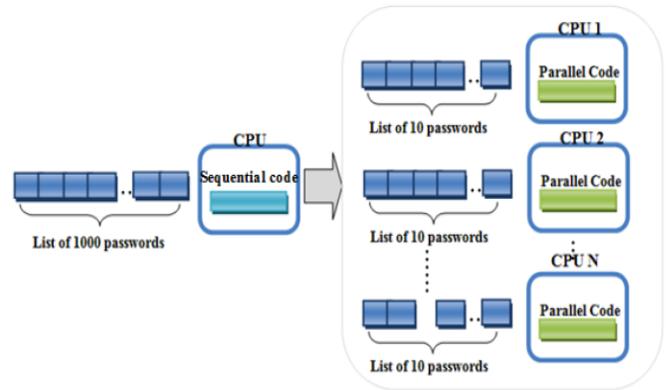


Fig. 1. Parallel and distributed computing paradigm.

III. PROPOSED MASSIVELY DISTRIBUTED VIRTUAL MACHINE

A. Massively Distributed Virtual Machine Architecture

The proposed massively distributed virtual machine is a new parallel and distributed computing environment, constituted over distributed heterogeneous nodes in distributed system. This virtual machine based micro-services model which is managed by cloud middleware, allows performing the parallel and distributed programs as services by cooperative micro-services team MsVPUs. For each deployed service, the Scientifics and researchers can take benefits of the flexibility of this virtual machine with the parallel computing models such as: SPMD, MPMD, and topologies (2D Mesh, 3D Mesh,...). Each MsVPU is an autonomous service that collaborates with the computing team using well determined communication mechanism for HPC. For example (Fig. 2), in order to perform the big data classification the well-known classification algorithms; c-means and Fuzzy c-means are implemented in this virtual machine as distributed classification service (Section 3) according to SPMD architecture. To do so, each team worker MsVPU will receive the input data from its team leader MsVPU, and perform the classification service and send the results back to its team leader in order to accomplish the execution of the application.

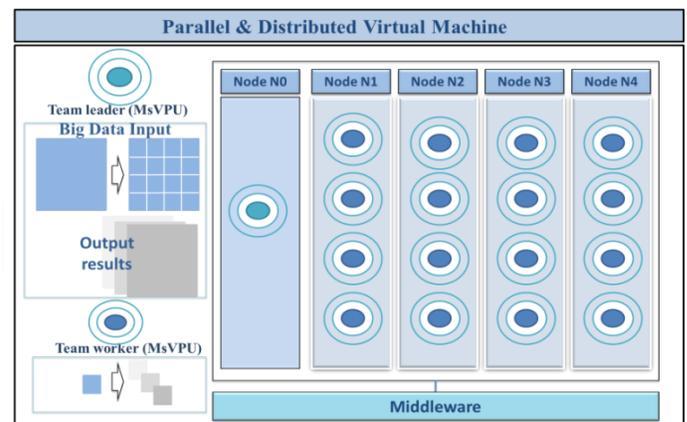


Fig. 2. SPMD distributed computing model based micro-services approach.

B. Distributed Computing Model based Main Components

In order to perform a high performance parallel and distributed computing, the proposed virtual machine model collaborate specific types of micro-services according to their tasks. When the parallel and distributed program is deployed on this virtual machine the micro-services team is created. This later is constituted by; Team leader (MsVPU) micro-service and the team workers (MsVPUs) that are distributed on each node to perform their corresponding services. Each team worker (MsVPU) encapsulates the program as service and collaborates with the other MsVPUs and provides the results to their Team leader (MsVPU) micro-service which manages and orchestrates the computing of its team while the execution of the program. This virtual machine allows deploying more than one parallel and distributed program by its integrated Proxy Ms Provider micro-service which works with the Load Balancer Ms micro-service in order to choose the appropriate team for each application request. The main principal micro-services of the model (Fig. 3) are presented as follows:

- **Proxy Ms Provider.** This micro-service is the mediator between the micro-services MsVPUs and the applications. The application requests are sent to this micro-service which communicate with the Load Balancer Ms in order to choose and send the request to the appropriate micro-service MsVPU. Then, the Proxy Ms Provider returns the results to the appropriate application.

- **Load Balancer Ms.** This micro-service is the one responsible of the management of the micro-services of the virtual machine. Each micro-service publishes its information (name, address, port, and number of CPUs) in this micro-service. So, this helps the Load Balancer Ms to get the node performance and ensure the load balancing of micro-services according to well defined load balancing strategies.
- **Team leader MsVPU.** This micro-service is the one responsible of the execution of the application requests. It cooperate with its team works (MsVPUs) in order to execute the parallel and distributed programs as services and sends the final results to the Proxy Ms Provider. This micro-service can be deployed in many distributed nodes.
- **Team worker MsVPUs.** This micro-service corresponds to a CPU. Each MsVPU receives the data from its team leader Ms and executes the service and returns the results to this later in order to compute the finale results.
- **DF Ms.** This micro-service centralizes the configuration of micro-services of the model. Each deployed micro-service will search for its configuration on this micro-service. So, the Proxy Ms Provider will easily follow the appropriate micro-services of the application request.

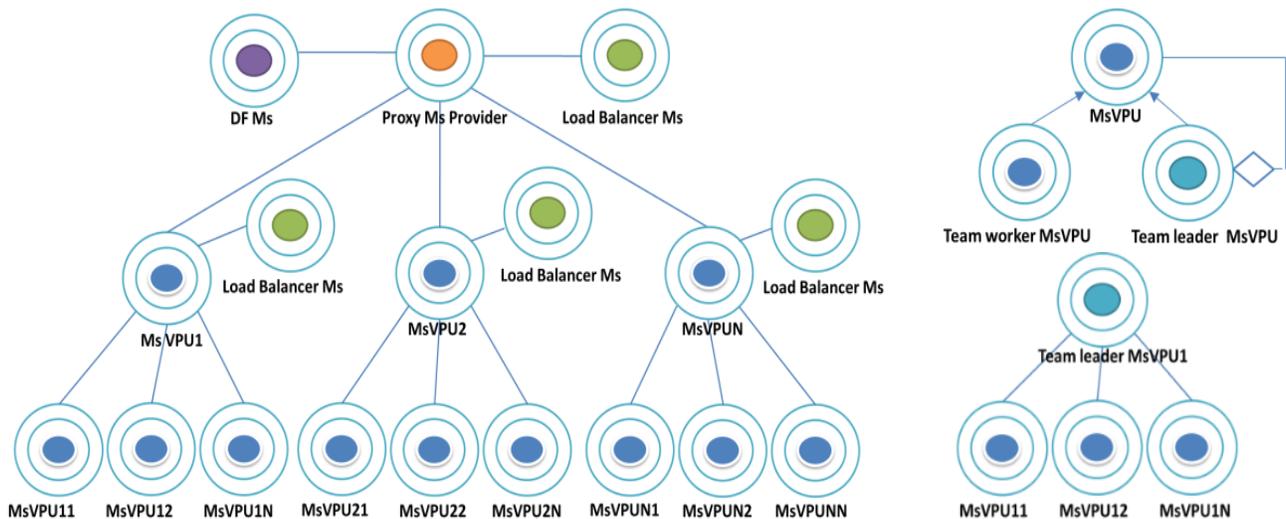


Fig. 3. Architecture of the Main components of the massively distributed virtual machine.

The UML diagram of the virtual machine model is illustrated in Fig. 4, which allows the MsVPUs micro-services to collaborate in the grid computing in order to perform the distributed services according to different programming models and parallel topologies.

The communication between the computing model main components is presented in the sequence diagram of Fig. 5. For example, in order to perform the parallel and distributed

computing service, the application sends the request with the input data to Ms Proxy Provider. This later sends this request to the Ms Load Balancer which determines the Team leader Ms that will perform this request, and sends its address to the Ms Proxy Provider which sends the input data to the right Team leader Ms in order to perform this request in collaboration with its team of MsVPUs. At the end, the final result is send back to the application by the Ms Proxy provider.

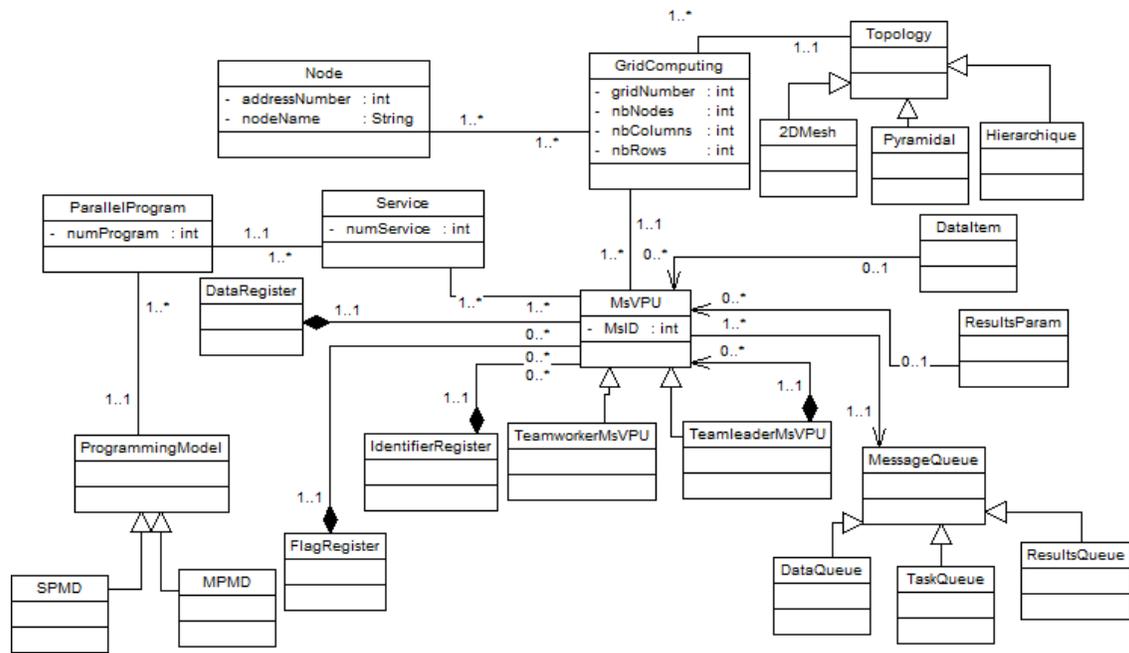


Fig. 4. UML diagram of the proposed massively distributed virtual machine model.

C. Massively Distributed Computing Middleware

The Massively Distributed computing Middleware (Fig. 6) is a new paradigm based micro-services, which allows dividing the complex tasks of the parallel programs to independent sub tasks as distributed micro-services deployed on the computing model of virtual machine. This computing model cooperates the micro-service team leader MsVPU and its micro-services

team workers MsVPU in order to perform the parallel programs on cloud computing platform. So, the scalability and efficiency of this middleware are illustrated by its two main modules; Communication Optimization Module for implementing the asynchronous communication mechanism and Load Balancing Module in order to manage the overloads between the micro-services.

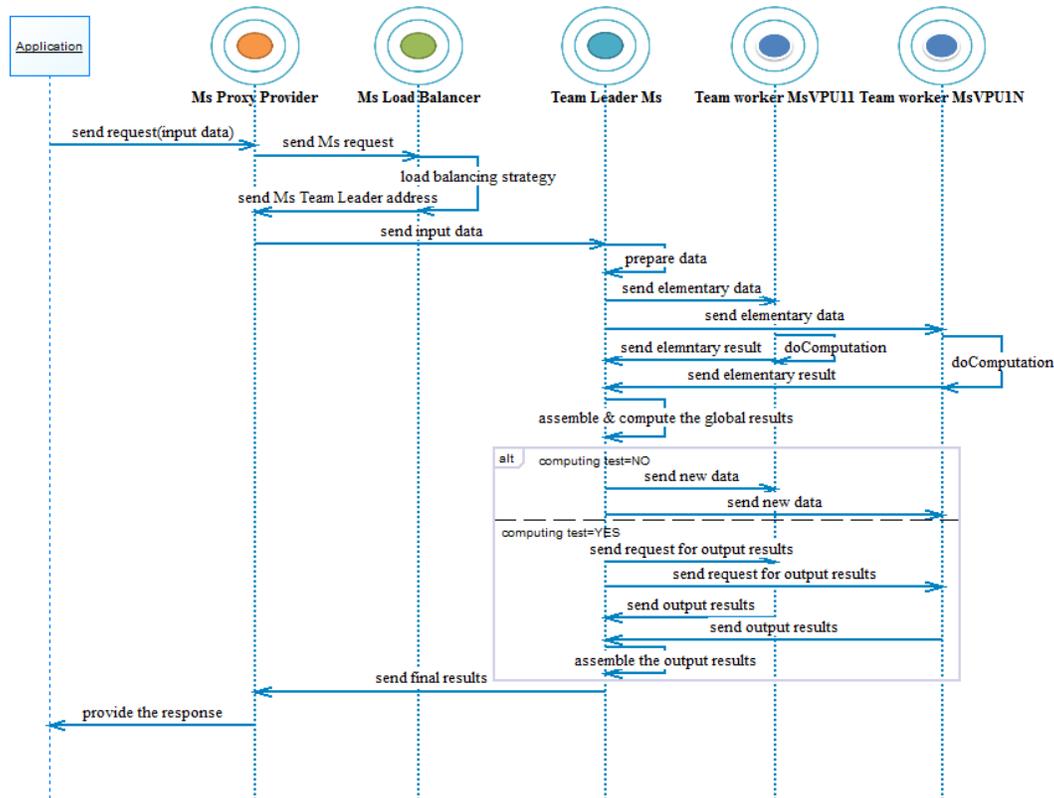


Fig. 5. Communication diagram of main components of the massively distributed virtual machine.

Communication Optimization Module This module ensures a lightweight communication mechanism between the micro-services of the computing model. This is done, by the implementation of the RabbitMQ messaging Framework in the computing model. So that the micro-services will use asynchronous communication by exchanging messages based on AMQP (Advanced Message Queuing Protocol) protocol. Furthermore, this module provides three types of message queues (data_queue, tasks_queue, and results_queue) which store and provide the exchanged messages between the micro-services. For example (Fig. 7) in order to perform an SPMD service, the Team leader Ms micro-service sends the computing data to the data_queue, and then this data is sent to the appropriate MsVPU micro-services. Each MsVPU will execute the service and send the results to the results_queue in order to be received by the Team leader Ms.

LoadBalancing Module This module provides a load balancing mechanism for the micro-services of the computing model by a specific micro-service the *Load Balancer Ms*. This later collaborates with the micro-services *TNPMs* (Team Node Performance Micro-service) which are deployed on each node in order to define the performance index of all the nodes of the distributed system, and their loads index. So, the Load Balancer Ms will get the set of TNPMs micro-services from the DF Ms micro-service, and execute the performance test in collaboration with TNPMs micro-services in order to define the required metadata for elaborating the load balancing strategy (Fig. 8) according to these three global steps:

- **Initial Performance Test of nodes** The Ms Load Balancer executes the performance test on the node N_0 , and then it sends the data D_0 to the TNPMs micro-services at t_0 . Each TNPMs micro-service performs the performance test on its data D_0 and sends the result R_i that is composed by (Computation Time T_{pi} , and the number of CPUs NC_i), to the Ms Load Balancer at $t_1(i)$.

These results will be used by the Ms Load Balancer in order to get the metadata {Execution Time TE ($TE_i=(t_1(i)-t_0)$), and Communication Latency TL ($TL_i= TE_i-T_{pi}$)} needed to define the initial performance index and the loads index respectively according to the following equations:

$$NPI_i^{Ms_0} = \frac{MIN(TE_i^{Ms_0})}{TE_i^{Ms_0}} \quad (1)$$

$$NLI_i^{Ms_0} = \text{round} \left(\frac{nbMs}{n} \times \frac{NPI_i^{Ms_0}}{NPI_1^{Ms_0}} \right) \quad (2)$$

$$\overline{NPI}_i^{Ms_0} = \frac{\sum_{i=0}^{n-1} NPI_i^{Ms_0}}{n} \quad (3)$$

- **Performance Index of the Nodes NPI** The Load Balancer Ms uses the metadata of the initial performance test {TE, TL, C_0 } and the metadata of MsVPU_s { C_k the complexity of service, and Z_k the amount of data exchanged between the node N_0 and N_i } in order to define the performance index NPI_i of each node N_i by :

$$NPI_i = \frac{MIN(TE_i^{Ms})}{TE_i^{Ms}} \quad (4)$$

Also, the execution time, and the latency of communication and the computational time can be estimated respectively by :

$$Tp_i^{Ms} = \frac{Tp_i^{Ms_0} \times C_k}{C_0} \quad (5)$$

$$TL_i^{Ms} = \frac{TL_i^{Ms_0} \times Z_k}{Z_0} \quad (6)$$

$$TE_i^{Ms} = (Tp_i^{Ms} / NC_i) + TL_i^{Ms} \quad (7)$$

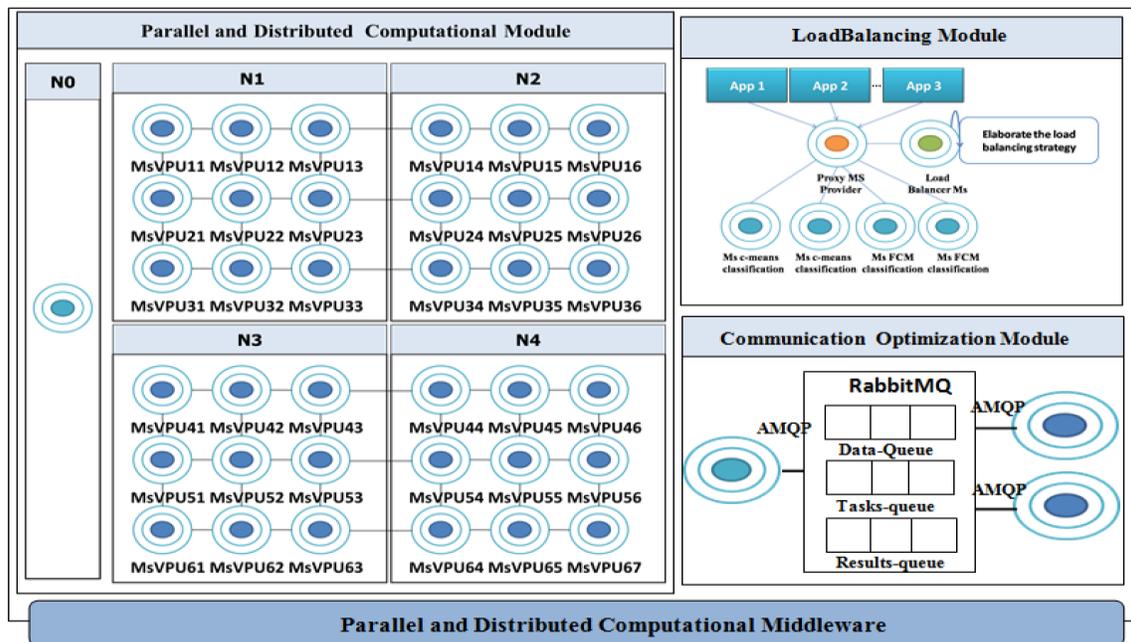


Fig. 6. Parallel and distributed computing middleware based micro-services modules.

- Load Index of the Node NLI:** The Load Balancer Ms computes the load index of each node N_i by (8) based on the total number nbMs of micro-services needed for performing the request, and the number n of nodes, and the performance index NPI. This micro-service can get the micro-services information (address of node, port number) in order to choose the appropriate micro-services on each node N_i for performing the application

request, by the way to maintain a balanced virtual machine.

$$NLI_i = \text{round} \left(\frac{nbMs}{n} \times \frac{NPI_i}{\overline{NPI}_i} \right) \quad (8)$$

where

$$\overline{NPI}_i = \frac{\sum_{i=0}^{n-1} NPI_i}{n} \quad (9)$$

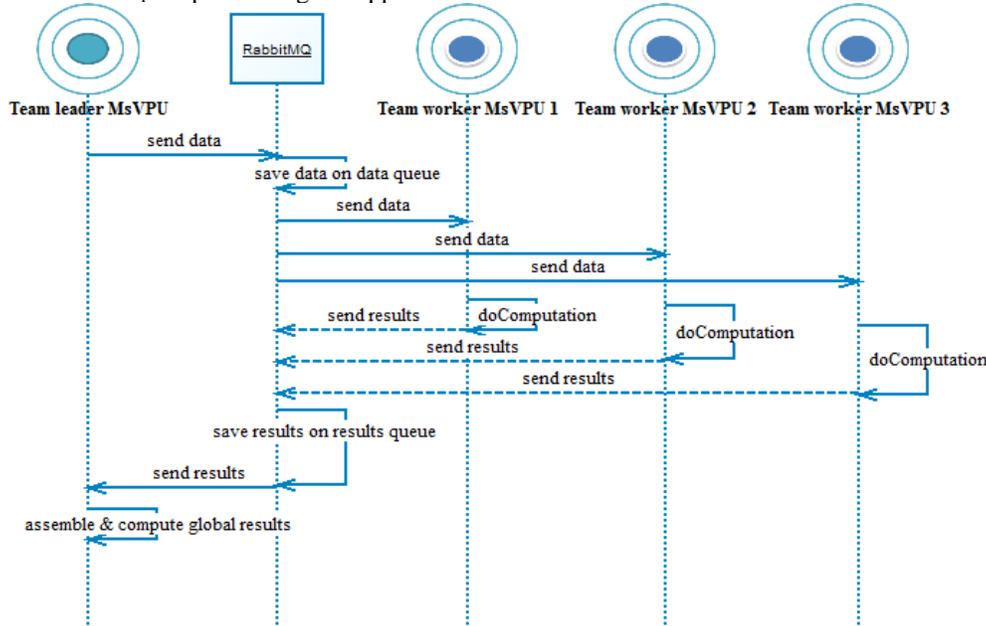


Fig. 7. Communication diagram of MsVPUs based asynchronous communication mechanism.

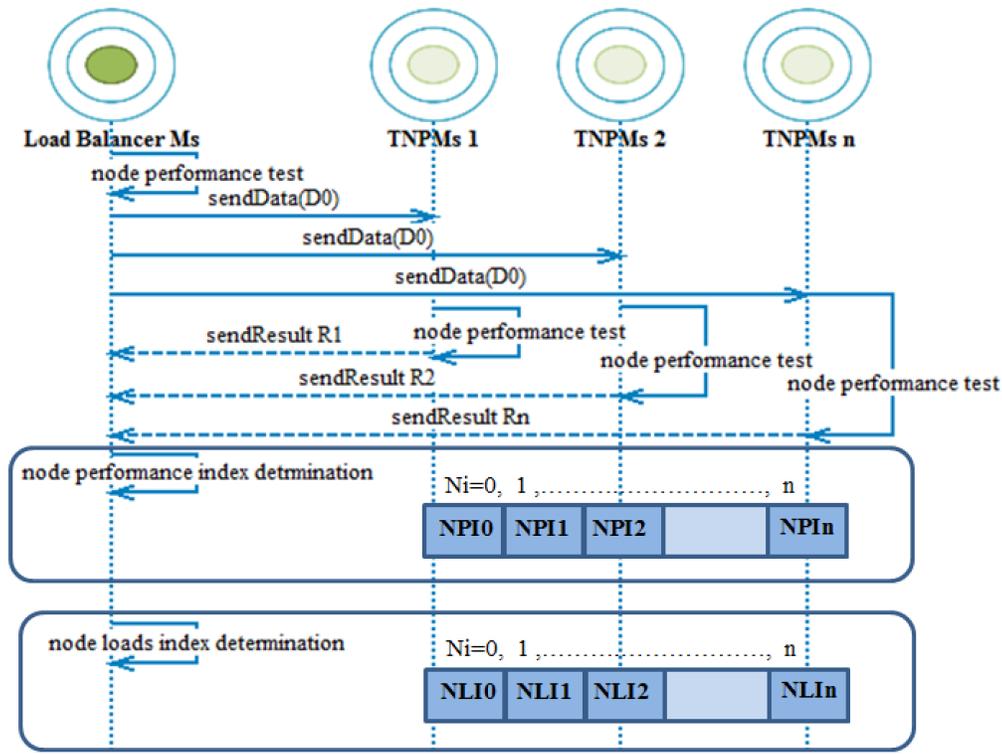


Fig. 8. Load balancing strategy based micro-services model.

IV. APPLICATION & RESULTS

For testing the scalability and efficiency of the proposed virtual machine model, the two well-known SPMD classification algorithms; c-means [12] and Fuzzy c-means [13] are implemented as distributed services using the Spring Cloud Middleware.

A. Distributed Implementation

The classification algorithms are implemented on the MsVPUs of the computing model according to the communication diagram in Fig. 9. This diagram presents the micro-services MsVPUs and their implemented services in order to perform the classification of big image. For example, in order to perform the classification of the image the c-means algorithm is implemented according to distributed implementation DSCM (Distributed Service C-means) as follows:

- The Team leader MsVPU divides the input image on $NS=me \times ne$ elementary images.
- The Team leader MsVPU sends the elementary images NS to the Team workers MsVPUs, one per team worker MsVPU(s).
- Each Team worker MsVPU(s) gets its elementary image EI, and performs its classification service.
- For each iteration t

{

1) The Team leader MsVPU sends the initial class centers to all the Team workers MsVPU(s).

2) Each Team worker MsVPU(s) gets the class centers values and performs the classification service (doClassificationService). This service allows the Team worker MsVPU(s) to perform the classification on its elementary image and computes and elementary results :

ER2(s,k) the sum of colors of each class centers c_k , which is computed by:

$$ER2(s,k) = \sum_{j, x_j \in C_k} x_j \quad (11)$$

ER3(s,k) the sum of the membership matrix of each class centers c_k , which is computed by:

$$ER3(s,k) = \sum_{j=1}^{pi} c_k \quad (12)$$

where **pi** is the number of pixels of the Team worker MsVPU elementary image EI.

ER1(s) the sum of distances of each class centers c_k , which is computed by:

$$ER1(s) = \sum_{j, x_j \in C_k} d(x_j, c_k) \quad (10)$$

At the end of the classification, each Team worker MsVPU(s) sends its elementary results ER1(s), ER2(s, k), ER3(s,k) to its Team leader MsVPU.

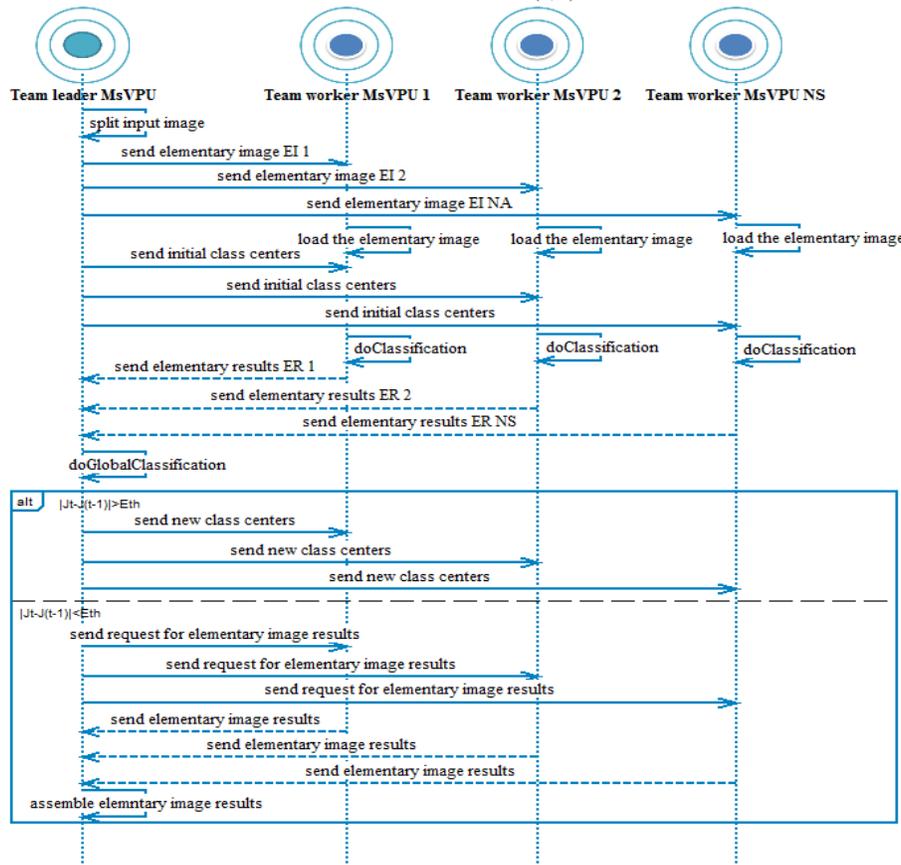


Fig. 9. Communication diagram of distributed big data classification model main components.

3) The Team leader MsVPU gets the elementary results of the Team workers MsVPUs and performs the global classification service (doGlobalClassificationService) which is based on performing the three following sub services :

Assembling the elementary results: When the Team leader MsVPU receives the elementary results (ER1(s), ER2(s,k), ER3(s,k)). This later computes the global results (GER1(k), GER2(k), GER3(k)), respectively by (13),(14),(15).

GER1(k) the global value of ER1(s) of all the Team workers MsVPUs.

$$GER1(k) = \sum_{s=1}^{NS} ER1(s) \quad (13)$$

GER2(k) the global value of ER2(s) of all the Team workers MsVPUs.

$$GER2(k) = \sum_{s=1}^{NS} ER2(s, k) \quad (14)$$

GER3(k) the global value of ER3(s) of all the Team workers MsVPUs.

$$GER3(k) = \sum_{s=1}^{NS} ER3(s, k) \quad (15)$$

Calculate the new class centers: The Team leader MsVPU computes the new value of class centers based on the value of GER2(k) and GER3(k) by (16).

$$c_k = \frac{GER2(k)}{GER3(k)} \quad (16)$$

Computes the objective function J_t : The Team leader MsVPU uses the computed value of GER1(k) to determine the objective function by (17).

$$J_t = \sum_{k=1}^c GER1(k) \quad (17)$$

4) Test of convergence of the algorithm ($|J_t - J_{(t-1)}| < E_{th}$). The Team leader MsVPU compare the difference between the obtained objective function J_t and the one obtained in the previous iteration with the error (E_{th}), if $|J_t - J_{(t-1)}| < E_{th}$ (end), else (repeat from step 1 with the new value of the class centers).

}// End of iteration t

- The Team leader MsVPU requests the segmented elementary output images from the Team worker MsVPUs in order to assemble and provide the c outputs images and the final results to the application by Proxy provider Ms.

B. Results

The scalability and the performance of the proposed model are illustrated through an SPMD application. This application has to process a satellite image of size (row, column)=(7280, 7750) pixels on three output images C1, C2, C3 as shown in Fig. 10. The two classification services; c-means and fuzzy c-means using the same initial class centers (1.2, 2.5, 3.8) are performed under this application. We conclude in Table 1 and Table 2, that the two services; DSCM and DSFCM converge dynamically to the same final class centers (4.866, 112.396, 163.370). Fig. 11 and 12 show the dynamic convergence and the error of the objective function of both services.

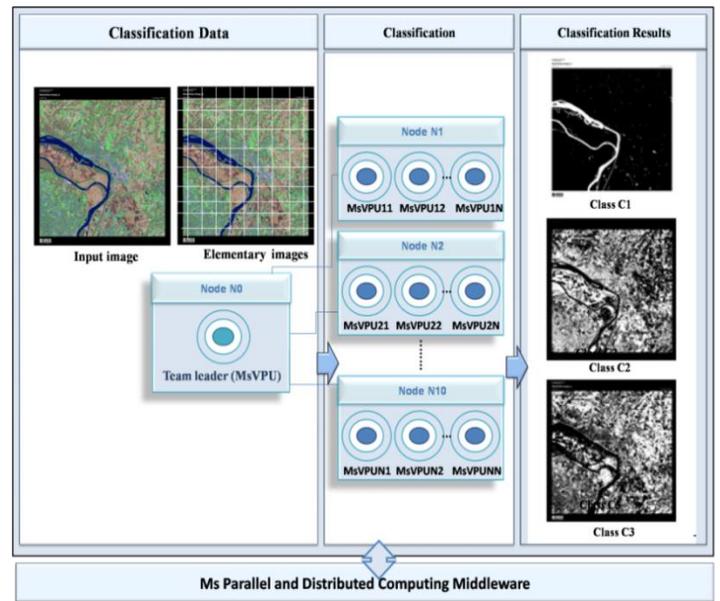


Fig. 10. Output classification image results using the proposed virtual machine based middleware.

TABLE I. DIFFERENT STATES OF THE DISTRIBUTED FUZZY C-MEANS SERVICE (DSCM) STARTING FROM THE CLASS CENTERS (C1, C2, C3) = (1.2,2.5,3.8).

Iteration	Value of each class center			Absolute value of threshold $ J_t - J_{t-1} $
	C ₁	C ₂	C ₃	
1	1,200	2,500	3,800	7,50E+00
2	0,001	2,253	132,119	1,27E+02
3	0,001	31,030	138,600	3,53E+01
4	0,219	47,574	140,414	1,86E+01
5	1,225	65,062	142,384	2,05E+01
6	3,481	87,135	145,767	2,77E+01
7	4,089	99,093	152,041	1,88E+01
8	4,379	105,239	156,876	1,13E+01
9	4,563	108,870	160,110	7,05E+00
10	4,706	110,646	161,733	3,54E+00
11	4,784	111,797	162,822	2,32E+00
12	4,866	112,396	163,370	1,23E+00
13	4,866	112,396	163,370	0,00E+00

For validating the performance of the proposed model, the classification time is analyzed for both services according to the involved number of MsVPUs in the classification in Fig. 13. We conclude that for both services the classification time achieves its minimum values of 26331 ms for DSCM and of 153970 ms for DSFCM using 32 MsVPUs. This number of MsVPUs is the required number for the classification of this case of image.

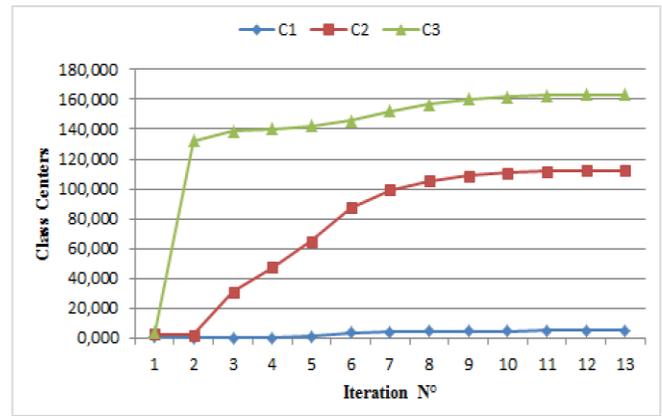
To illustrate the performance of the communication mechanism and the load balancing of this model, a compared study of the proposed model is performed with the mobile agent's model. In this study the corresponding application has to process 1000 elementary images of size (1024 × 786) pixels, by the way that 1000 MsVPU micro-services will execute the same service of complexity $C_k(x) = O(x^2)$ in parallel.

TABLE II. DIFFERENT STATES OF THE DISTRIBUTED FUZZY C-MEANS SERVICE (DSFCM) STARTING FROM THE CLASS CENTERS (C1, C2, C3) = (1,2,2.5,3.8).

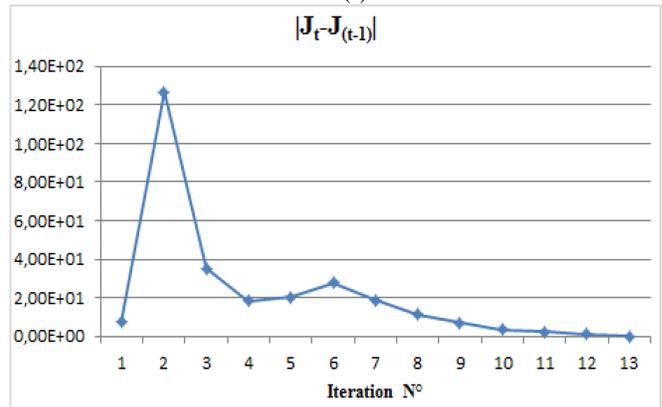
Iteration	Value of each class center			Absolute value of the error $ J_n - J_{n-1} $
	C ₁	C ₂	C ₃	
1	1,200	2,500	3,800	1,95E+09
2	56,561	123,684	128,646	1,03E+09
3	18,376	128,181	142,342	2,30E+08
4	6,348	122,692	153,589	1,37E+08
5	4,745	116,811	160,697	2,41E+07
6	4,344	114,478	163,436	1,28E+06
7	4,216	113,824	164,239	1,14E+06
8	4,180	113,666	164,460	4,42E+05
9	4,172	113,633	164,523	1,14E+05
10	4,170	113,630	164,543	3,39E+04
11	4,170	113,632	164,551	1,21E+04
12	4,170	113,634	164,555	5,26E+03
13	4,170	113,636	164,557	2,68E+03
14	4,170	113,637	164,558	1,51E+03
15	4,170	113,637	164,559	8,95E+02
16	4,170	113,638	164,559	5,42E+02
17	4,170	113,638	164,559	3,31E+02
18	4,170	113,638	164,560	2,03E+02
19	4,170	113,638	164,560	1,25E+02
20	4,170	113,638	164,560	7,69E+01
21	4,170	113,639	164,560	4,75E+01
22	4,170	113,639	164,560	2,91E+01
23	4,170	113,639	164,560	1,78E+01
24	4,170	113,639	164,560	1,11E+01
25	4,170	113,639	164,560	6,76E+00

From Fig. 14 and Table 3, it can be seen that the AVPU model achieves an acceptable load balancing with $\Omega_{MAX}^{EXP}(AVPU) - \Omega_{MIN}^{EXP}(AVPU) = 82,5517$ s with the error $\epsilon_i(AVPU) \in [0.00014, 0.03]$ and for the proposed model based MsVPU $\Omega_{MAX}^{EXP}(MsVPU) - \Omega_{MIN}^{EXP}(MsVPU) = 71,5271$ s with $\epsilon_i(MsVPU) \in [0.00014, 0.03]$ which means a gain of performance of $\varphi 1 = \frac{\Omega_{MAX}^{EXP}(AVPU) - \Omega_{MIN}^{EXP}(AVPU)}{\Omega_{MAX}^{EXP}(MsVPU) - \Omega_{MIN}^{EXP}(MsVPU)} = 1,15413$ compared to the AVPUs based model. This is due, to the lightweight communication mechanism of the micro-services compared to the mobile agents. So, the both models integrate the mechanism to ensure high performance computing.

From Table 4, it can be seen that the both models provide autonomous virtual computing units which enhance the processing power, and manage the computing challenges; heterogeneity of computing nodes and the message passing mechanism of computing units. So, the HPC applications can take advantages of these two models.

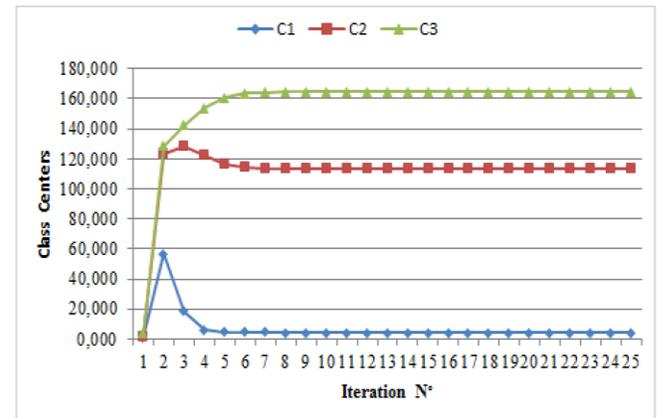


(a)



(b)

Fig. 11. Dynamic convergence of DSCM service with initial class centers (c1, c2, c3) = (1.2, 2.5, 3.8); (a) Class centers, (b) Error of the objective function.



(a)

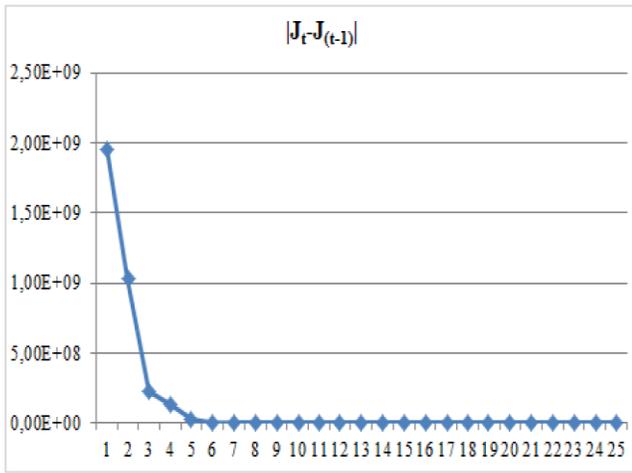


Fig. 12. Dynamic convergence of DSFCM service with initial class centers $(c_1, c_2, c_3) = (1.2, 2.5, 3.8)$; (a) Class centers, (b) Error of the objective function.

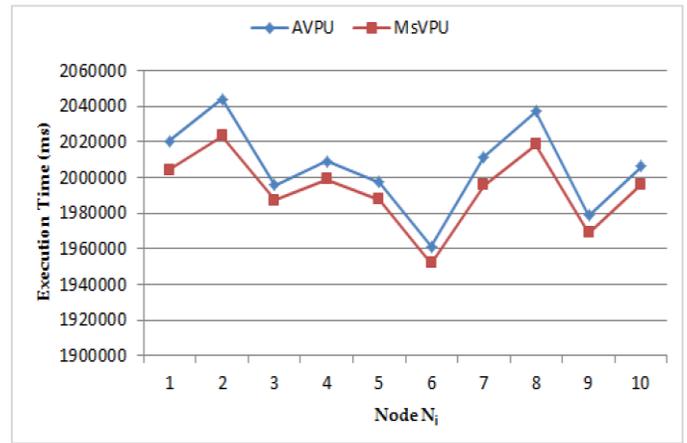


Fig. 14. Execution time for each node N_i , for the both computing models; Agent AVPU Model and Micro-service MsVPU Model.

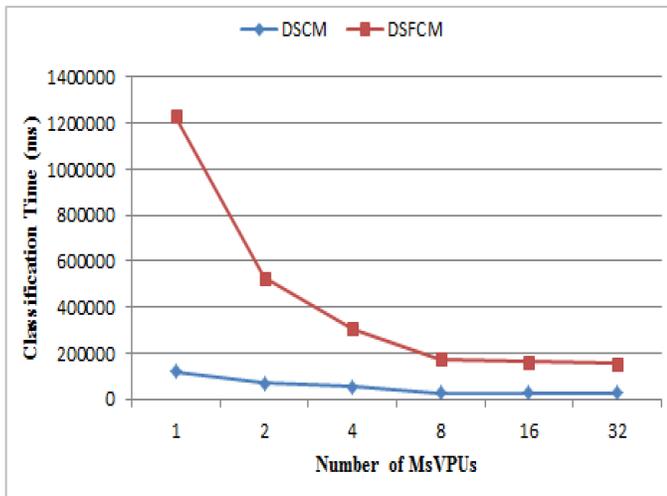


Fig. 13. Classification Time depending on the number of MsVPUs for c-means service DSCM and Fuzzy c-means service DSFCM.

TABLE III. COMPARISON OF EXECUTION TIME FOR EACH NODE N_i FOR THE BOTH COMPUTING MODELS; AGENT AVPU MODEL AND MICRO-SERVICE MSVPU MODEL

N_i	AVPU Model		MsVPU Model	
	\mathcal{L}_i^{EXP} (ms)	ε_i	\mathcal{L}_i^{EXP} (ms)	ε_i
0	2020513,677	0,007556677	1989694,791	0,007947277
1	2043923,005	0,019275406	2012147,181	0,01937828
2	1996003,738	0,004652171	1964820,136	0,004652169
3	2009285,591	0,001981942	1977301,264	0,001761339
4	1997420,298	0,003964798	1966256,613	0,003868472
5	1961371,238	0,021891009	1930789,454	0,021891011
6	2011818,492	0,003214181	1980248,984	0,003214181
7	2037281,587	0,015916015	2005008,187	0,015801354
8	1978621,624	0,013275925	1964954,997	0,004573581
9	2006293,217	0,000495296	1975019,785	0,000594413

TABLE IV. COMPARISON BETWEEN THE PARALLEL AND DISTRIBUTED COMPUTING MODELS; AGENT AVPU MODEL AND MICRO-SERVICE MSVPU MODEL.

	AVPU Model	MsVPU Model
Virtual Processing Unit	Mobile agent	Micro-service
Fault Tolerance	Agent clone mechanism	Micro-service platform libraries. Example for Spring cloud (Netflix Hystrix)
Load Balancing	Agent migration	Micro-service platform libraries. Example for Spring cloud (Eureka)
Communication	Asynchronous communication with ACL (Agent Communication Language) message.	Asynchronous communication with AMQP (Advanced Message Quering Protocol) Protocol.
Deployment	With Multi agents platform using JVM	With Micro-service containers using Cloud Computing.
Performance	High	High

V. RELATED WORKS

Several inspired researches have been proposed for scaling up the cloud computing [14]-[18]. For example in [18], the authors presented a new approach for horizontally scaling cloud resources, and in [19] for load balancing, and resources scheduling [20] in cloud. In [21] the authors presented the cloud concept and its emerged services that deal with the IoT trends, and they notice also that the applications with complex data-intensive computations are the best candidate to take advantages of cloud computing. Therefore, by applying the parallel and distributed simulation on cloud, the performance of these applications depends on the applied synchronization algorithms [22]. Our approach considers the performance of the HPC applications which are implemented on cloud architectures using micro-services, and deals with intensive computing units communication.

The cloud native applications [23], [24] with their related micro-services architectures can be promising methodologies for HPC applications in cloud computing. For example in [25] the authors presented a performance evaluation of micro-services architectures using containers: master-slave and nested-container, and in [26] they discussed the benefit of implementing micro-services architecture for emerging the telecom application. Also, the micro-services approach is implemented to digital curation infrastructure by devolving function into a set of micro-services which grants the deployment flexibility and simplify of the development and the maintenance [27]. These features of the micro-services architectures deal with the new trends of the HPC middleware [5], and ensure the scalability of the distributed applications [28].

VI. CONCLUSION

The massively distributed virtual machine model based micro-services for HPC. This model integrates a cooperative team of micro-services that are deployed as virtual computing components MsVPUs (Micro-service Virtual Processing Units) for performing the parallel programs as services according to different architectures and topologies. The MsVPUs are the virtual processors that enhance the processing power. Also, they use the asynchronous communication mechanism based AMQP protocol to optimize the communication cost of the model. This model implements a load balancing module for managing the micro-services and ensures the high performance computing. In this paper, the efficiency and the performance of this model are illustrated through an application of classification using the two services; c-means and Fuzzy c-means. In each node a specific number of MsVPUs are deployed according to the load balancing method. So, the MsVPUs cooperate in different nodes and execute the application by the way to ensure a balanced virtual machine with low communication cost. Compared to the mobile agents based model, the proposed model grants a lightweight communication mechanism which optimizes significantly the communication cost. Also, the proposed virtual machine capabilities provides the ability to extended its model to an elastic platform that will be deployed in Cloud as PaaS (Platform as a Service).

REFERENCES

- [1] M. D. Assunção, R. N. Calheiros, S. Bianchi, M. A. S. Netto, and R. Buyya, "Big Data computing and clouds: Trends and future directions," *J. Parallel Distrib. Comput.*, vol. 79, pp. 3–15, 2015.
- [2] M. Armbrust, A. Fox, R. Griffith, A. Joseph, and RH, "Above the clouds: A Berkeley view of cloud computing," Univ. California, Berkeley, Tech. Rep. UCB , pp. 07–013, 2009.
- [3] M. Zakarya and L. Gillam, "Energy efficient computing, clusters, grids and clouds: A taxonomy and survey," *Sustain. Comput. Informatics Syst.*, vol. 14, pp. 13–33, 2017.
- [4] A. Marathe, R. Harris, D. K. Lowenthal, B. R. de Supinski, B. Rountree, and M. Schulz, "Exploiting Redundancy and Application Scalability for Cost-Effective, Time-Constrained Execution of HPC Applications on Amazon EC2," *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, no. 9. pp. 2574–2588, 2016.
- [5] C. Engelmann, H. Ong, and S. L. Scott, "Middleware in Modern High Performance Computing System Architectures," in *Computational Science -- ICCS 2007: 7th International Conference, Beijing, China, May 27 - 30, 2007, Proceedings, Part II*, Y. Shi, G. D. van Albada, J. Dongarra, and P. M. A. Sloot, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 784–791.
- [6] H. El-Rewini and M. Abd-El-Barr, *Advanced Computer Architecture and Parallel Processing*. 2005.
- [7] R. Miller and Q. F. Stout, "Geometric Algorithms for Digitized Pictures on a Mesh-Connected Computer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 7, no. 2, pp. 216–228, 1985.
- [8] J. Wu, J. JaJa, and E. Balaras, "An Optimized FFT-Based Direct Poisson Solver on CUDA GPUs," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 3, pp. 550–559, 2014.
- [9] A. Rafique, G. A. Constantinides, and N. Kapre, "Communication Optimization of Iterative Sparse Matrix-Vector Multiply on GPUs and FPGAs," *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 1, pp. 24–34, Jan. 2015.
- [10] M. Youssfi, O. Bouattane, and M. O. Bensalah, "A Massively Parallel Re-Configurable Mesh Computer Emulator: Design, Modeling and Realization," *J. Softw. Eng. Appl.*, pp. 11–26, 2010.
- [11] M. Youssfi, O. Bouattane, F. Z. Benchara, and M. O. Bensalah Mohammed, "A Fast Middleware For Massively Parallel And Distributed Computing," *IJRCCCT*, vol. 3, no. 4, 2014.
- [12] O. Bouattane, B. Cherradi, M. Youssfi, and M. O. Bensalah, "Parallel c-means algorithm for image segmentation on a reconfigurable mesh computer," *Parallel Comput.*, vol. 37, no. 4, pp. 230–243, 2011.
- [13] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. Norwell, MA, USA: Kluwer Academic Publishers, 1981.
- [14] S. Niu, J. Zhai, X. Ma, X. Tang, W. Chen, and W. Zheng, "Building Semi-Elastic Virtual Clusters for Cost-Effective HPC Cloud Resource Provisioning," *IEEE Trans. Parallel Distrib. Syst.*, vol. 27, no. 7, pp. 1915–1928, 2016.
- [15] B. Mao, S. Wu, and H. Jiang, "Exploiting Workload Characteristics and Service Diversity to Improve the Availability of Cloud Storage Systems," *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, no. 7. pp. 2010–2021, 2016.
- [16] W. Xiao, W. Bao, X. Zhu, C. Wang, L. Chen, and L. T. Yang, "Dynamic Request Redirection and Resource Provisioning for Cloud-Based Video Services under Heterogeneous Environment," *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, no. 7. pp. 1954–1967, 2016.
- [17] H. Wang, Z. Kang, and L. Wang, "Performance-Aware Cloud Resource Allocation via Fitness-Enabled Auction," *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, no. 4. pp. 1160–1173, 2016.
- [18] D. Grimaldi, A. Pescapè, A. Salvi, s. santini, and V. Persico, "A Fuzzy Approach based on Heterogeneous Metrics for Scaling Out Public Clouds," *IEEE Transactions on Parallel and Distributed Systems*, vol. PP, no. 99. p. 1, 2017.
- [19] J. Zhao, K. Yang, X. Wei, Y. Ding, L. Hu, and G. Xu, "A Heuristic Clustering-Based Task Deployment Approach for Load Balancing Using Bayes Theorem in Cloud Environment," *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, no. 2. pp. 305–316, 2016.

- [20] X. Xu, L. Cao, X. Wang, X. Xu, L. Cao, and X. Wang, "Resource pre-allocation algorithms for low-energy task scheduling of cloud computing," *Journal of Systems Engineering and Electronics*, vol. 27, no. 2, pp. 457–469, 2016.
- [21] S. Sharma, V. Chang, U. S. Tim, J. Wong, and S. Gadia, "Cloud-based emerging services systems," *International Journal of Information Management*, 2016.
- [22] G. D'Angelo and M. Marzolla, "New trends in parallel and distributed simulation: From many-cores to Cloud Computing," *Simul. Model. Pract. Theory*, vol. 49, pp. 320–335, 2014.
- [23] N. Kratzke and P. C. Quint, "Understanding cloud-native applications after 10 years of cloud computing - A systematic mapping study," *J. Syst. Softw.*, vol. 126, pp. 1–16, 2017.
- [24] D. Namiot and M. Sneps-Snepe, "On Micro-services Architecture," *Int. J. Open Inf. Technol.*, vol. 2, no. 9, 2014.
- [25] M. Amaral, J. Polo, D. Carrera, I. Mohamed, M. Unuvar, and M. Steinder, "Performance Evaluation of Microservices Architectures Using Containers," 2015 IEEE 14th International Symposium on Network Computing and Applications. pp. 27–34, 2015.
- [26] M. Sneps-Snepe and D. Namiot, "Micro-service Architecture for Emerging Telecom Applications," *Int. J. Open Inf. Technol.*, vol. 2, no. 11, 2014.
- [27] S. Abrams, J. Kunze, and D. Loy, "An Emergent Micro-Services Approach to Digital Curation Infrastructure," *Int. J. Digit. Curation*, vol. 5, no. 1, 2010.
- [28] N. Dragoni, I. Lanese, S. T. Larsen, M. Mazzara, R. Mustafin, and L. Safina, "Microservices: How To Make Your Application Scale," 2017.

Energy Management Strategy of a PV/Fuel Cell/Supercapacitor Hybrid Source Feeding an off-Grid Pumping Station

Houssem CHAOUALI (*), Hichem OTHMANI, Mohamed Selméne BEN YAHIA, Dhafer MEZGHANI and Abdelkader MAMI

UR-LAPER, UR17ES11, Faculty of Sciences of Tunis,
University of Tunis El Manar, 2092 Tunis, Tunisia

Abstract—This work aims to develop an accurate energy management strategy for a hybrid renewable energy system feeding a pumping station. A developed model under Simulink environment is used to compare the performance of the pumping system when it is only fed by a photovoltaic generator, by a hybrid photovoltaic and fuel cell system and finally by a hybrid photovoltaic, fuel cell and a supercapacitor system. The developed control strategy is based on Fuzzy Logic control technique. Several simulations in different dramatic scenarios of working conditions show that the developed control strategy brought major enhancements in system performance and that the use of the supercapacitor makes economic profits by reducing the fuel cell production during critical solar irradiation periods.

Keywords—Energy management strategy; simulink; pumping station; photovoltaic generator; fuel cell generator; supercapacitor; fuzzy logic control technique

I. INTRODUCTION

In order to avoid more atmosphere pollution problems caused by conventional energy sources, scientists are continuously developing green energy sources and their applications which have touched almost every existing field such as transportation, domestic energy powering and industrial proceedings. Various environmentally friendly energy generators have been developed based on renewable energy sources, such as the sun and the wind, and although the major advantages they present, it is still impossible to avoid the dependence between this kind of power generation and weather conditions [1].

This dilemma imposed an open challenge for scientists which led to settling numerous solutions such as adding power electronics devices between the source and the load to adapt the flow between generated and requested energy. These adaptation systems are usually DC-DC converters [2], [3].

In case of an isolated, off-grid, areas where the renewable energy sources have become one of the most supplying solutions, studies focused on hybridizing different energy sources in order to ensure continuous production, so that, for example, photovoltaic (PV) generator can supply energy when wind is not available and a wind turbine can produce energy at moments of lack in solar irradiations. Of course, this hybrid settlement has many moments of zero energy production in case of absence of both wind and sun which is very common

scenario during the 24 hours of the day besides the problem of the high cost that this hybrid solution would impose using two types of generators at the same moment [4]-[7].

Thanks to the invention of the Fuel Cell (FC) technology and its rapid evolution, it became a stable, efficient and clean solution for continuous power generation [8]. Proton Exchange Membrane Fuel Cell (PEMFC) is a widely used type of FCs in different sectors especially transportation such as electrical cars. So the idea of using a PEMFC as a secondary source side by side with another classical renewable energy source, such as PV generator, has been studied in numerous works such as [9]-[11]. This hybrid solution provides a continuous clean and renewable energy production without, theoretically, dependence to climate conditions.

On the other hand, the cost of energy production would be much greater because of hydrogen consumption by the FC generator, which imposes the obligation of minimizing its activation as much as possible and avoid using it as primary source. In another hand, other studies propose to include a storage device along with this hybrid type of power generation to realize the economic objective [12]-[15]. Among these proposed solutions, [16] shows a study that proves economic profits by using a supercapacitor as a secondary source in an electric car mainly fed by PEMFC.

In this context, this study is investigating the best power generation topology for our system by comparing three possible topologies where the supervision of the generators is made by using a Fuzzy Logic (FL) energy management strategy that we developed for this purpose.

In a first place, this paper gives, in Section II, a presentation and modeling of the different used generators separately: the GSA-60 PV generator, the H-500 PEMFC and the Maxwell supercapacitor.

Then, in Section III, the three-phased pumping system which is composed by a voltage inverter, an asynchronous machine and a centrifugal pump is presented and modeled.

In the last, Section IV presents the developed FL control strategy for the two investigated topologies with hybridization of sources. The simulation results are presented in order to compare the system performance with these different combinations and the PEMFC utilization.

II. PRESENTATION AND MODELING OF THE DIFFERENT ENERGY SOURCES

A. Photovoltaic Generator

The main power source of the system, the photovoltaic generator, is a mixed parallel and series combination of a Kaneka GSA-060 module. In order to develop a model of this PV generator, we had to resort to the classical electrical presentation of it, shown in Fig. 1.

In another hand, Table 1 presents the different characteristics of the studied generator.

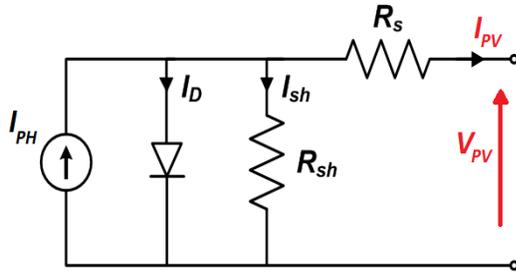


Fig. 1. Electrical presentation of a PV generator.

The generated current by the PV generator is expressed by (1): [16]

$$I_{PV} = I_{ph} - I_D - I_{Sh} \quad (1)$$

With:

$$I_{ph} = N_p \cdot [I_{CC} \frac{E}{E_r} + k_{isc} (T - T_r) \frac{E}{E_r}] \quad (2)$$

$$I_D = N_p \cdot I_s \left[\exp\left(\frac{V_{pv}}{N_s \cdot V_T}\right) - 1 \right] \quad (3)$$

$$I_{Sh} = \frac{V_{PV} + R_s I_{PV}}{R_{Sh}} \quad (4)$$

N_p : number of parallel strings.

N_s : number of modules in series.

TABLE I. KANEKA GSA-60 ARRAY FEATURES

Parameter	Value
N_p	2
N_s	5
P_{mpp}	600 W
V_{mpp}	335 V
I_{mpp}	1.8 A
V_{oc}	460 V
I_{sc}	2.38 A

Fig. 2 and 3 present the I-V and P-V characteristic curves for, respectively, a variable ambient temperature and a variable solar irradiance, of 2 parallel strings of 5 in-series connected panels.

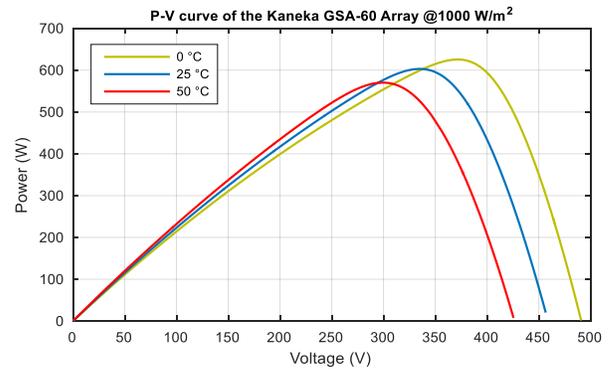
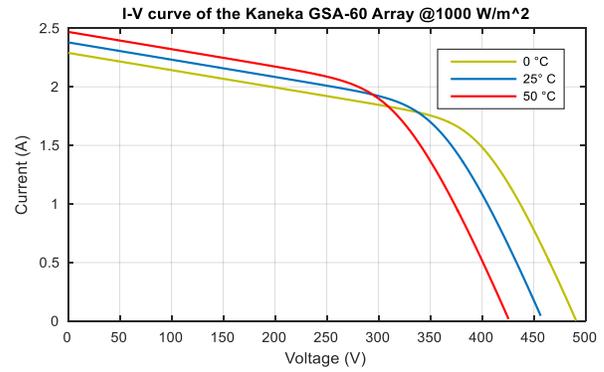


Fig. 2. The influence of ambient temperature variation on I-V and P-V characteristics of the Kaneka GSA-60 array.

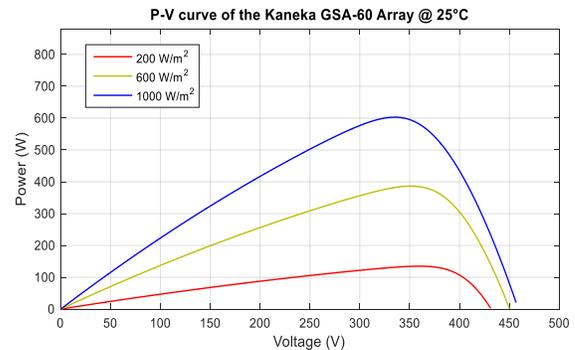
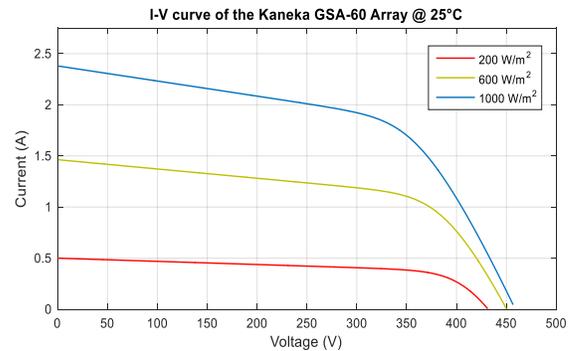


Fig. 3. The influence of solar irradiance variation on I-V and P-V characteristics of the Kaneka GSA-60 array.

B. PEM Fuel Cell

The PEM Fuel Cells are mainly composed of three parts. These parts are the Anode, the Cathode and, between these last two seats of chemical reactions, we find a conductive membrane called the Electrolyte which is the core of the PEMFC [17].

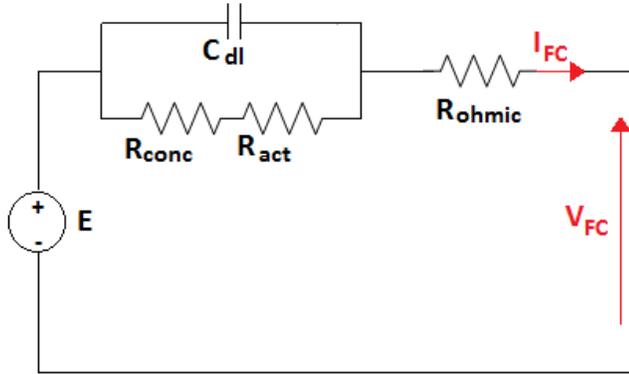


Fig. 4. PEMFC electrical circuit equivalent model.

where:

- R_{conc} : concentration Resistor.
- R_{act} : Activation Resistor.
- R_{ohmic} : Ohmic Resistor.
- C_{dl} : Double-Layer Capacitor.

Based on the electrical circuit presentation, in Fig. 4, and on Nernst equation as given in [18], [19], the PEMFC's generated voltage expression can be formulated as in (5).

$$V_{FC} = E - V_{con} - V_{act} - V_{ohm} \quad (5)$$

where:

- V_{FC} : Fuel Cell Output Voltage.
- E : Theoretical Potential of the Cell.
- V_{con} : Gases Concentration Voltage Losses given by (6).
- V_{act} : Activation Voltage Losses given by (7).
- V_{ohm} : Ohmic Voltage Losses given by (8).

$$V_{con} = -\frac{RT}{2.F} \ln \left(1 - \frac{I_D}{I_{Dmax}} \right) \quad (6)$$

$$V_{act} = \frac{RT}{2.\ell.F} \ln \left(\frac{I_{FC}}{I_0} \right) \quad (7)$$

$$V_{ohm} = I_{FC} \cdot R_{ohmic} \quad (8)$$

Where:

- R, T, F : perfect gas constant = 8.14 J/K/mol
- T : Operating temperature of the cell.
- F : Faraday constant = 96485 C/mol.
- I_D, I_{Dmax} : Current density and Current maximal density (A/cm^2)
- ℓ : Tafel slope for the activation losses.

- I_0 : Exchange current density during the activation (mA/cm^2).

For this work, we chose to use a model of a 500W PEM Fuel Cell Commercialized by FuelCellsEtc under the product code **H-500** which its different parameters are given in Table 2.

TABLE II. H-500 PEM FUEL CELL FEATURES

Parameter	Value
Rated Power	500 W
Number of Cells	24
Rated Performance	14.4V at 35A
Max Stack Temperature	60°C
Hydrogen Flow Rate at Maximum Output	6.5L/min
Hydrogen Pressure	0.45-0.55 bar
Hydrogen Purity Requirement	99.995 %
Start-Up Time	<= 30s

The next figures present the different characteristics of the PEMFC model that we are using, where, Fig. 5 presents the Voltage-Current characteristic curve and Fig. 6 presents the Power-Current characteristic curve.

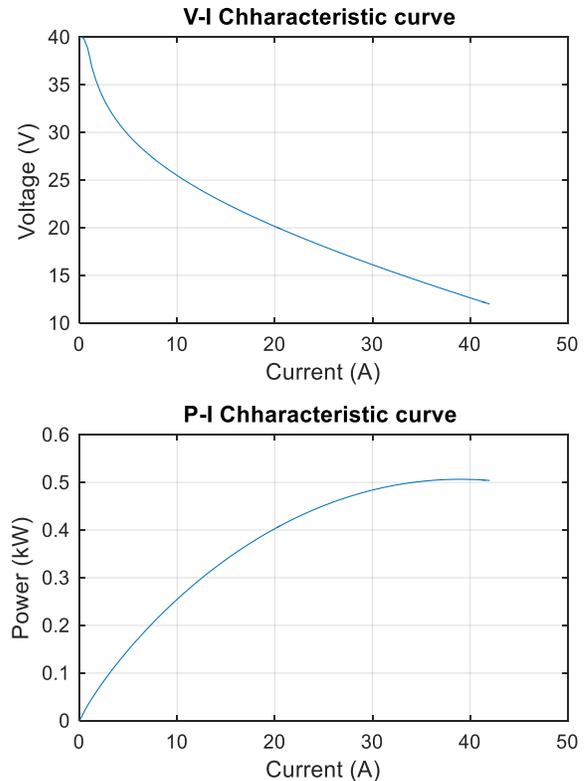


Fig. 5. V-I and P-I characteristics of the studied H-500 PEM fuel cell.

C. Super Capacitor

As a last energy source used in our system, the Maxwell BMOD0006-E160-B02 160V module is chosen and its different parameters are given in Table 3.

TABLE III. MAXWELL 160V ULTRA-CAPACITOR MODULE FEATURES

Parameter	Value
Rated Capacitance	5.8 F
Rated Voltage	160 V
Maximum Voltage	170 V
Maximum Current	170 A
Maximum ESR	240 mΩ
Maximum Stored Energy per Cell	0.35 Wh
Number of Cells	60

Fig. 6 presents the discharge equivalent circuit of the supercapacitor and its mathematical model is given in (9) as in [20].

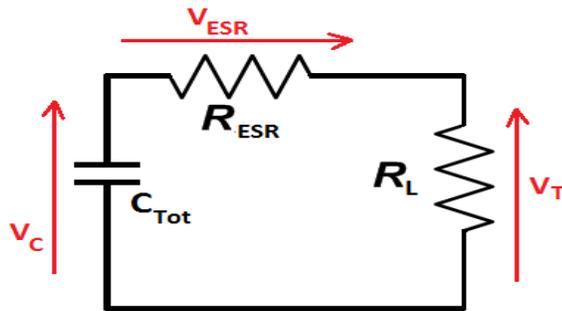


Fig. 6. Supercapacitor simplified discharge electrical circuit.

$$V_T(t) = V_C(t) + V_{ESR}(t) \quad (9)$$

Where:

- $R_{ESR} = N_S R_{esr}$ with N_S is the number of cells mounted in series and R_{esr} is single cell Equivalent Series Resistance.

- $C_{Tot} = \frac{1}{N_S} C_{Cell}$ with N_S is the number of cells mounted in series and C_{Cell} is single cell capacitance.

III. PRESENTATION AND MODELING OF THE PUMPING SYSTEM

A. General Overview of the Studied Load

The different sources we presented in the last sub-sections, are meant to supply a three-phased AC moto-pump, type Ebara Pra-050T, via a Moeller DV51 speed drive.

The speed drive, which is a three-phased inverter, is used to convert input voltage, either single phase AC voltage or DC voltage, into three-phased controllable AC voltages in order to control the speed of the asynchronous machine, also known as induction motor, which trains a centrifugal pump. By controlling the speed of the machine, we are controlling the flow rate of the water pumped by the moto-pump set [21].

Table 4 contains different technical specifications of both speed drive and moto-pump in study.

TABLE IV. TECHNICAL SPECS OF THE PUMPING SYSTEM

Moeller DV51 Speed Drive	
Maximum Power	2.2 KW
AC input	230 V
DC Input	400 V
Output Voltage	3 ~ 230 V
Ebara Pra-0.50T Moto-Pump	
Power	3 Hp ≈ 0.37 Kw
Voltage	3~ 240V
Nominal Current	1.8A
Frequency	50 Hz
P	1
Cos ρ	0.8
Maximum Speed	2850 rpm ≈ 300 rad/s
Maximum Flow rate	45 L/min

B. Modeling the Different Parts of the Pumping System

As explained, the load is composed of three different parts as shown in Fig. 7: a voltage inverter, a three-phased asynchronous machine, also called induction motor, and a centrifugal pump.

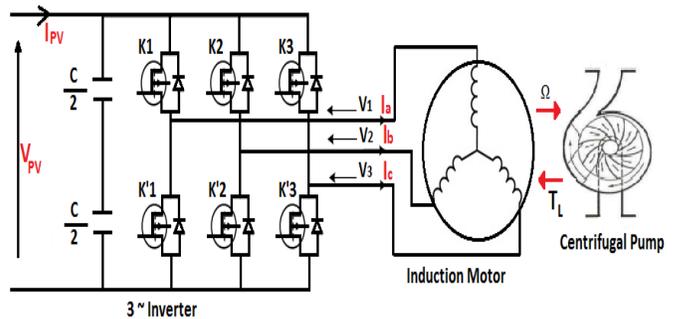


Fig. 7. 3~ pumping system equivalent model.

1) 3 Phased Speed Drive Modeling

Based on this electrical scheme, the different voltages can be expressed in (10) and the relation between the input and the three output currents is given by (11) [22].

$$\begin{bmatrix} V_1 \\ V_2 \\ V_3 \end{bmatrix} = \frac{V_{PV}}{3} \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix} \begin{bmatrix} K_1 \\ K_2 \\ K_3 \end{bmatrix} \quad (10)$$

$$I_{PV} = K_1 I_a + K_2 I_b + K_3 I_c \quad (11)$$

Where,

- I_{PV} and V_{PV} are respectively the current and voltage generated by the PV generator.

- $K_1, K_2, K_3, K'_1, K'_2$ and K'_3 : are the controlled switches of the 3 arms of the inverter.

2) Asynchronous Moto-Pump Modeling

The general presentation of the voltages at the stator in d, q frame is given by (12) and (13) [23].

$$V_{qs} = R_s \cdot I_{qs} + \frac{d\phi_{qs}}{dt} \quad (12)$$

$$V_{ds} = R_s \cdot I_{ds} + \frac{d\phi_{ds}}{dt} \quad (13)$$

Where,

ϕ_{qs} and ϕ_{ds} are the presentation of the stator flux in the d,q frame given by (14) and (15).

$$\phi_{qs} = L_s \cdot I_{qs} + M \cdot I_{qr} \quad (14)$$

$$\phi_{ds} = L_s \cdot I_{ds} + M \cdot I_{dr} \quad (15)$$

at the rotor of the machine, the voltage presentation is given by (16) and (17).

$$V_{qr} = R_r \cdot I_{qr} + \frac{d\phi_{qr}}{dt} \quad (16)$$

$$V_{dr} = R_r \cdot I_{dr} + \frac{d\phi_{dr}}{dt} \quad (17)$$

Where,

ϕ_{qr} and ϕ_{dr} are the presentation of the rotor flux in the d,q frame given by (18) and (19).

$$\phi_{qr} = L_r \cdot I_{qr} + M \cdot I_{qs} \quad (18)$$

$$\phi_{dr} = L_r \cdot I_{dr} + M \cdot I_{ds} \quad (19)$$

The mechanical model of the machine is expressed in (20).

$$T_{em} - T_L - f\Omega = J \frac{d\Omega}{dt} \quad (20)$$

Where,

f : Coefficient of viscous friction.

J : Inertia moment.

T_{em} : Electromagnetic torque.

Ω : Rotor speed.

T_L : Load torque.

The fact that the centrifugal pump presents a proportional relation between its resistive torque and the square of its speed, we can write the total electromagnetic torque of the moto-pump by replacing the new expression of T_L in (21) and finally obtain (21) [21];

$$T_{em} - K\Omega^2 - f\Omega = J \frac{d\Omega}{dt} \quad (21)$$

Where, K is The torque constant of the pump.

IV. PROPOSED ENERGY MANAGEMENT STRATEGY

A. Fuzzy Logic Control technique

Energy management in a system fed by hybrid power sources is important to realize an optimal energy production in term of generators life cycle (for the PEMFC), energy production cost, total and instantaneous cover of energy demand, etc. [16].

For that, numerous techniques have and still been used to achieve these different objectives. Among them, intelligent controllers such as FLC technique is presented as a reliable choice for this mission because since its first appearance in 1965, by Lotfi Zadeh, it witnessed rapid development and replaced almost all conventional techniques a wide range of applications [24].

A general working principle of an FLC based on Mamdani method is given by its flow chart in Fig. 8 [25].

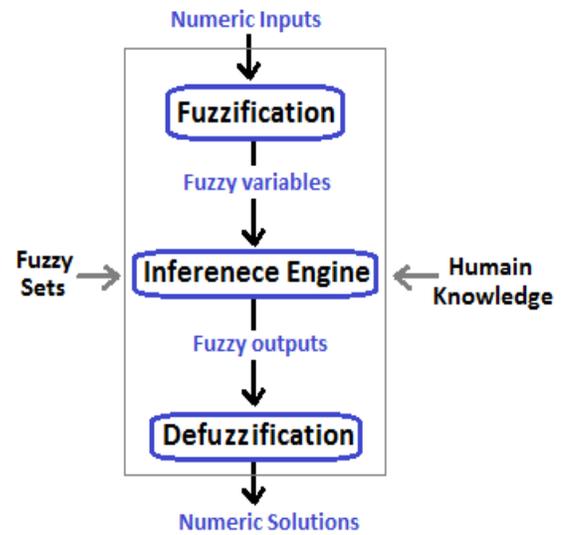


Fig. 8. Mamdani FLC working flow chart.

B. Developed FLC for Energy Management

We developed a management strategy based on FLC technique for two proposed topologies to compare the results. The difference between the two topologies is the existing of the supercapacitor in order to investigate its role in minimizing the PEMFC power generation, and thus, hydrogen consumption and total power production cost [16].

1) Topology 1: Standard PV-Pumping System

This topology is the standard system in study. It is composed of the PV generator supplying the load via the different converters and their controllers (MPPT and Speed).

2) Topology 2: Without Supercapacitor

This topology contains two energy sources, PV generator and the PEM fuel cell, and the load. The developed fuzzy algorithm for energy management in this topology uses two inputs given in (22) and (23) in order to determine the output which is one of the predefined modes in order to select the proper working configuration of the system.

$$P_{demand} = P_{PV} - P_{Load} \quad (22)$$

$$dP_{demand} = P_{demand}(k) - P_{demand}(k-1) \quad (23)$$

The different modes of system configuration are explained as next:

- Mode 1: The GPV is the only energy source and the PEMFC is not connected.
- Mode 2: The PEMFC is activated and the GPV is not connected.
- Mode 3: The GPV and PEMFC are both activated and must make sure that the GPV is the main generator and the PEMFC only works to compensate the lack of energy demanded by the load.

3) Topology 3 : with Supercapacitor

This topology contains the two energy sources, PV generator and the PEM fuel cell, along with the storage device, the supercapacitor, and the three-phased pumping system. The developed fuzzy algorithm for energy management in this topology uses an additional 3rd input which is the State Of Charge (SOC) of the supercapacitor besides the same two inputs previously given by (22) and (23) in order to determine the proper working mode and thus selecting the optimal working configuration of the system.

Different modes of system configuration are set as next:

- Mode 1: The GPV is the only energy source, the PEMFC and the Supercapacitor are not connected.
- Mode 2: The GPV is activated, the Supercapacitor is charging (condition: SOC<50%) and the PEMFC is not connected.
- Mode 3: The GPV is activated, the Supercapacitor is discharging (condition: SOC>50%) and the PEMFC is not connected.
- Mode 4: The GPV is activated, the Supercapacitor is discharging (condition: SOC>50%) and the PEMFC is activated.

All the modes must maintain these conditions: The main source must be the GPV, the supercapacitor only charges when the GPV power exceeds the demanded power, the PEMFC is only activated when the GPV power is not enough for load demand and the supercapacitor is discharged.

C. Results and Discussion

Fig. 9 shows the speed response with different studied topologies in constant irradiance equal to 1000 W/m².

Fig. 10 shows the speed response with different studied topologies in variable irradiance where the imposed simulation scenario is:

- From t=0s to t = 5s: constant at 1000W/m².
- At t = 5s: variation from 1000 W/m² to 200W/m².
- From t=5s to t = 10s: constant at 200W/m².

- At t = 10s: sudden variation of irradiance from 200W/m² to 800W/m².
- From t=10s to t = 15s: constant at 800W/m².

Fig. 9 and 10 proves that topology 3 has better performance in both variable and constant irradiance conditions.

Fig. 11 presents the on/off status of the PEMFC with both topologies 2 and 3 in the case of variable irradiance.

The obtained results show that even for a high irradiance (1000 W/m²), there has been a frequent activation of the PEMFC in topology 2 and a continuous activated status in low irradiance value. In another hand, with topology3, the PEMFC has not been activated even when the irradiance is at 200 W/m².

This can be explained by the fact that the GPV can deliver the needed power by itself when the irradiance is high and the supercapacitor, initially charged at 100%, would deliver the additional needed power when the irradiance is low.

Fig. 12 presents the evolution of SOC of the Supercapacitor initially charged. At the start, even the irradiance is high, the SOC makes a quick drop because of the nature of the load which demands more energy in its starting phase. Then, we can see clearly that during the low irradiance period, the supercapacitor is discharging to provide the difference between the generated power by the PV generator and load demand. After that, when the irradiance reaches high values in a second time, the supercapacitor is charging and its SOC is rising.

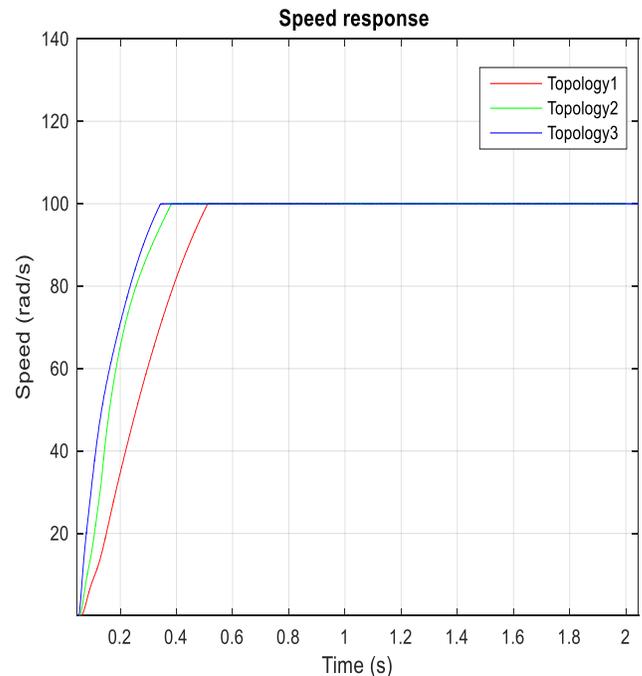


Fig. 9. Speed response with different topologies for constant irradiance.

V. CONCLUSION

The different results show a better performance of the system, when using the topology with photovoltaic generator, PEMFC and supercapacitor, in both constant and variable irradiance and a major economic profit because of the supercapacitor charge/discharge cycles supervision in by the Fuzzy Logic management strategy.

This work can be enhanced by adding Lithium Batteries bank to have a hybrid storage system which will ensure that the utilization of the PEMFC will be reduced even in long periods of solar irradiance absence, and thus less fuel consumption. In the control side, the management strategy can be based on Economic Model Predictive Control (EMPC) technique which is a new developed form of the classic MPC control technique. The investigation of the system performance by using this new storage scheme and control approach will make the subject of future works.

REFERENCES

- [1] M. H. Nehrir, C. Wang, K. Strunz, H. Aki, R. Ramakumar, J. Bing, Z. Miao, and Z. Salameh, "A review of Hybrid Renewable/Alternative Energy Systems for Electric Power Generation: Configurations, Control, and Applications", *IEEE Transactions On Sustainable Energy*, 2(4), pp. 392-403, November 2011. <http://dx.doi.org/10.1109/TSTE.2011.2157540>
- [2] JABALLAH Mohamed Akram, MEZGHANI Dhafer and MAMI Abdelkader, "Design and Simulation of Robust Controllers for Power Electronic Converters used in New Energy Architecture for a (PVG)/(WTG) Hybrid System", *International Journal of Advanced Computer Science and Applications(IJACSA)*, 8(5), 2017. <http://dx.doi.org/10.14569/IJACSA.2017.080531>
- [3] CHAOUALI Housseem, OTHMANI Hichem, MEZGHANI Dhafer, JOUINI Houda and MAMI Abdelkader, "Fuzzy logic control scheme for a 3 phased asynchronous machine fed by Kaneka GSA-60 PV panels", *IEEE International Renewable Energy Congress(IREC)*, Tunisia 2016. <https://doi.org/10.1109/IREC.2016.7478893>
- [4] KAABECHE Abdelhamid and IBTIOUEN Rachid, "Techno-economic optimization of hybrid photovoltaic/wind/diesel/battery generation in a stand-alone power system", *Solar Energy*, 103, pp. 171-182, 2014. <https://doi.org/10.1016/j.solener.2014.02.017>
- [5] Binayak Bhandari, Kyung-Tae Lee, Caroline Sunyong Lee, Chul-Ki Song, Ramesh K. Maskey and Sung-Hoon Ahn, "A novel off-grid hybrid power system comprised of solar photovoltaic, wind, and hydro energy sources", *Applied Energy*, 133, pp. 236-242, 2014. <https://doi.org/10.1016/j.apenergy.2014.07.033>
- [6] Ajay Kumar Bansal, Rajesh Kumar and R. A. Gupta, "Economic analysis and power management of a small autonomous hybrid power system (SAHPS) using biogeography based optimization (BBO) algorithm", *IEEE Transactions on Smart Grid*, 4(1), pp. 638-648, 2013. <http://dx.doi.org/10.1109/TSG.2012.2236112>
- [7] Lei Zhang and Yaoyu Li, "Optimal Energy Management of Wind-Battery Hybrid Power System With Two-Scale Dynamic Programming", *IEEE Transactions on Sustainable Energy*, 4(3), pp. 765-773, 2013. <https://doi.org/10.1109/TSTE.2013.2246875>
- [8] R. Chedid, H. Akiki, and S. Rahman, "A decision support technique for the design of hybrid solar-wind power systems," *IEEE Transactions on Energy Conversion*, vol 13(1), pp. 76-83, 1998.
- [9] PASKA Józef, BICZEL Piotr and KIOS Mariusz, "Hybrid power systems – An effective way of utilising primary energy sources", *Renewable Energy*, 34(11) pp. 2414-2421, 2009. <https://doi.org/10.1016/j.renene.2009.02.018>
- [10] Mehmed Eroglu, Erkan Dursun, Suat Sevencan, Junseok Song, Suha Yazici and Osman Kilic, "A mobile renewable house using

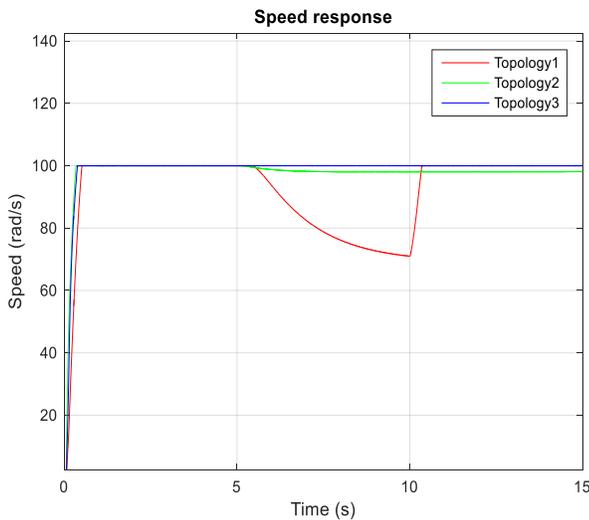


Fig. 10. Speed response with different topologies for variable irradiance.

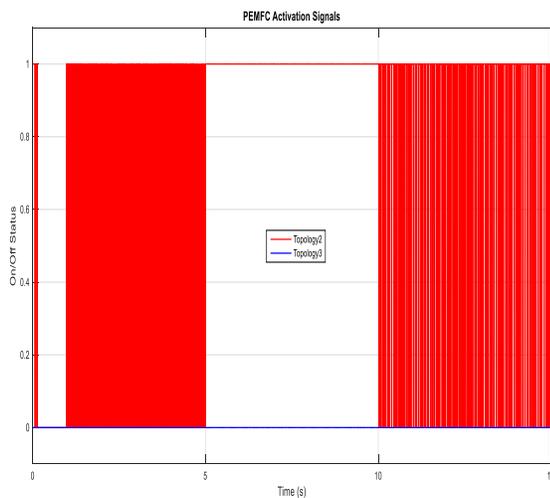


Fig. 11. PEMFC activation signal for a variable irradiance.

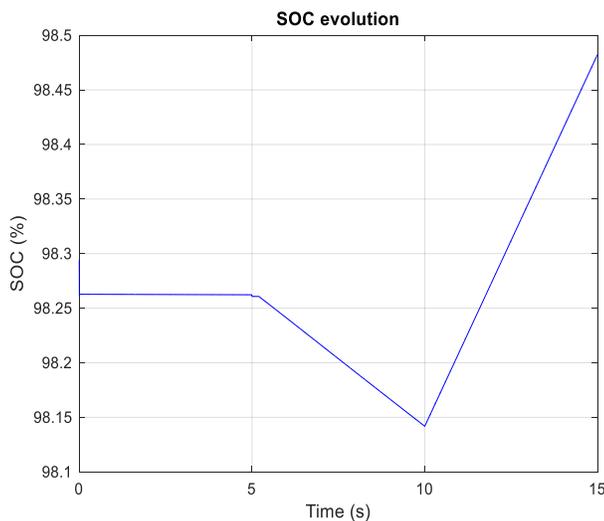


Fig. 12. SOC evolution of the super capacitor initially charged.

- PV/wind/fuel cell hybrid power system”, International Journal of Hydrogen Energy, 36(13), pp. 7985-7992, 2011.
<https://doi.org/10.1016/j.ijhydene.2011.01.046>
- [11] O.C. Onar, M. Uzunoglu, M.S. Alam and Dynamic modelling, “Design and Simulation of a Wind/Fuel Cell/Ultra-Capacitor-Based Hybrid Power Generation System”, Journal of Power Sources, 61(1), pp. 707-722, 2006.
<https://doi.org/10.1016/j.jpowsour.2006.03.055>
- [12] Phatiphat Thounthong, Arkhom Luksanasakul, Poolsak Koseeyaporn and Bernard Davat, “Intelligent Model-Based Control of a Standalone Photovoltaic/Fuel Cell Power Plant With Supercapacitor Energy Storage”, IEEE Transactions on Sustainable Energy, 4(1), pp. 240-249, 2013.
<http://dx.doi.org/10.1109/TSTE.2012.2214794>
- [13] Akbar Maleki and FathollahPourfayaz, “Sizing of stand-alone photovoltaic/wind/diesel system with battery and fuel cell storage devices by harmony search algorithm”, Journal of Energy Storage, 2, pp.30-42, 2015.
<https://doi.org/10.1016/j.est.2015.05.006>
- [14] Hassan El Fadil, Fouad Giri, Josep M. Guerrero and Abdelouahad Tahri, “Modeling and Nonlinear Control of a Fuel Cell/Supercapacitor Hybrid Energy Storage System for Electric Vehicles”, IEEE Transactions on Vehicular Technology, 63(7), pp. 3011 - 3018, 2014.
<http://dx.doi.org/10.1109/TVT.2014.2323181>
- [15] MehdiAnsarey, Masoud Shariat Panahi, Hussein Ziarati and Mohammad Mahjoob, “Optimal energy management in a dual-storage fuel-cell hybrid vehicle using multi-dimensional dynamic programming”, Journal of Power Sources, 250, pp. 359-371, 2014.
<https://doi.org/10.1016/j.jpowsour.2013.10.145>
- [16] ANDARI Wahib, GHOZZI Samir, ALLAGUI Hatem and MAMI Abdelkader, “Design, Modeling and Energy Management of a PEM Fuel Cell / Supercapacitor Hybrid Vehicle” International Journal of Advanced Computer Science and Applications(IJACSA), 8(1), 2017.
<http://dx.doi.org/10.14569/IJACSA.2017.080135>
- [17] BEN SALEM W., MZOUGH D., ALLAGUI H. and MAMI A., “The bond graphs to the study of interactions between the PEM fuel cell and static converters”, 17th International Conference on Sciences and Techniques of Automatic Control and Computer Engineering (STA), pp.423-428, Tunisia 2016.
<http://dx.doi.org/10.1109/STA.2016.7952004>
- [18] BEN YAHIA Mohamed Sélmene, ALLAGUI Hatem, BOUAICHA Arafet and MAMI Abdelkader, “Fuel Cell Impedance Model Parameters Optimization using a Genetic Algorithm”, International Journal of Electrical and Computer Engineering(IJECE), 7(1), pp. 184-193, 2017.
- [19] ALLAGUI Hatem, MZOUGH Dhaia, BOUAICHA Arafet and MAMI Abdelkader, “Modeling and Simulation of 1.2 kW Nexa PEM Fuel Cell System”, Indian Journal of Science and Technology, 9(9), 2016.
<http://dx.doi.org/10.17485/ijst/2016/v9i9/85299>
- [20] A. B. Cultura II, Z. M. Salameh, “Modeling, Evaluation and Simulation of a Supercapacitor Module for Energy Storage”, International Conference on Computer Information Systems and Industrial Applications (CISIA 2015), Thailand 2015.
- [21] MEZGHANI Dhafer, OTHMANI Hichem, SASSI Fares, MAMI Abdelkader and DAUPHIN-TANGUY Geneviève, “A New Optimum Frequency Controller of Hybrid Pumping System: Bond Graph Modeling-Simulation and Practice with ARDUINO Board” International Journal of Advanced Computer Science and Applications(IJACSA), 8(1), 2017.
<http://dx.doi.org/10.14569/IJACSA.2017.080112>
- [22] MEZGHANI Dhafer, OTHMANI Hichem, SASSI Fares, MAMI Abdelkader and G. Dauphin-Tanguy, “A New Optimum Frequency Controller of Hybrid Pumping System: Bond Graph Modeling-Simulation and Practice with ARDUINO Board”, International Journal of Advanced Computer Science and Applications(IJACSA), 8(1), pp. 78-87, 2017.
<http://dx.doi.org/10.14569/IJACSA.2017.080112>
- [23] OTHMANI Hichem, SASSI Fares, MEZGHANI Dhafer and MAMI Abdelkader, “Comparative Study between Fuzzy Logic Control and Sliding Mode Control for Optimizing the Speed Department of a Three Phase Induction Motor”, International Review of Automatic Control, 9(3), pp. 175-181, 2016.
<https://doi.org/10.15866/ireaco.v9i3.9269>
- [24] N. Ammasai Gounden, Sabitha Ann Peter, Himaja Nallandula, S.Krithiga, “Fuzzy logic controller with MPPT using line-commutated inverter for three-phase grid-connected photovoltaic systems”, Renewable Energy, 34(3), pp. 909-915, 2009.
<http://dx.doi.org/10.1016/j.renene.2008.05.039>
- [25] Housseem Chaouali, Hichem Othmani, Dhafer Mezghani and Abdelkader Mami, “Enhancing classic IFOC with Fuzzy Logic technique for speed control of a 3~ Ebara Pra-50 moto-pump”, 17th International Conference on Sciences and Techniques of Automatic Control and Computer Engineering (STA), pp.423-428, Tunisia 2016.
<http://dx.doi.org/10.14569/10.1109/STA.2016.7951985>

Object's Shape Recognition using Local Binary Patterns

Muhammad Wasim

Department of Computer Science
Usman Institute of Technology
Karachi, Pakistan

Abdul Aziz

Department of Computer Science
Hamdard University
Karachi, Pakistan

Syed Faisal Ali

Department of Computer Science
Usman Institute of Technology
Karachi, Pakistan

Adnan Ahmed Siddiqui

Department of Computer Science
Hamdard University
Karachi, Pakistan

Lubaid Ahmed

Department of Computer Science
Usman Institute of Technology
Karachi, Pakistan

Fauzan Saeed

Department of Computer Science
Usman Institute of Technology
Karachi, Pakistan

Abstract—This paper discusses the concept of object's shape identification using local binary pattern technique (LBP). Since LBP is computationally simple it has been utilized successfully for recognition of various objects. LBP which has the potential to be used in various identification related fields was applied on a number of different shaped objects, the process converted the given image in to 3x3 binary matrices and several rounds of computation yields the final decision parameter, which is known as merit function. This parameter was then exploited to uniquely identify the shape of different objects.

Keywords—Local binary patterns; object shape recognition; security technologies; content based recognition

I. INTRODUCTION

Today people all around the world are facing a number of challenges related to health, education and specially security. Almost each country of the world is facing different type of threats and indiscipline activities. Because of this unsecure environment and criminal activities hundreds of people are harmed and killed, daily. Terminating, fire, murder and bomb impact are normal exercises nowadays. Security agencies are trying their level best to encounter such types of threats. Object shape recognition is a standout amongst the most difficult and demanding territory of research nowadays. There are numerous applications and research works have been carried out to recognize the shape of objects. Using surveillance camera as a part of open spots, air terminals, lodgings and markets, objects recognition turns out to be more useful technology to maintain a strategic distance from any criminal occurrences in these territories. Object's shape recognition framework required just the shape of any object regardless of various hues, color, size or patterns. Automated computerized object shape recognition is not an easy task.

The proposed work discussed the object's shape recognition system using LBP technique. The object recognition system described in this paper was divided into two modules: Image Registration Module and Image

Identification Module. In first module of system, image of an object, whose shape was required to be recognized was captured using a digital camera and image was then stored in computer system database for the purpose of identification and profiling. After filtering, the LBP technique was applied on captured image to produce a merit function to recognized objects. This merit function was stored into system database along with the actual image of an object. In second module the shape of different objects were recognized as shown in Fig. 1.

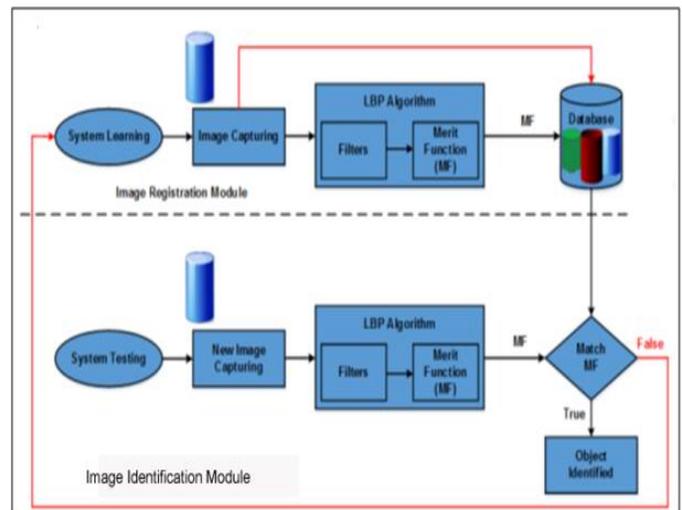


Fig. 1. System modules of proposed work.

Once image was captured, background subtraction technique was applied to extract the object. Using re-sizing method, a captured image was transformed into a standard sized image and then converted into a binary image. Using LBP technique, which was based on image matrix information, a merit function was solved that can be used as a decision parameter to identify the shape of an object. Fig. 2 shows the working of all phases.

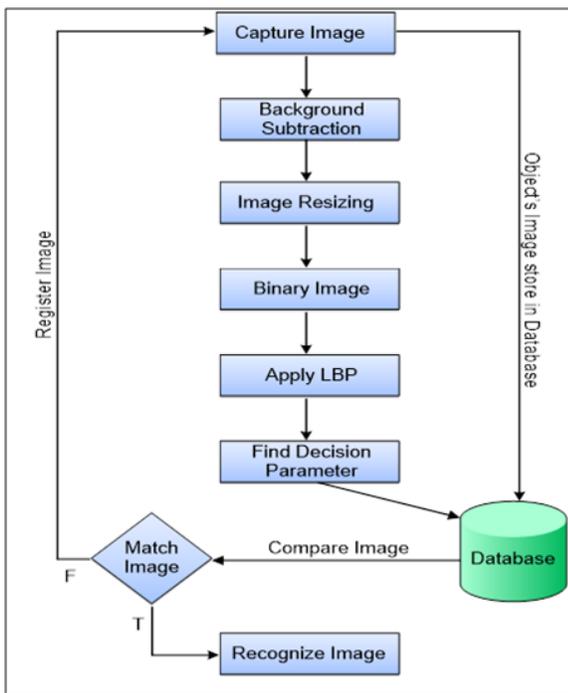


Fig. 2. System working diagram.

II. LOCAL BINARY PATTERN

LBP technique is very modest and exceptionally productive method for texture cataloging of diverse objects. In this efficient technique of LBP, the binary patterns of object are converted into some equivalent numeric numbers, which can be treated as decision parameter to recognize an object. At initial level this method was discussed a corresponding way to analyze the contrast of pictures. Key embodiment of this technique was set as 8 neighboring pixels, considering the center pixel as core. A mathematical model, which was based on neighbor pixels with respect to center pixel, provided the weights of the network, which was finally totaling the results. In given Fig. 3 the process to calculate merit function (decision parameter) is discussed. For this purpose, a wooden square piece was selected an object. A matrix of order 3×3 was generated in the first step. By applying LBP technique, a merit function was generated, which calculated a value of 112. Same process was applied for other matrices and finally a decision parameter was solved to recognize that object.

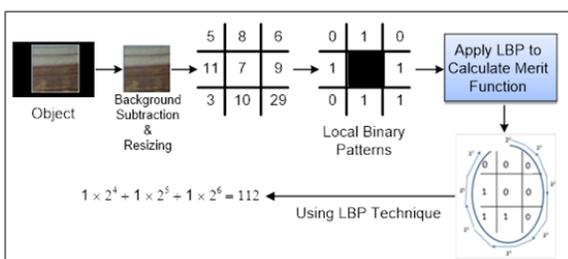


Fig. 3. Calculation of merit function.

A histogram of object's binary pattern also validated the process of verification of different objects. LBP generated various binary codes of different shapes of objects. Using mathematical technique, these binary patterns converted into

decision parameter (a numeric value) for the purpose of recognition. LBP has turned out to be by and large utilized as a part of image processing and computer vision areas because of its high discriminative strength, broadmindedness in contradiction of light variations and computational easiness. Some very common LBP applications are:

- Object's feature extraction to find the nature of shape.
- Textures classification and segmentation of different objects [1].
- Pattern recognition of different objects [2], [3].
- To analyze biomedical images.
- To extract the features of human face for identification [4]-[6].

The LBP technique has the potential to use in many applications with an acceptable accuracy. Its computational complexity to identify shape of different objects is low. This technique has no major impact with pose variation of objects or change in illumination [7], [8]. Logically LBP applications can be classified into two domains — local and global. A global approach of LBP is used to identify the shape of objects and local approach of LBP or the combination of both approaches provided detail information and can be used to recognized human faces.

III. LITERATURE REVIEW

Object recognition is one of the major areas of interest for researcher these days. There are a number of research contributions by the researchers. As for LBP is concerned, this technique is used in variety of applications such as face recognition, lung's cancer detection, prediction of facial age of human, etc.

A bottom-up approach is presented in [9] to identify nature of fascinated objects. To develop an operative and vigorous identification technique, the suggested methodology is accomplished by extracting the features of objects. In [10], authors described an effective way to identify several objects from images by means of a region resemblance identification is offered. In this technique objects are segmented into regions based on identical features. In [11], authors showed an effective technique to identify specific object from an image. The concept is based on a mixture of certain operators (contains a set of conventional parameters). To get the appropriate results, these parameters are required to adjust in a specific order. In [12], a relative learning is presented for two-dimensional and three-dimensional using LBP technique to diagnose lungs cancer from CT scan images. The technique was tested on a number of lungs CT images from "Japan Society of Computer Aided Diagnosis of Medical Images". Authors of [13] presented a method of face retrieval based on LBP technique. The key idea is to identify significant faces from the huge datasets using content based approach instead of metadata. An implementation of a vigorous face detection technique based on Integral-Haar-histograms with Circular-multi-block Local Binary Operator with comparatively better efficiency is presented in [14]. Hierarchical-age-estimation method is proposed in [15], which comprises local and global

information of human faces. LBP technique was used to extract the local facial features of human faces.

IV. METHODOLOGY

In this paper, the LBP technique is used to identify the shape of different objects. To achieve this goal a standard object image is first divided into small segments. Using the concept of thresh-holding LBP set an image in 3×3 order matrices. Each matrix generated a specific value, which were stored in another temporary matrix. Same operation was applied to temporary matrix and finally generated a merit function, which returned a constant value. This constant value was changed with respect to different shape of objects and treated as a main source of object recognition. The details of this methodology to calculate merit function of different objects are described in Section VI. A histogram of binary patterns before and after processing was also generated for the purpose of validation [16], as given in Fig. 4.

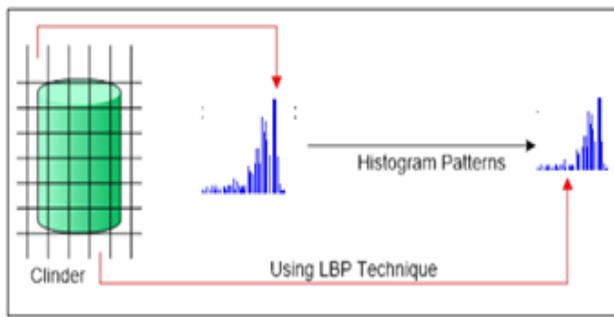


Fig. 4. Histogram of LBP.

This technique may be solved by considering divers neighbor pixel sizes. It depends upon the size of image to select 4, 8 or 16 neighbor pixels patterns for LBP as shown in Fig. 5. Binary patterns values (v) were generated for different patterns, for example $x_1^{(i)}$ is the first binary value generated by first neighbor pixel in 16 neighbor pixels patterns. Similarly $x_n^{(i)}$ is the last binary value generated by LBP. The basic local binary value using all 16 neighbor pixels can be generated using (1).

$$v = \sum_{k=1}^i x_k \quad (1)$$

Where 'i' is the total number of neighbor pixels under study.

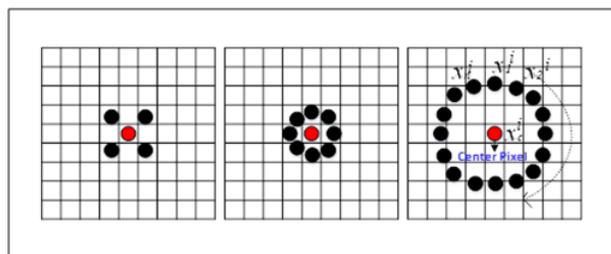


Fig. 5. Neighbor Pixels (4, 8 and 16) in LBP Technique.

The concept of LBP was demonstrated in [17], [18] by hypothesis that a surface consists of two corresponding features — the patterns and its anatomy.

V. PHASES AND WORKING OF THE SYSTEM

Following are the work example of LBP technique tested in Image Processing Research Lab (IPRL). The LBP technique applied on a number of objects and some of them (Cylinder, sphere, square and triangular wooden objects) are reported here. Fig. 6(a) shows the working of system using LBP technique. Fig. 6(b) described the calculation of merit function along with histogram patterns of cylinder object. In this figure an image of a cylindrical was captured, cropped and then converted into binary image. The LBP technique was applied using 3×3 neighborhood pixels of the image and generated a histogram along with the merit function for object (cylinder). Similarly, the same process was applied on other objects like square, sphere and triangular wooden objects to calculate their merit functions and corresponding histogram patterns, as shown in Fig. 6(b)-(d), respectively.

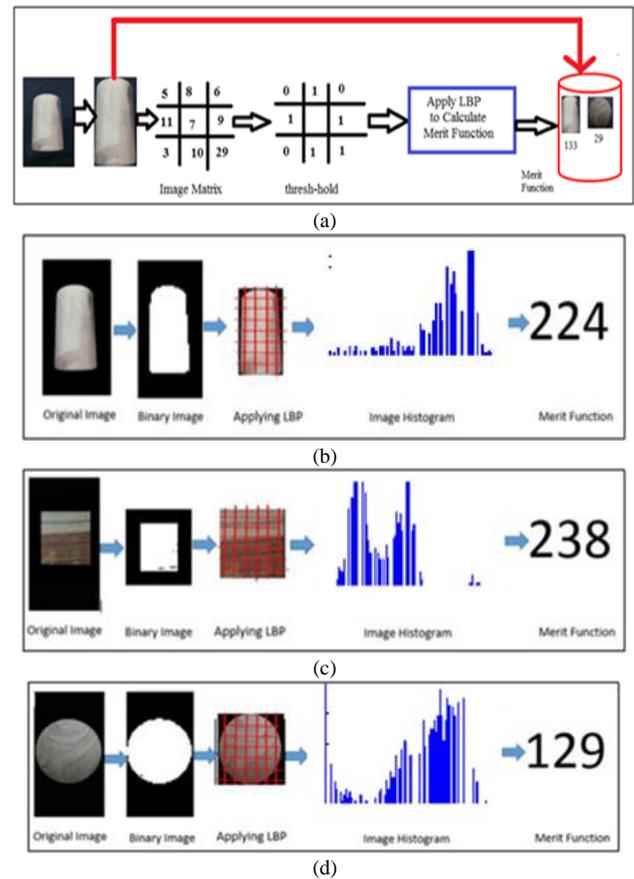


Fig. 6. (a). System working diagram using LBP technique, (b). Generation of merit function for cylinder object, (c). Generation of merit function for square object, (d). Generation of merit function for sphere object

VI. CALCULATION OF MERIT FUNCTION

At starting the design algorithm first selected a (3×3) matrix from captured image matrix. Then compared the center pixel value of the (3×3) matrix with its neighborhood values and convert the matrix in binary form by describing a condition (threshold); the value that is less than or equals to the center value will be assigned '0' and the greater value will be assigned '1'. The new value can be obtained through the calculation of the binary matrix. The matrix is added in a

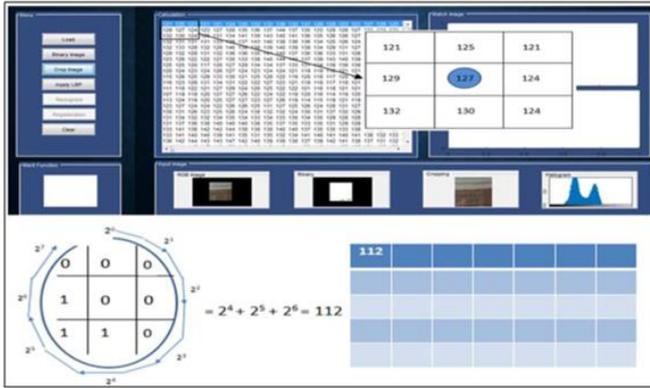
clockwise manner by the increasing powers of each code starting from 0 and ranging to 7 by using the formula:

$$x = 20 + 21 + 22 + 23 + 24 + 25 + 26 + 27$$

$$x = 0 + 2 + 0 + 8 + 0 + 0 + 0 + 0$$

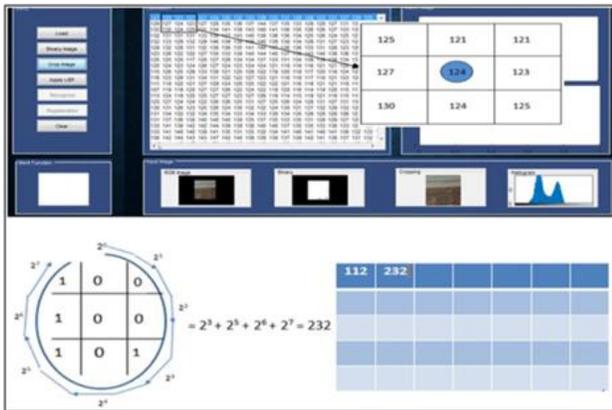
$$x = 10$$

Now the calculated value will be inserted into the first index of the new matrix, as shown in Fig. 7(a).



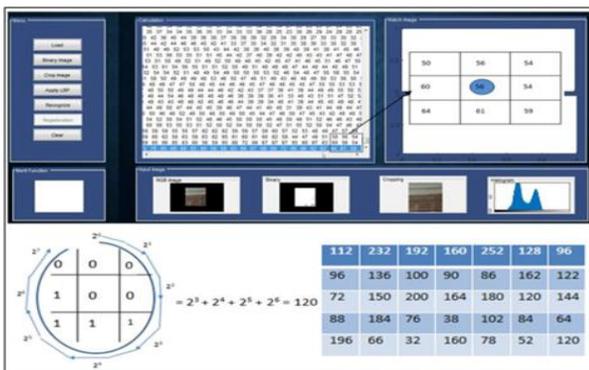
(a)

Then moving towards the second step, shift one index right and apply the same technique as described in the first step, and given in Fig. 7(b).



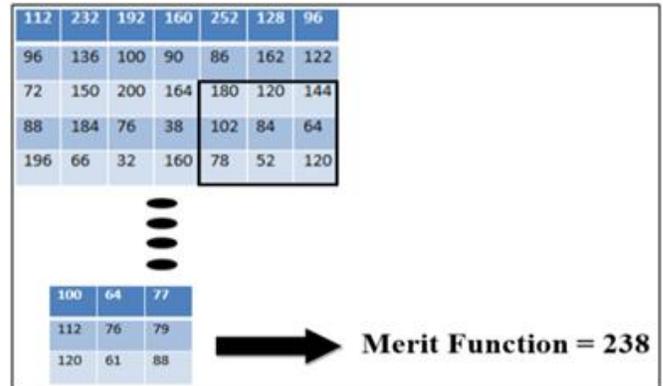
(b)

Continue shifting one by one pixel using same technique for 3 × 3 matrix and stored the value in a designed matrix. A final updated matrix along with all calculated values is shown in Fig. 7(c).



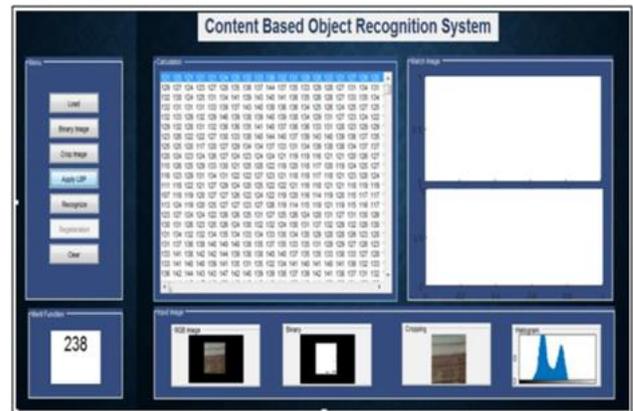
(c)

Now, pick the 3×3 matrix from the updated matrix and apply the same technique and generated new calculated values. Repeated same for all the matrix values and generated a final merit function. This process is shown in Fig. 7(d).



(d)

Based on all calculated values the merit function of the object is used to recognize the objects, as shown in Fig. 7(e).



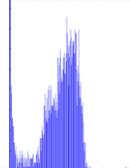
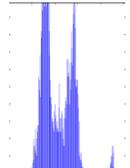
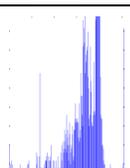
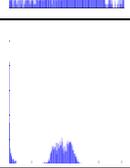
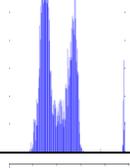
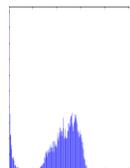
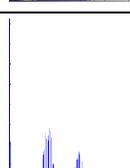
(e)

Fig. 7. (a) First index calculation of LBP Technique, (b) Second index calculation of LBP Technique, (c) Final updated matrix along with all calculated values, (d) Calculation of merit function, (e) Calculate merit function.

VII. RESULTS

The LBP system recognized the objects on the basis of the respective calculated merit function. The system was tested on different objects (sphere, cylinder, square, and triangular). The merit function value based on designed LBP technique was found as: Sphere=129, Square=238, Cylinder=224, Triangle=131, Rectangle=230, Ellipse=210 and Hexagonal=155. The results showed reasonable differences between the merit function values for different objects for the purpose of object's shape recognition. These results are summarized in Table 1 below:

TABLE I. RESULTS OF DIFFERENT OBJECTS

S No	Objects	Image	Merit Function	Histogram
1	Sphere		129	
2	Square		238	
3	Cylinder		224	
4	Triangle		131	
5	Rectangle		230	
6	Ellipse		210	
7	Hexagonal		155	

VIII. CONCLUSION

The object recognition based on LBP method was an operative technique in the domain of security, surveillance, medical and industrial applications. The system based on LBP is very simple and reliable technique to recognize the shape of different objects.

IX. FUTURE WORK

In contents of future direction, this concept of LBP can be used in a number of applications such as to find the face

symmetry of human faces, to analyze the back shape symmetry of human body, content based object recognition, etc.

REFERENCES

- [1] Wang, L., He, D.C.: Texture classification using texture spectrum. Pattern Recognition. pp. 905–910 (1990).
- [2] Mäenpää, T.: The local binary pattern approach to texture analysis—extensions and applications. PhD thesis, Acta Universitatis Ouluensis C 187, University of Oulu (2003).
- [3] A. Hadid, T. Ahonen, M. Pietikinen, "Face analysis using local binary patterns" in Handbook of Texture Analysis, U.K., London:Imperial College Press, pp. 347-373, 2008.
- [4] T. Ahonen ; A. Hadid ; M. Pietikainen. Face Description with Local Binary Patterns: Application to Face Recognition. IEEE Transactions on Pattern Recognition, 28(12) (2006).
- [5] W. Y. Zhao, R. Chellappa, A. Rosenfeld, and P. J. Phillips. Face recognition: A literature Survey. UMD CfAR Technical Report CAR-TR-948, Center for Automation Research, University of Maryland, (2002).
- [6] Rahim, Md Abdur, et al. "Face recognition using local binary patterns (LBP)." Global Journal of Computer Science and Technology (2013).
- [7] T. Ahonen, A. Hadid & M. Pietikäinen, Face Description with Local Binary Patterns: Application to Face Recognition. Draft, (2006).
- [8] T. Ojala, M. Pietikäinen & T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with Local Binary Patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence. 24(7), (2002): 971-987.
- [9] Chenini, Hanen. A Bottom-up Approach for Visual Object Recognition on FPGA based Embedded Multiprocessor Architecture. International Journal of Advanced Computer Science and Applications (IJACSA), 8(5) (2017): 474-482.
- [10] El Idrissi, Abdellatif, et al. A Multiple-Objects Recognition Method Based on Region Similarity Measures: Application to Roof Extraction from Orthophotoplans. International Journal of Advanced Computer Science and Applications (IJACSA), 6(11) (2015): 292-303.
- [11] Qaffou, Issam, Mohamed Sadgal, and Aziz Elfazziki. A New Automatic Method to Adjust Parameters for Object Recognition. arXiv preprint arXiv:1211.6971. International Journal of Advanced Computer Science and Applications (IJACSA), (2012)..
- [12] Arai, Kohei, Yeni Herdiyeni, and Hiroshi Okumura. Comparison of 2D and 3D local binary pattern in lung cancer diagnosis. International Journal of Advanced Computer Science and Applications (IJACSA), 3(4) (2012): 89-95.
- [13] Khoi, Phan, Lam Huu Thien, and Vo Hoai Viet. Face Retrieval Based On Local Binary Pattern and Its Variants: A Comprehensive Study. International Journal of Advanced Computer Science and Applications (IJACSA), 7(6) (2016): 249-258.
- [14] Suri, P. K., and Er Amit Verma. Robust face detection using circular multi block local binary pattern and integral haar features. International Journal of Advanced Computer Science and Applications (IJACSA), Special Issue on Artificial Intelligence (2011).
- [15] Günay, Asuman, and Vasif V. Nabiyev. Facial Age Estimation based on Decision Level Fusion of AAM, LBP and Gabor Features. International Journal of Advanced Computer Science and Applications (IJACSA), 6(8) (2015).
- [16] Mäenpää, T., Pietikäinen, M.: Texture analysis with local binary patterns. In: Chen, C.H., Wang, P.S.P. (eds.) Handbook of Pattern Recognition and Computer Vision, 3rd ed., World Scientific, Singapore (2005): 197–216.
- [17] Ojala, Timo, and Matti Pietikäinen. "Unsupervised texture segmentation using feature distributions." Pattern Recognition 32.3 (1999): 477-486.
- [18] Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with Classification based on feature distributions. Pattern Recognit. 29(1), (1996): 51–59.

Feature Extraction and Classification Methods for a Motor Task Brain Computer Interface: A Comparative Evaluation for Two Databases

Oana Diana Eva

Faculty of Electronics, Telecommunications and
Information Technology
“Gheorghe Asachi” Technical University
Institute of Computer Science
Romanian Academy
Iasi, Romania

Anca Mihaela Lazar

Faculty of Medical Bioengineering
“Grigore T. Popa” University of Medicine and Pharmacy
Iasi, Romania

Abstract—A comparative evaluation is performed on two databases using three feature extraction techniques and five classification methods for a motor imagery paradigm based on Mu rhythm. In order to extract the features from electroencephalographic signals, three methods are proposed: independent component analysis, Itakura distance and phase synchronization. The last one consists of: phase locking value, phase lag index and weighted phase lag index. The classification of the extracted features is performed using linear discriminant analysis, quadratic discriminant analysis, Mahalanobis distance based on classifier, the k-nearest neighbors and support vector machine. The aim of this comparison is to evaluate which feature extraction method and which classifier is more appropriate in a motor brain computer interface paradigm. The results suggest that the effectiveness of the feature extraction method depends on the classification method used.

Keywords—Brain computer interface; independent component analysis; Itakura distance; phase synchronization; classifiers

I. INTRODUCTION

Brain Computer Interface (BCI) provides a new communication method for people who are suffering of motor disabilities [1]. A BCI system acquires brain signals, analyzes them and translates them into commands for external devices (wheelchair, neuroprosthesis, etc.). The most commonly studied signals generated from brain activity are electrical signals. The electroencephalography (EEG) records the electrical activity by using electrodes placed on the scalp.

Motor imagery produces reliable and distinct features in the brain activity that can be used by BCI systems. When a user performs a mental activity as left/right hand movement imagination without physically executing the movements, changes called event related desynchronizations (ERD) and event related synchronizations (ERS) appear in the sensorimotor area in the corresponding signal power of Mu or beta rhythms. Mu rhythm represents an oscillation of the EEG signal in the frequency band 8-12 Hz and it is affected by movements and movement imagery [2]. There are different features extraction methods for EEG signals suited to discriminate the motor tasks in a BCI paradigm. Among these, the independent component analysis [3], [4], Itakura distances

[5]-[7] and phase synchronization methods [8]-[10] are chosen in order to be used for classification with linear discriminant analysis [11], quadratic discriminant analysis [12], Mahalanobis distance [13], the k-nearest neighbors [14], [15] and support vector machine [16], [17].

In Section II there are described the databases used in the comparative study. Section III is reserved to the methods used in the proposed assessing. The results obtained for the used databases are presented in Section IV and Section V contains the conclusions of the paper.

II. DATABASES

In the evaluation of efficiency of feature extraction and classification methods, two databases are used: the database composed of EEG signals recorded in our laboratory and the BCI competition 2002 database downloaded from the internet [18]. The databases description is listed in Table 1.

TABLE I. DATABASES DESCRIPTION

Database details	Our database	BCI Competition 2002
Number of subjects	40	9
Aquisition system	gMobilab+ module and BCI 2000 platform	Unknown
Paradigm description	Left and right arrows are displayed successively on a monitor. The subjects must carefully look at the arrows and try to imagine the left or right hand movement indicated by the arrow.	
Used channels	CP ₃ , CP ₄ , P ₃ , C ₃ , Pz, C ₄ , P ₄ , Cz.	FC ₁ , FC ₂ , FC ₃ , FC ₄ , C ₁ , C ₂ , C ₃ , C ₄ , CP ₁ , CP ₂ , CP ₃ , CP ₄

III. METHODS

The chosen feature extraction methods are presented for short. For detail information, the mentioned references may be studied.

Independent component analysis is used for spatial filters substitution. The proposed method consists in using the same spatial filter obtained by applying ICA method for relaxation state and for imagining motor tasks [19].

The Itakura distance for imagination of the left hand and the relaxation (rest) state is as follows [7]:

$$ID_{REST-LEFT} = \log \left(\frac{MSE_{y_{REST}, y_{LEFT}}}{MSE_{y_{REST}, y_{REST}}} \right), \quad (1)$$

where the mean square error $MSE_{y_{REST}, y_{LEFT}}$ and $MSE_{y_{REST}, y_{REST}}$ are:

$$MSE_{y_{REST}, y_{LEFT}} = (a^{LEFT})^T R_{y_{REST}}(p) a^{LEFT}, \quad (2)$$

$$MSE_{y_{REST}, y_{REST}} = (a^{REST})^T R_{y_{REST}}(p) a^{REST} \quad (3)$$

and $R_{y_{REST}}(p)$ is the autocorrelation matrix of $y_{REST}(n)$, $y_{REST}(n)$ is the output of an autoregressive (AR) model system with an input of $x_{REST}(n)$.

The autoregressive model is characterized by:

$$y(n) = -\sum_{k=1}^p a_k y(n-k) + e(n), \quad (4)$$

a_k are the parameters of the model, p , the model order and $e(n)$ the prediction error.

There are similar relations for the Itakura distance for movement imagination of the right hand and the relaxation state.

The left symmetric Itakura distance is [20]:

$$ID_{LEFT} = \frac{1}{2} (ID_{REST-LEFT} + ID_{LEFT-REST}). \quad (5)$$

The left normalized Itakura distance is defined as [21]:

$$NORM_{ID_{REST-LEFT}} = \frac{(ID_{REST-LEFT} - \min(ID_{REST-LEFT})) * 100}{\max(ID_{REST-LEFT}) - \min(ID_{REST-LEFT})}. \quad (6)$$

Phase locking value (PLV) [22], phase lag index (PLI) [23] and weighted phase lag index (wPLI) [24] are used to measure the synchronization between two signals $x(t)$ and $y(t)$.

PLV characterizes the stability of the phase difference between instantaneous phases $\varphi_x(t)$ and $\varphi_y(t)$:

$$PLV = \left| \langle e^{j\Delta\varphi(t)} \rangle \right| \quad (7)$$

$$\Delta\varphi(t) = \varphi_y(t) - \varphi_x(t).$$

The phase lag index [23] is defined by:

$$PLI = |\langle \text{sign}[\Delta\theta(t_k)] \rangle|, \quad (8)$$

sign is the signum function and $\langle . \rangle$ denotes the average over the time.

The weighted phase lag index is calculated using [11]:

$$wPLI = \frac{|I(X)|}{|I(X)|} = \frac{|I(X) \text{sign} I(X)|}{|I(X)|}, \quad (9)$$

where $I(X)$ is the imaginary component of the cross spectrum between two signals $x(t)$ and $y(t)$.

The used methods are described in detail in [24]-[26].

IV. RESULTS

In this section there are presented both comparisons between some features extraction methods and comparisons between some classification methods used for EEG signals recorded in a BCI motor task paradigm. The results are reported on two EEG databases: the 2002 BCI Competition database and our own database.

A. Database of EEG Signals Recorded in Our Laboratory

The methods used in feature extraction used for our database are: independent component analysis, Itakura distance, symmetric Itakura distance and measures for phase synchronization. For ICA three algorithms (INFOMAX, SOBI and JADE) are used. Concerning Itakura distance and symmetric Itakura distance, 6 and 10 order AR models are handled. PLV, PLI and wPLI are applied measures for phase synchronization. LDA, QDA, MD, kNN (k=1:5) and SVM are the methods we have utilized in order to classify the detected features.

In Table 2, the mean and maximum correct classification rates acquired for each of the mentioned feature extraction methods are presented. For ICA, Itakura distance and symmetric Itakura distance methods, maximum classification rates were obtained for LDA, QDA and MD. The lowest classification rates were achieved for PLI, PLV and wPLI. The mean classification rates are in the range of 59.06% (for wPLI) and 89.43% (for symmetric Itakura distance). The highest mean and maximum values of the classification rates were obtained using QDA.

TABLE II. THE MEAN AND MAXIMUM CLASSIFICATION RATES FOR ICA, ITAKURA DISTANCE, SYMMETRIC ITAKURA DISTANCE AND PHASE SYNCHRONIZATION METHODS WITH LDA, QDA AND MD CLASSIFIERS (ON OUR DATABASE)

Method		Classification rates					
		LDA		QDA		MD	
		Mean ± standard deviation [%]	Max [%]	Mean ± standard deviation [%]	Max [%]	Mean ± standard deviation [%]	Max [%]
ICA	INFOMAX	81,3 ± 12,74	97,73	83,6 ± 15,9	100	82,28 ± 15,82	100
	SOBI	78,8 ± 15,63	97,78	79,3 ± 17,66	100	79,64 ± 17,52	100
	JADE	83,90 ± 12,39	100	82,61 ± 19,52	100	83,62 ± 15,59	100
Itakura Distance	Model Order 6	82,40 ± 12,60	100	88,19 ± 9,74	100	86,62 ± 11,28	100
	Model Order 10	83,35% ± 11,94	100	88,33 ± 10,22	100	86,62 ± 9,78	98,33
Symmetric Itakura Distance	Model Order 6	81,35% ± 15,25	100	87,85 ± 12,48	100	86,75 ± 12,35	98,33
	Model Order 10	84,04 ± 12,54	100	89,43 ± 10,03	100	87,15 ± 10,23	100
Phase synchronization	PLI	64,78 ± 7,09	82,12	73,98 ± 6,64	85,28	73,08 ± 6,35	84,67
	PLV	64,62 ± 7,18	82,48	73,99 ± 6,67	85,64	73,03 ± 6,51	84,31
	wPLI	59,06 ± 3,62	66,67	64,08 ± 4,67	72,51	63,06 ± 4,44	71,78

From the analysis of data in Table 2, the outcomes are as follows:

- For ICA method, JADE algorithm performs the best classification rates for LDA and MD classifiers.
- For Itakura distance and symmetric Itakura distance methods, 10 order AR model with QDA classifier

presents the best performance.

- For PLI, PLV and wPLI, QDA classifier attends the highest classification rates.

In Table 3, the mean and maximum correct classification rates obtained for the each of the mentioned methods with kNN classifier are presented.

TABLE III. THE MEAN AND MAXIMUM CLASSIFICATION RATES FOR ICA, ITAKURA DISTANCE, SYMMETRIC ITAKURA DISTANCE AND PHASE SYNCHRONIZATION METHODS WITH KNN CLASSIFIER (ON OUR DATABASE)

Method		kNN Number of neighbors	Classification rates	
			Mean ± standard deviation [%]	Max [%]
ICA	INFOMAX	1	81,76 ± 13,77	100
		2	81,76 ± 13,76	100
		3	81,79 ± 13,75	100
		4	81,80 ± 13,74	100
		5	81,83 ± 13,74	100
	SOBI	1	82,25 ± 13,78	100
		2	82,21 ± 13,79	100
		3	82,17 ± 13,82	100
		4	82,14 ± 13,84	100
		5	82,11 ± 13,87	100
	JADE	1	84,61 ± 13,81	99,80
		2	84,61 ± 13,81	99,80
		3	84,62 ± 13,80	99,80
		4	84,62 ± 13,79	99,80
		5	84,63 ± 13,78	99,81
Itakura Distance	Model order 6	1	84,69 ± 9,92	97,50
		2	84,04 ± 9,90	97,78
		3	83,56 ± 9,49	97,08
		4	83,34 ± 9,50	97,00
		5	82,58 ± 9,62	96,94
	Model order 10	1	85,00 ± 9,91	97,50
		2	84,46 ± 10,18	97,22
		3	84,12 ± 10,33	97,08
		4	83,93 ± 10,56	97,00
		5	83,40 ± 10,68	96,94
Symmetric Itakura Distance	Model order 6	1	84,55 ± 11,54	99,17
		2	83,81 ± 11,61	99,44
		3	83,50 ± 11,52	99,17
		4	83,43 ± 11,56	99,33
		5	82,93 ± 11,50	98,61
	Model order 10	1	86,10 ± 16,96	99,17
		2	85,71 ± 17,06	99,44
		3	85,00 ± 16,89	98,75
		4	84,80 ± 16,91	98,67
		5	84,22 ± 16,85	97,78
Phase synchronization	PLI	1	92,74 ± 3,42	96,66
		2	92,83 ± 3,40	96,71
		3	92,89 ± 3,39	96,75
		4	92,97 ± 3,38	96,80
		5	92,98 ± 3,38	96,82
	PLV	1	92,73 ± 3,41	96,57
		2	92,83 ± 3,38	96,63
		3	92,89 ± 3,38	96,67
		4	92,97 ± 3,36	96,70
		5	92,99 ± 3,37	96,72
	wPLI	1	83,15 ± 6,83	92,94
		2	83,27 ± 6,87	93,06
		3	83,33 ± 6,87	93,16
		4	83,41 ± 6,91	93,28
		5	83,42 ± 6,90	93,33

From the analysis of data in Table 3, the findings are as follows:

- For ICA method, JADE algorithm performs the best classification rates.
- For Itakura distance and symmetric Itakura distance methods, 10 order AR model offers the best performance.
- For PLI, PLV and wPLI, there are not essential differences between the classification rates.

The mean and maximum correct classification rates obtained for each of the mentioned methods with SVM classifier are organized in Table 4.

TABLE IV. THE MEAN AND MAXIMUM CLASSIFICATION RATES FOR ICA, ITAKURA DISTANCE, SYMMETRIC ITAKURA DISTANCE AND PHASE SYNCHRONIZATION METHODS WITH SVM CLASSIFIER (ON OUR DATABASE)

Method		SVM	
		Classification rates	
		Mean ± standard deviation [%]	Max [%]
ICA	INFOMAX	82,29 ± 17,28	100
	SOBI	81,10 ± 18,07	100
	JADE	86,25 ± 14,56	100
Itakura Distance	Model order 6	82,39 ± 12,61	98,33
	Model order 10	83,10 ± 16,51	95,37
Symmetric Itakura Distance	Model order 6	80,88 ± 16,90	98,33
	Model order 10	83,24 ± 17,23	96,67
Phase synchronization	PLI	92,69 ± 5,48	99,27
	PLV	92,88 ± 5,24	99,64
	wPLI	82,00 ± 7,06	92,70

TABLE V. THE MEAN AND MAXIMUM CLASSIFICATION RATES FOR ICA, NORMALIZED ITAKURA DISTANCE AND PHASE SYNCHRONIZATION METHODS WITH LDA, QDA AND MD CLASSIFIERS (ON BCI COMPETITION 2002 DATABASE)

Method		Classification rates					
		LDA		QDA		MD	
		Mean ± standard deviation [%]	Max [%]	Mean ± standard deviation [%]	Max [%]	Mean ± standard deviation [%]	Max [%]
ICA	INFOMAX	81,64 ± 13,04	97,56	85,62 ± 16,99	100	81,81 ± 18,17	100
	SOBI	98,80 ± 10,91	100	94,10 ± 10,16	100	92,08 ± 11,32	100
	JADE	79,91 ± 16,88	100	86,54 ± 16,52	100	83,83 ± 14,30	96,96
Normalized Itakura Distance	Model Order 6	76,67 ± 8,38	82,82	72,86 ± 7,87	80	74,92 ± 9,02	83,33
	Model Order 10	80,99 ± 7,76	91,11	78,89 ± 8,44	86,67	79,63 ± 6,16	88,89
Phase synchronization	PLI	74,01 ± 8,18	86,42	82,24 ± 7,07	93,21	98,83 ± 1,32	100
	PLV	74,07 ± 8,20	86,42	82,92 ± 7,31	93,21	98,49 ± 1,34	100
	wPLI	76,61 ± 6,37	85,80	77,85 ± 6,14	88,27	95,88 ± 3,72	99,38

The best classification rates for kNN classifier (Table 6) are the following:

- For ICA method, SOBI algorithm.
- For normalized Itakura distance, 10 order AR model.
- For phrase synchronization methods, PLV and PLI.

From the analysis of data in Table 4, we can conclude that:

- For ICA method, JADE algorithm performs both the highest maximum classification rate and highest mean classification rate.
- For Itakura distance and symmetric Itakura distance methods, 10 order AR model offers the best performance.
- For PLV offers the best classification rates.

B. BCI Competition 2002 Database

The methods of features extraction are the same as those for our database, except the normalized Itakura distance instead of Itakura distance and symmetric Itakura distance. It was chosen to test the method based on the normalized Itakura distance because the results obtained following the Itakura distance calculation method without the normalization procedure did not offer optimal classification rates.

The same classification methods as in the case of our database were applied.

The mean and maximum classification rates obtained with LDA, QDA and MD, kNN (k=1:5), SVM classifiers are illustrated in Tables 5, 6 and 7, respectively.

Concerning the mean classification rates, from Table 5, we conclude that:

- For ICA, SOBI algorithm with LDA, QDA and MD classifiers lead to the best results.
- For normalized Itakura distance, 10 order AR model with LDA, QDA and MD classifier performed the best classification rates.
- For all the phase synchronization methods the highest classification rates were performed with MD classifier.

Looking at the results from Table 7, for SVM classifier, the best classification rates are the following:

- SOBI algorithm for ICA method.
- The AR model with order 10 for normalized Itakura distance method.
- PLV index for phase synchronization methods.

TABLE VI. THE MEAN AND MAXIMUM CLASSIFICATION RATES FOR ICA, NORMALIZED ITAKURA DISTANCE AND PHASE SYNCHRONIZATION METHODS WITH KNN CLASSIFIER (ON BCI COMPETITION 2002 DATABASE)

Method		kNN	Classification rates	
		Number of neighbors	Mean ± standard deviation [%]	Max [%]
ICA	INFOMAX	1	81,69 ± 19,01	100
		2	81,69 ± 19,03	100
		3	81,70 ± 19,03	100
		4	81,70 ± 19,06	100
		5	81,70 ± 19,08	100
	SOBI	1	87,02 ± 13,03	100
		2	87,07 ± 12,99	100
		3	87,12 ± 12,94	100
		4	87,16 ± 12,90	100
		5	87,20 ± 12,86	100
	JADE	1	79,26 ± 17,71	95,99
		2	79,31 ± 17,75	96,03
		3	79,36 ± 17,81	95,96
		4	79,42 ± 17,82	95,88
		5	79,48 ± 17,83	95,81
Normalized Itakura Distance	Model order 6	1	68,89 ± 14,17	86,67
		2	66,67 ± 15,50	82,22
		3	70,79 ± 16,89	86,67
		4	71,11 ± 16,77	88,89
		5	73,02 ± 16,40	88,89
	Model order 10	1	72,59 ± 10,36	84,44
		2	71,85 ± 11,91	82,22
		3	73,33 ± 9,16	82,22
		4	72,84 ± 9,01	80
		5	75,56 ± 9,55	86,67
Phase synchronization	PLI	1	99,06 ± 0,87	99,89
		2	99,06 ± 0,86	99,89
		3	99,05 ± 0,84	99,89
		4	99,06 ± 0,83	99,90
		5	99,05 ± 0,83	99,90
	PLV	1	99,01 ± 0,88	99,89
		2	99,03 ± 0,86	99,89
		3	99,04 ± 0,85	99,89
		4	99,06 ± 0,83	99,90
		5	99,04 ± 0,83	99,90
	wPLI	1	97,34 ± 1,33	98,46
		2	97,35 ± 1,33	98,48
		3	97,32 ± 1,30	98,51
		4	97,29 ± 1,28	98,54
		5	97,21 ± 1,27	98,56

TABLE VII. MEAN AND MAXIMUM CLASSIFICATION RATES FOR ICA, NORMALIZED ITAKURA DISTANCE AND PHASE SYNCHRONIZATION METHODS WITH SVM CLASSIFIER (ON BCI COMPETITION 2002 DATABASE)

Method		SVM	
		Classification rates	
		Mean ± standard deviation [%]	Max [%]
ICA	INFOMAX	82,27 ± 15,88	100
	SOBI	92,59 ± 12,16	100
	JADE	84,65 ± 15,70	100
Normalized Itakura Distance	Model order 6	74,29 ± 11,02	84,44
	Model order 10	75,56 ± 12,01	88,89
Phase synchronization	PLI	98,56 ± 1,06	100
	PLV	98,63 ± 1,22	100
	wPLI	96,91 ± 1,15	98,77

In order to compare our results to related works, some impediments appear. The major one is related to our dataset. As our database is not publically one, there are not reported any other results using the EEG recordings from this database. The results obtained on BCI 2002 competition dataset are consistent with other works. In [27] where a time-frequency approach is investigated are reported smaller classification rates than the classification rates obtained with methods presented. As concerning the BCI competition dataset, comparing different algorithms at present is still difficult, but as in [28] a global remark could be settled that the best choice of the classifier for a motor task BCI depends on the feature extraction method used in that system.

V. CONCLUSIONS

The research evaluated three feature extraction methods and five classification methods on two different databases. The algorithms are simply to apply and can be exploited by the motor imagery paradigms.

In order to have a proper preparation, the subjects from our database executed first the hand movements and then the hand movement imagination. For the subjects from the BCI competition 2002 database it is mentioned that they were well trained.

Overall the highest classification rates are obtained with QDA and with kNN classifier.

The best feature extraction methods are the phase synchronization, Itakura distance and ICA.

The results point out that the effectiveness of the feature extraction method depends on the classification method used and there is not a best method that outperforms all the others.

The future work implies the developing of a new database which will contain EEG signals achieved from people with disabilities and testing the proposed methods on that database.

ACKNOWLEDGMENT

This work was supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CNCS – UEFISCDI, project number PN-II-RU-TE-2014-4-0832 “Medical signal processing methods based on compressed sensing; applications and their implementation”.

REFERENCES

- [1] J. R. Wolpaw, “Brain-computer interfaces as new brain output pathways”, *The Journal of Physiology*, 2007, 579.3: 613-619, 2007.
- [2] J. R. Wolpaw, E. W. Wolpaw, “Brain-computer interfaces: principles and practice”, OUP USA, 2012.
- [3] A. Kachenoura, L. Albero, L. Senhadji, P. Comon, “ICA: a potential tool for BCI systems”, *IEEE Signal Processing Magazine*, 25(1), 57-68, 2008.
- [4] A. Hyvärinen A. “Independent component analysis: recent advances”, *Phil Trans R Soc A* 371: 20110534, <http://dx.doi.org/10.1098/rsta.2011.0534>, 2013.
- [5] F. Ebrahimi, M. Mikaili, E. Estrada, H. Nazeran, “Assessment of Itakura distance as a valuable feature for computer-aided classification of sleep stages”, In *Engineering in Medicine and Biology Society*, pp. 3300-3303, IEEE, 2007.
- [6] Benesty J., Sondhi M. M., Huang Y., *Springer Handbook of Speech Processing*, Springer Science & Business Media, 2007

- [7] E. Cardoso et al., “A contribution for the automatic sleep classification based on the Itakura-Saito spectral distance”, *Emerging Trends in Technological Innovation*, 374-381, 2010.
- [8] C.J. Stam, G. Nolte, A. Daffertshofer, “Phase lag index: assessment of functional connectivity from multi channel EEG and MEG with diminished bias from common sources”, *Hum. Brain Mapp.*, vol. 28, pp. 1178–1193, 2007.
- [9] X. Bao, J. Hu, “Phase synchronization for classification of motor imagery EEG”, *Journal Of Information &Computational Science*, 5(2), 949-955, 2008.
- [10] M. Vinck, R. Oostenveld et al., “An improved index of phase-synchronization for electrophysiological data in the presence of volume-conduction, noise and sample-size bias”, *Neuroimage*, vol. 55(4), pp. 1548–1565, 2011.
- [11] B. Blankertz, S. Lemm, M. Treder, S. Haufe, K.R. Müller, “Single-trial analysis and classification of ERP components—a tutorial”, *NeuroImage*, 56(2), 814-825, 2011.
- [12] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, B. Arnaldi, “A review of classification algorithms for EEG-based brain-computer interfaces” *J. Neural Eng.*, 4, pp. R1-R13, 2007.
- [13] A. Bashashati, M. Fatourehchi, R.K. Ward, G.E. Birch, “A survey of signal processing algorithms in brain-computer interfaces based on electrical brain signals”, *Journal of Neural engineering*, 4(2), R32, 2007.
- [14] S. Ge, R. Wang, D. Yu, “Classification of four-class motor imagery employing single-channel electroencephalography”, *PLoS one*, 9(6), e98019, 2014.
- [15] A. Ahangi, M. Karamnejad, N. Mohammadi, R. Ebrahimpour, N. Bagheri, “Multiple classifier system for EEG signal classification with application to brain-computer interfaces”, *Neural Computing and Applications*, 23(5), 1319-1327, 2013.
- [16] L. Xiang, Y. Dezhong, D. Wu, L. Chaoyi, “Combining spatial filters for the classification of single-trial EEG in a finger movement task”, *IEEE Trans. Biomed. Eng.*, 54, 821–831, 2007.
- [17] SM Zhou, JQ Gan, F. Sepulveda, “Classifying mental tasks based on features of higher-order statistics from EEG signals in brain-computer interface”, *Information Sciences*, 178(6):1629-40, 2008.
- [18] A.Osman, A. Robert. “Time-course of cortical activation during overt and imagined movements.” *Proc. Cognitive Neuroscience Annu. Meet.*, New York 1: 1842-1852, 2001.
- [19] Y. Wang, Y.T. Wang, T.P. Jung, “Translation of EEG spatial filters from resting to motor imagery using independent component analysis,” *PLoS One*, vol. 7, no. 5, p. e37665, Jan. 2012.
- [20] E. Estrada, H. Nazeran, F. Ebrahimi, M. Mikaeili, “Symmetric Itakura Distance as an EEG Signal Feature for Sleep Depth Determination”, *American Society of Mechanical Engineers*, pp.723-724, 2009.
- [21] E. Estrada, P. Nava, H. Nazeran, K. Behbehani, J. Burk, E. Lucas, “Itakura distance: a useful similarity measure between EEG and EOG signals in computer-aided classification of sleep stages”, *Engineering in Medicine and Biology Society*, pp. 1189-1192, 2005.
- [22] V. Gogunutla, Y. Wang, K.C. Veluvolu, “Phase synchrony in subject-specific reactive band of EEG for classification of motor imagery tasks”, *InEngineering in Medicine and Biology Society (EMBC)*, 2784-2787, 2013.
- [23] C.J. Stam, G. Nolte, A. Daffertshofer, “Phase lag index: assessment of functional connectivity from multi channel EEG and MEG with diminished bias from common sources”, *Hum. Brain Mapp.*, vol. 28, pp. 1178–1193, 2007.
- [24] O.D. Eva, A.M. Lazar, “Channels selection for motor imagery paradigm - An Itakura distance based method”, *E-Health and Bioengineering Conference (EHB)*, IEEE, pp. 1-4, 2015.
- [25] O. D. Eva, D. Tarniceriu, “Substitution of spatial filters from relaxation to motor imagery for EEG based brain computer interface”, *System Theory, Control and Computing*, pp. 147-150, 2015.
- [26] O. D. Eva, “Detection and Classification of Mu Rhythm using Phase Synchronization for a Brain Computer Interface” *International Journal of Advanced Computer Science and Applications (IJACSA)*, 7(12), 2016.

- [27] N.F. Ince, A.H. Tewfik, S. Arica, "Extraction subject-specific motor imagery time-frequency patterns for single trial EEG classification", *Computers in biology and medicine*, 37(4), pp.499-508, 2007.
- [28] H. Bashashati, RK Ward, GE Birch, A. Bashashati, "Comparing Different Classifiers in Sensory Motor Brain Computer Interfaces", Yao D, ed. *PLoS ONE*. 2015;10(6).

Creating and Protecting Password: A User Intention

Ari Kusyanti

Department of Information Technology
Universitas Brawijaya
Malang, Indonesia

Yustiyana April Lia Sari

Department of Information System
Universitas Brawijaya
Malang, Indonesia

Abstract—Students Academic Information System (SAIS) is an application that provides academic information for the students. The security policy applied by our university requires the students to renew their SAIS password based on the university's policy. This study aims to analyze SAIS users' behavior by using six variables adapted from Protection Motivation Theory (PMT), which are Perceived Severity, Perceived Vulnerability, Fear, Response Efficacy, Response Cost and Intentions. The data was collected from 288 SAIS users as respondents. The data analysis method used is Structural Equation Modeling (SEM) analysis. The study result shows that the factors affecting the intention of changing the passwords are perceived severity, fear, response efficacy, and response cost.

Keywords—Students Academic Information Systems (SAIS); SEM; intention; PMT

I. INTRODUCTION

Students Academic Information System (SAIS) enables students to access and process their academic information, such as students' personal information, study plan, courses including exam schedules and grades, and also financial information including registration/tuition fee. Since SAIS is containing sensitive and confidential information about students, authentication process is needed to protect student's privacy and to secure student's SAIS account. Users' authentication or verification problem occurs when password to log in to the system is considered unsafe. The users often use simple and predictable words for passwords like their own names or their birth dates. To prevent unwanted parties knowing users' passwords, the university has made a new policy regarding the password-creating process [1].

Surely our university imposes its own policy concerning password-creating process for SAIS account. All new freshmen who have just received SAIS account with default username and password are required to change their passwords due to the university policy. Soon after the short notice from the university, all sophomores, juniors, and seniors also demanded to change their passwords as well. The policy requires the password to be a combination of at least 8 characters minimum of letter and number. Furthermore, suggestions and notifications will appear when a user is going to set the password, i.e. "use the combination of letters and numbers", "password is good", "password is strong", and "8 characters minimum, with letters and numbers combination". Those notifications will appear to inform the user whether he has made a good password according to the password-creating policy.

This study is similar to a study that has examined the account of students academic information systems at Carnegie Mellon University (CMU) named Andrew account. In December 2009, all Andrew accounts users received an email to change their password for the security of personal information. The password policy applied to Andrew's account contains at least eight characters, and includes at least one uppercase, one lowercase, one digit, and one symbol. The password will also be subject to a dictionary check. If the user does not change the password according to the new policy, the user becomes unable to access their Andrew account.

Several studies have examined how password policy affects user behavior. The result is that although users are aware of security issues but users rarely change their passwords [2]. A survey reported that 90% of 152 computer system users leaked their passwords. The survey also found that users tend to use simple passwords and passwords are used from time to time [3]. A survey conducted by SafeNet found that about half of the respondents wrote down their passwords and about 80% had 3 or more passwords [4].

In determining what factors influence the user to create strong password according to the policy applied, this research is using a model of Protection Motivation Theory (PMT). This model is best suited to investigate the protection motivations of users associated with user behavior in password-creating process. According to the PMT, someone wants to do something because it has its own protection motivation. Protection Motivation Theory (PMT) model consists of two processes, namely, threat-appraisal process and coping-appraisal process. Both processes have each variable that will affect the purpose of implementing strong password-creating process. Therefore, measures of behavioral intention are the typical dependent variable in the PMT. Two meta-analyses of the PMT show that it has been useful in predicting health-related intentions [5].

There is a recent studies on password-creating process by [6] entitled "Encountering Stronger Password Requirements: User Attitudes and Behaviors" which observing the attitude and behaviors related to the use and password-creating process. However, [6] did not include theoretical model that portraying the factors that affect user to create a strong password. Therefore, this research intends to use the research model taken from a study entitled "Am I Really at Risk? Determinants of Online Users' Intentions to Use Strong Passwords" by [7], that studied about perceived severity, perceived vulnerability, fear, response efficacy, response cost which affecting intention using framework from Protection Motivation Theory (PMT). Moreover, the intention variable is adapted from the study of

[8] which is used to measure the SAIS users' intention tendencies. This research objective is to examine whether perceived severity, perceived vulnerability, fear, response efficacy, response cost influence SAIS users' intention in creating a strong password.

The outline of this research is in Section 1 explains the background of the issues raised, while Section 2 describes the research model to be used along with the formulation of the hypothesis. Afterwards, Section 3 describes the data analysis and presented in the form of data, and in Section 4 is the discussion exposure from the results of data analysis that has been done. Finally Section 5 is the exposure of the conclusions from the results of data analysis that has been obtained.

II. MODEL STRUCTURE AND HYPOTHESIS

This research is confirmatory research based on model and hypothesis by [7] and [8]. The data is analyzed using Structural Equation Modeling (SEM). There are two stages in this SEM analysis: measurement model and structural model. Measurement model is used to determine the relationship between indicator and variables while structural model shows the relationship between latent variables.

The variables that are used in this research are described as follows along with the hypotheses.

A. Definition of each variable

1) Perceived Severity (PS)

Generally, Perceived Severity is used to scrutinize individual's reaction to life-threatening objects. If individual does not aware about how dangerous the threat is, therefore there is no motivation to protect themselves and no behavioral change. The violation of passwords can cause sensitive information and personal data leakage [7].

2) Perceived Vulnerability (PV)

When a user chose a weak password, the password is generally a common word and easily predicted [7]. The users believe that only people who has classified information or people who are distraught by the hackers whom should be aware about computer's securities [9].

3) Fear (FEAR)

Fear is an emotional response to a threat which can cause a change in attitude and behavior [10]. The anxious users will be motivated to use strong passwords. Those users tend to do anything to secure their account and change their passwords regularly.

4) Response Efficacy (RE)

Stronger passwords can protect online accounts better. Apart from using strong passwords, regularly changing password can help securing online accounts from malicious hackers [7].

5) Response Cost (RC)

According to [7] Response Cost refers user's time and work spent on creating and recreating passwords. Most users often forgot their passwords and having a hard time to remember their passwords. Using strong passwords and changing it from time to time can cause discomfort for users.

That is why most users use one password for many accounts instead of creating different password for each account.

6) Intention (SI)

According to [8] intention is used to measure how strong users' intention is in protecting their online accounts.

B. Hypothesis for the Variables

Perceived by the severity assess how severe is a threat that affects individual's life. The more serious is the individual to feel the negative impact of a threat, the more he will perform the recommended actions. If individual does not consider the impact of a severe threat to his life then there is no motivation protection measures undertaken. Using the Protection Motivation Theory (PMT), researchers showed perceived severity had a significant relationship with the behavior of the protection of such implementing measures as in [11] and [12]. Thus, the hypothesis is:

H1: Perceived severity is positively related to the intention of implementing online password protection.

In protecting an online account, password regarded as a vulnerability to threats. First, the hacker can employ a variety of techniques to attack the user's password. For example, hackers can use keyword-based attacks - a dictionary word, the technique of using the program to guess passwords by finding possible combinations include common words, slang and popular phrases. Since computer users tend to choose to use a bad password, word-based attacks would be very efficient [9]. Passwords can also be unpredictable after studying an individual's personal information such as birthdays, spouse or spouse's name, pet's name. Peoples who have a high degree of vulnerability felt to be more concerned with security or protection of their password [13]. Hence, the hypothesis is:

H2: Perceived vulnerability is positively related to the intention of implementing online password protection.

Fear refers to fears triggered by a threat. Fear is an emotional response to a threat that can cause a change individual's behavioral intentions [14]. If users are afraid of the threat of attack to guess passwords or hacked by others, they will be more likely to spend more effort in maintaining and updating their passwords. There is a positive relationship between fear with compliance with recommended action [15]. Fears increase user intention to use strong passwords. If users are afraid password will be hacked by someone else, they will be more likely to spend a lot of effort to renew their passwords. Therefore, the hypothesis is:

H3: Fear is positively related to the intention of implementing online password protection.

Response efficacy evaluates how effective coping responses suggested in reducing the threat. In implementing behavioral protection, the individual must make sure that the protective behaviors that they do will be effective in protecting them against the threat. In addition, using strong passwords to protect online accounts, renew regular password also help protect online accounts from malicious hackers. Individuals will be more involved in the protection behavior if they believe that their extra effort to create a secure password is valuable

[16]. It is also stated that response efficacy is positively related to protection behavior. Therefore, we hypothesized:

H4: Response efficacy is positively related to the intention of implementing online password protection.

Response cost measure effort including time, money, etc., that individual must pay when doing behavioral protection. As a result, response cost reduces the possibility of selecting the recommended action. In information security, [17] found that the barriers of implementing security practices negatively related to the attitude of individual [17]. Creating and updating passwords regularly adding user inconvenience. In addition, various online accounts owned by the user cause higher costs response. This is the reason users reuse their passwords for the same account to minimize the cost of the response in using a strong password. Thus, the hypothesis is:

H5: Response cost is negatively related to the intention of implementing online password protection.

Based on the above hypotheses, the research model is developed as shown in Fig. 1.

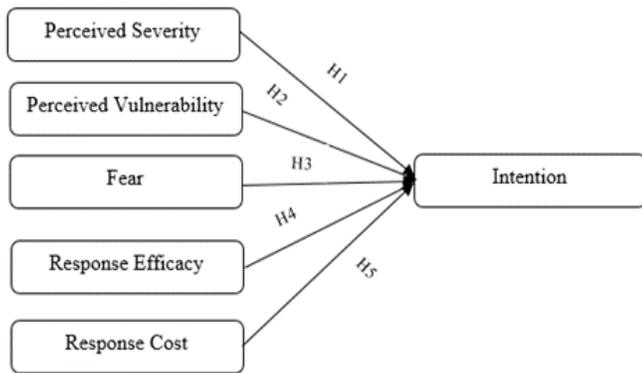


Fig. 1. Research model.

The model in Fig.1 will be used to depict the relationship between latent variables. This research is analyzing six latent variables and sixteen manifested variables (indicators).

III. DATA ANALYSIS

Statistical analysis that is used for this research is SEM. SEM is used to analyze the collected data from questionnaire. The complete questionnaire can be seen in Appendix (Table 6). The respondents of this study are all students whom actively use Students Academic Information Systems (SAIS).

A. Descriptive Analysis

A total of 300 questionnaires obtained from students who are actively using Students Academic Information Systems (SIAS). The characteristic of respondents is shown in Table 1.

B. Missing Data and Outlier

Based on Little’s MCAR, there is no missing data in this study. Mahalanobis distance is used to determine outlier data. Data which has mahalanobis distance of more than 34,805 is considered the outlier and need to be withdrawn. From 300 questionnaires collected, there are 12 outlier data, so the eligible data to be analyzed are 288 data.

TABLE. I. CHARACTERISTIC OF RESPONDENTS

Age	Count	%	Gender	Count	%
17	2	0.67	Male	0	0
			Female	2	0.67
18	1	0.33	Male	0	0
			Female	1	0.33
19	14	4.67	Male	7	2.33
			Female	7	2.33
20	155	51.67	Male	83	27.67
			Female	72	24
21	115	38.33	Male	52	17.33
			Female	63	21
22	13	4.33	Male	6	2
			Female	7	2.33
Count	300	100		300	100

C. Reliability Test

Reliability test is performed based on Cronbach alpha for every latent variable. Details of latent variable with its corresponding Cronbach alpha can be seen in Table 2.

TABLE. II. CRONBACH ALPHA VALUE

Factor	Cronbach Alpha
Limit Value	>0,6
PS	0.861
PV	0.853
FEAR	0.905
RE	0.660
RC	0.729
INTENTION	0.758

D. Factor Analysis

According to [18] the test of Kaiser-Meyer-Olkin (KMO) and Bartlett’s test is used to determine whether the sample data used in the study is sufficient to analyze certain factors. The KMO result is 0.704, so it can be said to have a good criteria. Then for Bartlett’s test is 0.000, so it can be said to be highly significant in accordance with the criteria of [18] (Sig. <.001).

E. Normality Test

Normality test aims to evaluate whether the regression model, the variable spam or residuals have a normal distribution. If this assumption is violated, the statistical tests to be invalid for a number of small samples [19]. The descriptive statistics for the latent factors or constructs revealed that the values for the Skewness and Kurtosis were lower than ± 2 for both statistics, which confirmed that there was no major issue of non-normality of the data [20]. Based on the tests that have been done, 288 data are normally distributed.

F. Levene’s Test

Levene’s test is used to determine whether the research data obtained is homogeneous or not [21], so it can be used for subsequent statistical analysis. Data are considered homogeneous if Sig. > 0.05 contrary, if Sig. < 0.05 then the data is considered not homogeneous. All variables in this study are said to meet homogeneous criteria.

G. Overall Model Fit

First step in SEM, which is a measurement model, is performed to determine the relationship between latent variables and its indicators by evaluating overall model fit. Overall model fit test results can be seen in Table 3. Based on

Table 3, the study has met all the determined limits. It can be concluded that the research method is fit and can be proceed for structural.

TABLE. III. GOODNESS OF FIT INDICES (GOFI) VALUES

Indeks	Criteria	Value	Info
Chi-square	>0,05	250,362	Good
CMIN/DF	1.00 < CMIN/DF < 3.00	2,813	Good
GFI	>0.9	0.913	Good
RMSEA	<0.05 good fit <0.08 acceptable fit	0.078	Acceptable Fit

Convergent validity is the extent to which observed variables of a particular construct share a high portion of the variance in common [22]. In addition, [18] suggested that average variance extracted (AVE) estimation should be greater than 0.5. AVE results can be seen in Table 4, in which all variables are said to meet the criteria.

TABLE. IV. AVE RESULTS

Construct	AVE
PS	0.608
PV	0.728
FEAR	0.563
RC	0.443
RE	0.603
INTENTION	0.631

H. Structural Model Fit

The next step in SEM is structural model fit. Path analysis is used to perform the advanced test which is structural model fit. This test is used to determine the relationship between latent variable to the model. The results of structural model fit can be seen in Table 5.

TABLE. V. STRUCTURAL MODEL RESULTS AND SEM MODEL HYPOTHESIS

Hypothesis	P	Result
	<0.05	
INTENTION ← PS	.004	Accepted
INTENTION ← PV	.055	Rejected
INTENTION ← FEAR	***	Accepted
INTENTION ← RE	***	Accepted
INTENTION ← RC	***	Accepted

The indicators of structural model fit test are the value of estimate, critical ratio, and p-value which can be seen completely in Table 5. In pursuant to Table 5, the relationship between variables with p-value less than 0.05(*) has a strong relation and the hypothesis is accepted.

IV. RESEARCH RESULT

A. Discussion on Hypothesis 1

Hypothesis 1 is accepted. It can be concluded that the respondents considered the threat of severe violations password for his life, so that users tend to change their behavior by using a strong password. It shows that in this study that Perceived

Severity (PS) has significant influence over users' intention in creating password (INTENTION). Therefore, in this study Hypothesis 1 is received.

B. Discussion on Hypothesis 2

As Hypothesis 2 is rejected, it shows the respondents do not concern about their password-creating process to protect their accounts. They also do not aware about possible danger from hackers. The result shows that there is no change in user behavior in creating a strong password. Therefore, the Perceived Vulnerability (PV) does not affect significantly users' intention in creating password (INTENTION).

C. Discussion on Hypothesis 3

Hypothesis 3 is accepted which means that the respondents are alarmed about the harm and the threats from the use of weak passwords. This can increase users' intention to create stronger password in order to secure their accounts. It proves that fear (FEAR) significantly affect users' intention in creating password (INTENTION).

D. Discussion on Hypothesis 4

Hypothesis 4 is accepted which shows that the respondents are aware that the use of strong passwords can secure their accounts from hackers. This can increase their intentions to create stronger passwords. It proves that Response Efficacy (RE) can significantly affect users' intention in creating password (INTENTION).

E. Discussion on Hypothesis 5

Hypothesis 5 is accepted, it can be concluded that the respondents consider that frequently updated password is not just waste of time and requires no effort, they believe it can improve their security so that it can affect respondents' intentions in creating stronger passwords. It shows that in this study the Response Cost (RC) has a significant influence on users' intention in creating password (INTENTION).

V. CONCLUSION

Based on the data analysis it can be concluded that factors affecting users to create strong passwords are; perceived severity, fear, response efficacy, and response cost. Respondents considered the threat of severe password violations so that users tend to change their behavior by creating a strong password. The respondents had fear to the threats that could be caused by the easily predicted passwords; hence the strong passwords are created. Respondents are sure that creating powerful and strong passwords would protect their account from the hackers; that increasing respondents' intention in creating strong ones. The respondents consider that frequently updated password is not just waste of time and requires no effort, they believe it can improve their security so that it can affect respondents' intentions in creating stronger passwords.

Although this research is only focused on Students Academic Information System (SAIS), many other applications, such as social network account, e-commerce account, email account, etc., also require a strong password to protect it. In that sense, this research only represents a first step in the direction of evaluating users' intention in protecting their

account. In ongoing and future work, it can be extended to a broader scope and platforms.

In addition, the result of this research can raise users' security awareness in term of protecting their online accounts as the security awareness is an important necessity for any organization, including university. Users are needed to be informed regarding their online safety to prevent a lot of potential problems that could damage the infrastructure and the organization as a whole.

APPENDIX

TABLE. VI. COMPLETE QUESTIONNAIRE

Item	Construct Indicator (measured on five-point, Likert-type scale)	References
Perceived Severity	<ol style="list-style-type: none"> How severe do you think the consequence will be if someone guessed your passwords? How severe do you think the consequence will be if someone cracked your passwords? How severe do you think the consequence will be if someone obtained your passwords? 	Adapted from [23] cited in [7]
Perceived vulnerability	<ol style="list-style-type: none"> What are your chances of someone guessing your passwords? What are your chances of someone cracking your passwords? What are your chances of someone obtaining your passwords? 	Adapted from [24] cited in [7]
Fear	<ol style="list-style-type: none"> The thought of having someone guess my passwords makes me nervous The thought of having someone crack my passwords makes me nervous The thought of having someone obtain my passwords makes me nervous 	Adapted from [25] cited in [7]
Response Cost	<ol style="list-style-type: none"> If I use strong passwords, they will be difficult for me to remember. If I update my passwords often, they will be difficult for me to remember If I use unique password on each account, they will be difficult for me to remember 	Adapted from [11] cited in [7]
Response Efficacy	<ol style="list-style-type: none"> I can protect my online accounts better if I use strong passwords I can protect my online accounts better if I update my passwords often I can protect my online accounts better if I use unique passwords for each 	Adapted from [26] cited in [7]

	online accounts	
Intention	<ol style="list-style-type: none"> I intend to make a strong password I intend to use strong password in the future I intend to update the password as often as possible 	[8]

REFERENCES

- Tim UPPTI (2007) Buku Panduan Layanan Teknologi Informasi Untuk Mahasiswa. Malang.
- P. Inglesant and M. A. Sasse. The true cost of unusable password policies: password use in the wild. In ACM Conference on Human Factors in Computing Systems 2010, pages 383{392, 2010.
- J. Leyden. Office workers give away passwords for a cheap pen. The Register, 2003.
- SafeNet. 2004 annual password survey results. SafeNet, 2005.
- Floyd, D. L., S. Prentice-Dunn, and R. W. Rogers. 2000. A meta analysis of research on protection motivation theory. Journal of Applied Social Psychology 30: 407-429.
- Shay, R., Komanduri, S., Kelley, P.G., Leon, P.G., Mazurek, L.M., Bauer, L., Christin, N., Cranor, L.F., Encountering Stronger Password Requirements : User Attitudes and Behaviors, 2010.
- Zhang, Lixuan and McDowell, William C.(2009) 'Am I Really at Risk? Determinants of Online Users' Intentions to Use Strong Passwords', Journal of Internet Commerce, 8: 3, 180 — 197
- Shin, D.H., (2010) The effects of trust, security and privacy in social networking: A security-based approach to understand the pattern of adoption. Vol. 22 No. 5, pp 428-438
- Campbell, J., D. Kleeman, and W. Ma. 2007. The good and not so good of enforcing password composition rules. Information Systems Security 16 (1): 2-8.
- Weirich, D., and M. A. Sasse. 2001. Pretty good persuasion: A first step towards effective password security in the real world. Proceedings of the 2001 Workshop on New Security Paradigms, Cloudscrofl, NM, September 10-13.
- Woon, I. M. Y., G. W. Tan, and R. T. Low. 2005. A protection motivation theory approach to home wireless security. Proceedings of the 26th International Conference on Information Systems, Las Vegas, NV, December 11-14, 367-380.
- Lee, Y., and K. R. Larsen. 2009. Threat or coping appraisal: Determinants of SMB executives' decision to adopt anti-malware software. European Journal of Information Systems 18: 177-187.
- Weirich, D., and M. A. Sasse. 2001. Pretty good persuasion: A first step towards effective password security in the real world. Proceedings of the 2001 Workshop on New Security Paradigms, Cloudscrofl, NM, September 10-13.
- LaTour, M. S., and H. J. Rotfeld. 1997. There are threats and (maybe) fear-caused arousal: Theory and confusions of appeals to fear and fear arousal itself. Journal of Advertising 26:45-59.
- Sutton, S. R. 1982. Fear-arousing communications: a critical examination of theory and research. In Social psychology and behavioral medicine, ed. J. R. Eiser, 303-337. London: Wiley.
- Gurung, A., X. Luo, and Q. Liao. 2009. Consumer motivation in taking action against spyware: An empirical investigation. Information Management and Computer Security 17 (3): 276-289.
- Herath, T., and H. R. Rao. 2009. Protection motivation and deterrence: A framework for security policy compliance in organizations. European Journal of Information Systems 18:106-125.
- Field, A. 2009. Discovering statistics using spss 3rd ed. [e-book]. Sage Publications. DOI= http://fac.ksu.edu.sa/sites/default/files/ktb_lktrwny_shml_fy_lhs.pdf.
- Ghozali, Imam. 2005. Aplikasi Analisis Multivariate dengan program SPSS, Badan Penerbit Universitas Diponegoro, Semarang.

- [20] Chandio, F. H. 2011. Studying Acceptance of Online Banking Information System: A Structural Equation Model. London: Brunel University.
- [21] Levene. 1960. Contributions to Probability and Statistics. Stanford University Press. CA.
- [22] Hair, et al. 2006. Multivariate Data Analysis 6th Ed. New Jersey: Pearson Education
- [23] Plotnikoff, R. C., and N. Higginbotham. 2002. Protection motivation theory and exercise behavior change for the prevention of coronary heart disease in a high-risk, Australian representative community sample of adults. *Psychology, Health and Medicine* 7 (1): 87–98.
- [24] Pechmann, C., C. Zhao, M. E. Goldberg, and E. T. Reibling. 2003. What to convey in antismoking advertisements for adolescents: The use of protection motivation theory to identify effective message theme. *Journal of Marketing* 67:1–18.
- [25] Milne, S., S. Orbell, and P. Sheeran. 2002. Combining motivational and volitional interventions to promote exercise participation: Protection motivation theory and implementation intentions. *British Journal of Health Psychology* 7:163–184.
- [26] Maddux, J. E., and R. W. Rogers. 1983. Protection motivation and self-efficacy: A revised theory of fear appeals and attitude change. *Journal of Experimental Social Psychology* 19 (5): 469–479.

Analyzing the Social Awareness in Autistic Children Trained through Multimedia Intervention Tool using Data Mining

Richa Mishra

Sir Padampat Singhania University
Udaipur, Rajasthan, India

Divya Bhatnagar

Sir Padampat Singhania University
Udaipur, Rajasthan, India

Abstract—This study focuses on creating a guideline for the ASD children by simulating the situation and analyzing the understanding of ASD (Asperger Syndrome) children over social skills by using a multimedia intervention tool designed for this purpose. 84 ASD individuals belonging to NGOs and clinics were selected for studying their social and cultural awareness. Autistic kids were taught social skills using specially designed multimedia intervention tool, in a controlled environment under the supervision of special educators or parents. Data mining technique was used to extract knowledge from the data collected after intervention. The results were analyzed to understand the impact of the designed multimedia intervention tool and share with special educators and parents of autistic children. The proposed multimedia intervention tool is inexpensive and user friendly. Integration of this tool has been observed to improve the quality of training an individual with autism traits. The overall growth in social communication of the ASD children under observation was observed to be 26.19%. There were substantial variances between age groups, training set and behavior parameters on any of the measures at follow-up. It was considered that an intervention starts at early age and proves beneficiary to ASD children. The study is establishing the remarkable benefits of designed multimedia intervention tool to train the ASD children.

Keywords—Asperger Syndrome (ASD); multimedia intervention tool; social skills; autism; computer aided training; autistic children

I. INTRODUCTION

Autism spectrum disorder is one of the most thought-provoking application of technology in the study, diagnosis, and treatment of disease [1]. As is the case with other psychological illnesses, the condition of autism is particularly incorporeal and complicated, providing no obvious, direct way of utilizing technology or conducting technological study to improve the condition of a patient with autism [2].

Autism is illustrated by insufficiencies in social communication and interaction, and abnormal and recurring behavior. Reasoning abilities in people with autism diverge between those with standard to above standard intelligence, to marginal and minor mental retardation, and others those function within the reasonable to extremely mentally retarded range [3].

Furthermore, the social aspect of the disease does not impart itself to treatment using any physical apparatus or

trivial scientific methods. Still, significant effort has gone into the exploration of technological aid in diagnosis and treatment of the disease, of improving the everyday life of an autistic person [4].

Computer technology plays an important role in the treatment of ailments, illness and disabilities. We continually seek and explore new ways of improving the health to empower those with disabilities to lead a life close to the normal. Multimedia Game or computer aided game helps in increasing Empathy Spectrum Quotient, Systemizing spectrum Quotient [5], [6].

Importance is also placed on present changes, with previous methods serving more or less successful strategies for managing with autism through multimedia technology [7]. The Study considered the detail analysis of autistic children on by using data mining as decision making tool [8]-[10] and multimedia intervention tool [11]-[13].

This study focuses on measuring social skills and analyzing the impact of multimedia intervention tool on the social behavior of ASD children.

II. RELATED WORK

1) A number of interventions have provided home-deliverable software or DVD packages that use photos, multimedia tool and document to teach face recognition and the recognition of emotion from faces. The majorities have found that subjects improve within the trained environment, but that these improvements do not generalize to performance of the same task in a ‘real-world’ environment. The purposes for this are discussed. One recent study [14] provided DVD packages to younger children (aged 4 to 7) and reported surprisingly strong generalization in a number of areas.

2) Other work here involves cameras or webcams than can be used to recognize emotion based on a combination of visual (facial) cues and top-down predictive emotional models. The wearable camera can either be pointed outwards, allowing a user to receive the computerized predictions about the emotional states of the person he or she is talking to, or pointed inwards [15] thus allowing the wearer to receive automatized predictions of what they are communicating about their own mental states.

3) Other projects use different media to explore interactive plays in ASD. The Reactive Colours project [16] is a digital play environment in which children with autism stand in front of a screen and explore different sorts of touch interface. Touching the screen evokes different combinations of auditory and visual cues – in one activity, dragging your finger across the screen creates a trail of bubbles and an ethereal sound; in another, hitting the screen results in cymbal crashes and huge splurges of colour.

The appeal of such a playful environment is intuitively understandable. It is a highly predictable, beautifully realized environment that rewards the kinds of repetitive, perfectly contingent interactions that some people with autism have been reported to prefer [17]. Although there is no direct training component, such endeavours are important as assistive technologies [18].

III. METHODOLOGY

In an attempt to collect the data about ASD individuals, few clinics and NGO's of Gujarat and Mumbai were visited. Special educators and parents were contacted to collect information on ASD children, and provide training on the multimedia intervention tool designed for this purpose. Medical history, responses of ASD children to prompts, and the observations before and after intervention were recorded.

The ASD children were pre conditioned to use mobile application in a controlled environment in the presence of parents and therapists, all were instructed to use commercially available apps to make ASD children familiar and acquainted with touch screen of tablets and mobile screens before starting the interventions.

A. Data Collection

Data, that are collected between January 2015 to March 2017; from a NGOs and clinics for the total sample of age groups by child's gender, to train or to improve quality life of autistic children. However study focuses on particularly assigned interventions. Data Mining is used to examine a range of parameters while sessions like competence, positive and negative likelihood ratios were reported.

B. Participants

The ASD groups comprised four groups aged from 3-30 for the experimental setup. The control group comprised NV (novice), BG (beginner), AD (adolescent) and AL (adulthood). After screening for ASD and other disability like PDD-NOS, PDD, ADHD conditions, 7 participants were excluded as they belonged to other disability. Fig. 1 shows relation of age groups and sex in regard of different behaviours of ASD children. The remaining 54 males and 23 females were matched on high guidance & teaching. In Table 1, the detail of participant's for discrete trials is shown. Data Mining is used to see the performance with respect to age group and gender, Study revealed significant differences between the age groups.

C. Dataset

a) Data collected for study

TABLE I. THE DETAILS OF PARTICIPANTS' FOR DISCRETE TRIALS

Baseline Characteristics	Clinic based treatment model			Special Educator managed treatment model		
	Total	Male	Female	Total	Male	Female
Age group & sex						
(0-3) NV	1	1	0	9	7	2
(4-11) BG	26	19	7	10	5	5
(12-19) AD	19	13	6	4	2	2
(20+) AL	8	7	1	0	0	0

Certain apprehension has been communicated, though, that the extrinsic provocation provided by the computer game might undermine any intrinsic interest a student might have in the subject matter.

By taking into due consideration of medical history and other recorded parameters the duration of sessions are decided for each individual ASD children. In Table 2, the details of participant's Dataset collected on mentioned parameters by using multimedia intervention tool are shown.

b) Attribute identification

TABLE II. ATTRIBUTES TO CALCULATE THE GROWTH IN ASD CHILD

Attributes	Description
AQ	Autism spectrum Quotient
EQ	Empathy Spectrum Quotient
SQ	Systemizing spectrum Quotient
FQ	Friendship Questionnaire
RQ	Relatives Questionnaire

D. Intervention

The multimedia intervention tool and exploration on computer-aided training (CAT) for the ASD child assesses the degree to which individualization has been addressed. This type of CAT includes short video clips or educational games which gives the student an ambition to reach or a problem to solve and requires the student to learn concepts through trial and error. For example, the ongoing simulation "Society" tells about Indian society followed by a game play consisting of multiple-choice questions (MCQs), to know the understanding of ASD kids based on video clips. Multimedia intervention tool based CAT is to provide individualized instruction, then it benefit from what is known about one-to-one teaching, an customized form of training which is the most effective form of instruction.

Assessments were done flexibly as per mood and nature of each child for study of FSCQ (Final score of social communication questionnaire), which includes scores of AQ,

EQ, FQ, RQ and SQ of ASD children. In Table 3, detailed explanation of the terms for calculating the growth of ASD children is shown.

TABLE III. TERMS USED TO EVALUATE THE PERFORMANCE OF ASD CHILDREN

S.no.	Terms	Explanation
1	FSCQ	Final score of social communication questionnaire FSCQ=(EQ'-EQ) + (SQ'-SQ) + (FQ'-FQ) + (RQ'-RQ) (If score greater than 15 then 2 otherwise 1) then converted into H and L
2	FS1	Final score of level 1 (ASD children plays MCQ by using designed multimedia intervention tool and score accordingly)
3	FS2	Final score of level 2 (ASD children has to type the answer and submit to score so at this level writing skills are also playing a considerate part)
4	EQ EQ'	Pre score of Empathy Spectrum Quotient Post score of Empathy Spectrum Quotient
5	SQ SQ'	Pre score of Systemizing spectrum Quotient Post score of Systemizing spectrum Quotient
6	FQ FQ'	Pre score of Friendship Questionnaire Post score of Friendship Questionnaire
7	RQ RQ'	Pre score of Relatives Questionnaire Post score of Relatives Questionnaire

$$\begin{aligned} \text{Rise in EQ} &= (EQ' - EQ) & (1) \\ \text{Rise in SQ} &= (SQ' - SQ) & (2) \\ \text{Rise in FQ} &= (FQ' - FQ) & (3) \\ \text{Rise in RQ} &= (RQ' - RQ) & (4) \end{aligned}$$

$$\begin{aligned} \text{FSCQ} &= (EQ' - EQ) + (SQ' - SQ) + (FQ' - FQ) + \\ & (RQ' - RQ) & (5) \\ \text{FSCQ} &\text{ is termed as Growth for the experimental setup.} \end{aligned}$$

Pre and post scores of discrete trials conducted through the designed multimedia intervention tool were taken into consideration.

E. Classification

Classification is a two-step process in which classification algorithms were applied on training and test data to classify the performance of ASD children. Various evaluation parameters were studied to find the accuracy of the classifiers.

The algorithms used in the study for classification are J48, SVM, JRip, Voted perceptron and Multilayer perceptron.

F. Experimental Setup

Weka is open source data mining tool and it is broadly used in data mining applications. From the above dataset of ASD children result.csv file was created. The file was then loaded in Weka explorer. Sample of 84 ASD children were taken for the implementation. The classify console permits to use classification algorithms to the selected dataset, to estimate accuracy and allow to visualize the model. The decision tree, SVM, Voted and multilayer perceptron were

applied in Weka. Under the test options, the 10-fold cross validation is chosen to classify.

IV. RESULTS

The result obtained for Final score of social communication questionnaire, i.e., FSCQ was obtained as under:

The improvement of 3.17% was observed in Empathy Spectrum Quotient EQ.

The improvement of 53 % was observed in Systemizing Spectrum Quotient SQ.

The improvement of 19% was observed in Friendship Questionnaire FQ.

The improvement of 28% was observed in Relatives Questionnaire RQ.

The growth in social communication skills of the observed ASD individuals was measured as 26.19%.

Social communication and cognitive skills tasks has shown significant improvement by depicting from the scores of FS1, FS2 and FSCQ. There is enough evidence to show that multimedia intervention tool for autistic children are independent of gender but impacts of age groups exist.

In Fig. 2, the knowledge represented by decision tree can be extracted in the form of IF-THEN rules to predict the growth in FSCQ by J48 (Final score of social communication questionnaire).

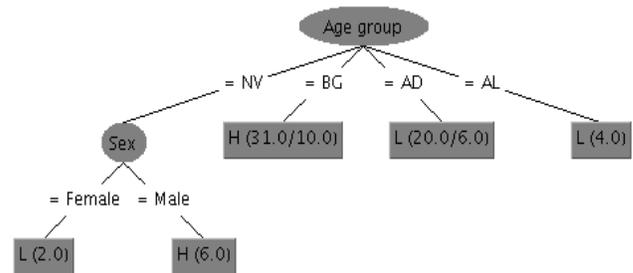


Fig. 1. Decision tree (Class attribute: Child Growth, L: Low and H: High).

The classification rules are as under:

1. If Age group = NV AND Sex=Female THEN Growth=L
2. If Age group = NV AND Sex= Male THEN Growth= H
3. If Age group = BG THEN Growth= H
4. If Age group = AD THEN Growth= L
5. If Age group = AL THEN Growth= L

From the above set of rules a conclusion emerges the ASD children of age group BG and female of NV group learns by using multimedia intervention tool and shows remarkable growth in performance of FSCQ. Table 4 presents the results

of evaluation parameters for various Classifiers. Hence voted perceptron stated as more accurate than other classifiers.

TABLE IV. EVALUATION PARAMETERS OF VARIOUS CLASSIFIERS

Evaluation Criteria	J48	Multilayer Perceptron	SVM	Voted
Accuracy	58.73%	79.37%	82.54%	88.89%
Kappa statistic	0.1727	0.5882	0.6516	0.7769
Mean absolute error	0.4597	0.195	0.1746	0.1077
Root mean squared error (RMSE)	0.5457	0.4013	0.4179	0.3168

Fig. 2 shows that the graphical representation of accuracy results of ASD children performance based on dataset. It

clearly reveals that Voted is the best classifier for analyzing the ASD children performance.

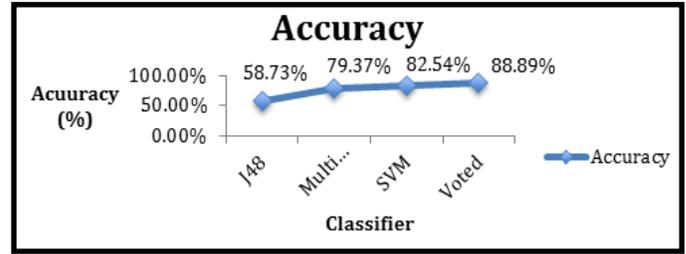


Fig. 2. Accuracy of the classifiers.

Table 5 shows the accuracy by Class. The Voted algorithm outperforms all other classifiers.

TABLE V. ACCURACY BY CLASS FOR VARIOUS CLASSIFIERS

Classifier	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
J48	0.833	0.000	1.000	0.833	0.909	0.844	Low
	1.000	0.167	0.868	1.000	0.930	0.844	High
	0.921	0.087	0.931	0.921	0.920	0.844	Average
SVM	0.867	0.212	0.788	0.867	0.825	0.827	Low
	0.788	0.133	0.867	0.788	0.825	0.827	High
	0.825	0.171	0.829	0.825	0.825	0.827	Average
Voted	0.867	0.091	0.897	0.867	0.881	0.927	Low
	0.909	0.133	0.882	0.909	0.896	0.927	High
	0.889	0.113	0.889	0.889	0.889	0.927	Average
Multilayer Perceptron	0.833	0.242	0.758	0.833	0.794	0.884	Low
	0.758	0.167	0.833	0.758	0.794	0.884	High
	0.794	0.203	0.797	0.794	0.794	0.884	Average

Fig. 3 presents ROC obtained for different classifiers.

Voted has highest area curve compared to different classifiers. Therefore, the Voted is the efficient classification technique.

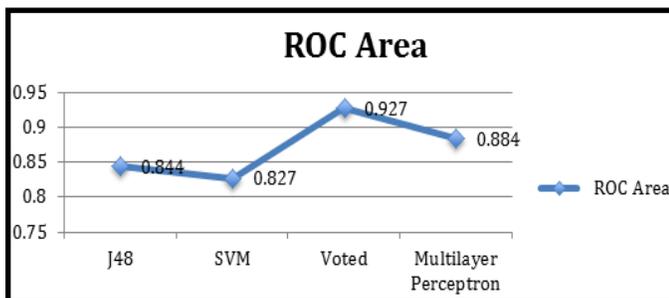


Fig. 3. ROC area of classifiers.

V. CONCLUSION

The study evaluated the expediency of machine learning algorithms in classifying the performance of ASD children. It was also discovered that Voted Perceptron performs best when employed in the study. This study is quite helpful for the

analysis of the behaviour of ASD children and providing timely suggestions to special educators and parents. The multimedia intervention tool designed for this study is inexpensive and affordable. Integration of the tools will enhance the quality of Personal attention to an individual with autism. The improvement of 3.17% was observed in Empathy Spectrum Quotient, the improvement of 53% was observed in Systemizing spectrum Quotient, the improvement of 19% was observed in Friendship Questionnaire and the improvement of 28% was observed in Relatives Questionnaire thus total growth in terms of Final score of social communication questionnaire was observed as 26.19% in ASD children.

Multimedia intervention tool intrusion is found to be greatly successful in treating disorders like anxiety and autism as parents of NGOs and clinical associates in India found multimedia a good tool for increasing concentration and eye contact and observed the rise in mood constancy in ASD children. The study is establishing the remarkable benefits of designed multimedia intervention tool to train the ASD children.

Currently intervention tool is on android platform further can be developed on iOS platform.

In future we plan to focus on developing a mechanism to read brain waves of autistic children so that we can have a comparative study between collected data through mobile intervention tool and measured brain waves of individual child. Certainly we will do utmost to achieve the goal of giving a tool to monitor the behaviour of autistic children. Creating video chat forum for autistic children from where they can build their active social life.

We look forward to have customized products like books, CD's and counselling sessions via mobile intervention tool thus will help parents and special educators to deal with autistic children efficiently.

ACKNOWLEDGMENT

Authors extend sincere thanks to Dr. Shushma Bhandarkar (Developmental Therapist) at Spandan School for mentally challenged, Mrs. Sushila Ben (Speech Therapist) at Disha Special School & Autism Center and Mr. J.C. Kathrani, Anmol Special School for their cooperation provided to conduct preliminary investigations and guidance for preparing medical records.

REFERENCES

- [1] Tamara C. Daley (2003) "From symptom recognition to diagnosis: children with autism in urban India" *Social Science & Medicine*, Elsevier doi:10.1016/S0277-9536(03)00330-7
- [2] Wass, S.V. & Porayska-Pomsta, K. "The uses of didactic and cognitive training technologies in the behavioral treatment of Autism Spectrum Disorders"
- [3] Rogers, S., Vismara, L. (2008) "Evidence-based comprehensive treatments for early autism" *Journal of Clinical Child & Adolescent Psychology* 37, 8–38
- [4] Iris Manor-Binyamini (2011) "Parental Coping with Developmental Disorders in Adolescents within the Ultraorthodox Jewish Community in Israel" *J Autism Dev Disord*, Springer Science+Business Media.
- [5] E. I. Barakova and T. Lourens, (2010) "Expressing and interpreting emotional movements in social games with robots," *Personal Ubiquitous Comput.*, vol. 14, pp. 457–467.
- [6] T. Lourens, R. van Berkel, and E. Barakova, (2010) "Communicating emotions and mental states to robots in a real time parallel framework using laban movement analysis," *Robot. Auton. Syst.*, vol. 58, pp. 1256–1265.
- [7] McConnellSR (2002) "Interventions to facilitate social interaction for young children with autism" review of available research and recommendations for educational intervention and future research. *J Autism Dev Disord* 2002;32:351–72.
- [8] M.S. Mythili and A.R.Mohamed Shanavas (2014) "A Novel Approach to Predict the Learning Skills of Autistic Children using SVM and Decision Tree" (IJCSIT) *International Journal of Computer Science and Information Technologies*, Vol. 5 (6), 2014, 7288-7291.
- [9] Thilini Ariyachandra et al. (2016) "Analytics In Behavioral Intervention Education" *Issues in Information Systems* Volume 17, Issue III, pp. 236-243
- [10] T. Miller et al. (2006) "Using a Digital Library of Images for Communication: Comparison of a Card-Based System to PDA Software," presented at First International Conference on Design Science Research in Information Systems and Technology, Claremont, CA.
- [11] Valentina Bartalesi et al. (2014) "An Analytic Tool for Assessing Learning in Children with Autism" Springer International Publishing Switzerland UAHCI/HCI 2014, Part II, LNCS 8514, pp. 209–220.
- [12] D. J. Moore, P. McGrath, and J. Thorpe, (2000) "Computer Aided Learning for people with autism - a framework for research and development," *Innovations in Education and Training International*, vol. 37, pp. 218-228.
- [13] M.S. Mythili and A.R.Mohamed Shanavas (2014) "A Novel Approach to Predict the Learning Skills of Autistic Children using SVM and Decision Tree" (IJCSIT) *International Journal of Computer Science and Information Technologies*, Vol. 5 (6), 2014, 7288-7291
- [14] Felix D. C. C. Beacher et al., "Sex Differences and Autism: Brain Function during Verbal Fluency and Mental Rotation", June 2012 | Volume 7 | Issue 6 | e38355
- [15] Golan, O., Ashwin, E., Granader, Y., McClintock, S., Day, K., Leggett, V. & Baron-Cohen, S. (2009). Enhancing Emotion Recognition in Children with Autism Spectrum Conditions: An Intervention Using Animated Vehicles with Real Emotional Faces. *Journal of Autism and Developmental Disorders* 40(3): 269-279.
- [16] El. Kaliouby, R., Picard, R. & Baron-Cohen, S. (2006) Affective computing and autism. *Progress in Convergence: Technologies for Human Wellbeing* 1093: 228-248.
- [17] Keay-Bright, W. (2006) Re Activities (c): autism and play. *Digital Creativity* 17(3): 149-156.
- [18] G.Rajendranand P.Mitchell (2000) "Computer mediated interaction in Asperger's syndrome: the Bubble Dialogue program," *Computers and Education*, vol. 35, pp. 189-207.

Context Aware Fuel Monitoring System for Cellular Sites

*Mohammad Asif Khan, Ahmad Waqas, Qamar Uddin Khand, Sajid Khan

*Department of Electrical Engineering
Department of Computer Science
Sukkur IBA University
Airport Road, Sukkur-65200, Pakistan

Abstract—The past decade has been very productive for cellular operators of Pakistan, as their subscribers have grown exponentially with increase in revenue. After this wave of rising, the operators have now reached to saturation level, with the highest teledensity of all time. These Cellular Networks consist of Cell sites, which need electrical power to run. Because of electrical power shortage in Pakistan, the power needs of cell site are fulfilled by the use of electrical power generators which are installed on each site. These generators run on fossil fuel, a large amount of which is being theft from sites. This has very negative impact on Network availability and Operator's operational expenditure. To cope with this major issue of fuel theft, an embedded system is being designed and tested. This paper highlights this issue of the telecom sector and discusses the design and results of the proposed system. This system would reduce the cell site operational cost and will increase its availability in the service area.

Keywords—Fuel theft; fuel sensor; fuel management; remote monitoring

I. INTRODUCTION

Pakistan telecom sector has passed peak era and has entered into a position where they have started saturation in terms of subscriber penetration and now cellular operators have started optimization of their resources to reduce operational expenditure (OPEX). Operators consume most of their OPEX on the operation of their basic network entity called Base Transceiver Station (BTS). Thousands of BTS are installed by cellular operators throughout the country in the shape of BTS towers. There is temperature sensitive equipment in those BTS sites. Numbers of air-conditions are installed at the site for maintaining required temperature of that temperature sensitive equipment. The operation of these BTS towers requires heavy electrical power, which is being supplied to sites from National Electrical Grid.

Pakistan is suffering from severe Electrical shortage for past fifteen years. Load shedding from National Grid creates trouble for continuous functioning of BTS sites. In big cities of Pakistan there is scheduled electrical power shutdown from Grid for 8 to 10 hours and in small cities, towns 12 to 18 hours are being observed [1]. This large scale regular power shutdown has very negative impact on revenue, availability, operation, and maintenance of cellular networks.

To cope with a shortage of electrical power, Cellular Operators use electrical power generator as a secondary source

of power for cell sites. Operators are spending a handsome amount of money for fuelling these generators. The system that is used by cellular companies for fuel filling and monitoring is manual. Because of this manual fuel system, a large amount of fuel is being theft from the site by either site guard or filler of fuel [2]. In the current manual filling system, it is very hard for Operator to identify who has performed the fuel theft.

To cope with this major issue of the telecom sector, a system is designed and discussed in this paper called the Context-Aware Fuel Monitoring System. This intelligent system updates the site engineer about different levels of the fuel and any abnormal activity if occur at the fuel tank of BTS. This context-aware system updates the site engineer through a text message that what type of anomaly has occurred with the fuel tank. This system informs the site engineer about levels of fuel, such as, what is current level of fuel, at which time of the day the generator is filled with fuel, how much amount of fuel is filled, how much liters it has consumed while running.

The main purpose behind this system is, it will report fuel theft to site engineer, also it will maintain fuel log of the site for budgeting, operation, and maintenance. By using this system, the site engineer will have information about the fuel status, and record of the log, also the system will monitor fuelling continuously for reporting any anomaly. The prototype of the system is being developed and deployed at the test site, results were gathered to test different fuel theft conditions and remote monitoring. It has been found that this system greatly reduces the fuel theft and OPEX of the Network also it increases Network availability.

This research paper is divided into seven sections. Section 2 is about the related work carried out on such issues by different researchers. Section 3 of the paper is about the electrical power setup at cellular sites. Section 4 explains the problem statement of research, whereas Section 5 discusses the design, implementation, and benefits of the system. Section 6 is used to show the results of the system in form of graphs, whereas Section 7 is a conclusion and future work of the system.

II. RELATED WORK

The researchers have done good work to cope with the issue of fuel theft, done at various places like cell sites, vehicles, stations, etc. In this regard Kunal Dhandel et al., have done a survey on fuel level measurement techniques used at

various places [4]. This work represents only a survey with no idea of implementation. R Surendra et al. have done work to remotely access the cell site, and ON/OFF the generators [5], the difference between this work and the proposed system is it also cope with fuel theft where the referenced work only remotely access site generator. Ghenga et al. have worked on the development of a mechanical sensor to accurately measure the fuel level and report it back to the remote mobile phone and aplicom [6], this work has only developed fuel sensor to be used for measuring fuel level.

The research is also going on fuel theft issue in another place instead of cell sites. Riny Sulityowti et al. have worked on the remote monitoring of fuel tank by the use of the android application, for that purpose they have used Bluetooth and GSM technology [7]. Nandini Hiremath et al. have been working to detect fuel theft in a tank used anywhere, they have used ARM7 micro controller with fuel sensor to report the theft or any leak in the tank [8]. Deo Ashwini et al. worked on the computer based monitoring of fuel tank and its reporting on the web with the help of web API [9]. Alwyn Hoffman et al. have worked to identify different issue relating to fuel used in logistics [10]. Most of these systems are not coping with all kind of fuel theft also these designs don't look scalable in terms of other features which can be incorporated in the proposed system in future.

There has been a significant amount of work on cellular sites monitoring and its patents are filed. Bejiman Stump has worked on the monitoring of cellular sites power and the discharge of the battery. They are checking different sites parameter to check the feasible conditions for battery discharge, in another case, they will run the power generator [11]. Philippe Gagnon et al. have worked on the monitoring of battery backup power remotely. This work is only working on the monitoring of battery bank and evaluating the other parameters of remote site those will impact the backup time [12].

III. ELECTRICAL POWER AT CELLULAR SITE

The cellular sites provide service coverage to the subscribers using different equipment which are being installed there. At the Cellular site, there is two major equipment which consumes power; cooling systems and telco equipment. If the power of cell site is shut down then the result is no availability in its coverage area. Availability of cell site has remained the key concern of cellular operator due to its direct impact on revenue. Therefore, continuous supply of power to a cell site is very important for any cellular operator.

Keeping in view the importance of electrical power at the cell site, telecom operators provide redundant electrical supply to the site for keeping it up all time. Normally in Asia Pacific region, the electrical power supply from national grid varies from 12KVA to 40KVA [3]. The electrical connectivity at any cell site is shown in Fig. 1, where the main supply comes from national grid in AC current form and goes to Automatic Transfer Switches (ATS). Similarly, backup power is also supplied from Generator set, which runs on the fuel, present in the fuel of tank attached with it. Generator's electrical power supply also goes to ATS system. The function of ATS is a switch to the available power source. In ATS, if all supplies are

up then, normally the primary power supply is set from the national electrical grid. In case the Grid supply goes off, then ATS switches to generator supply. Most of the telecom equipment operates on DC power, therefore rectifier is placed between equipment and ATS supply.

The third electrical power backup is battery bank. Batteries are charged and used when both AC supplies go off. These batteries also help in the switching between Grid supply to Generator supply and vice versa. Most of the under developed countries like Pakistan faces severe shortage of Electrical power at national power grid, and as result half of the time.

Electrical power is shut down by the government for load balancing. To cope with shortage of electrical power from national grid, cellular sites availability highly depends on generators electrical power, which is running on fuel.

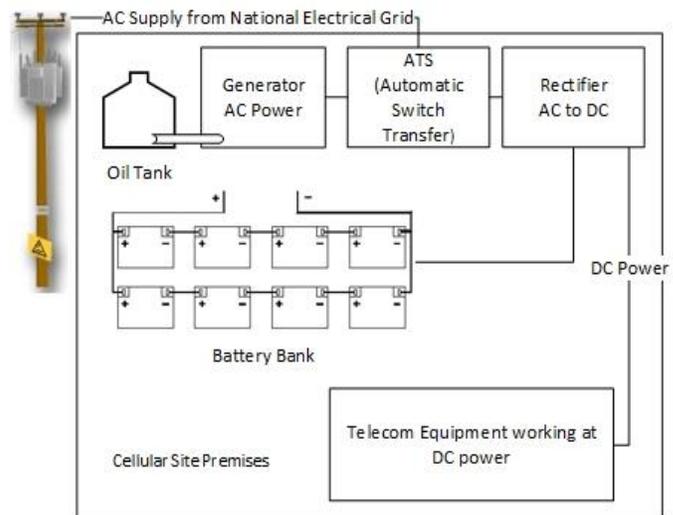


Fig. 1. Electrical power supply at cellular site.

IV. FUEL THEFT AT CELLULAR SITES

The cellular sites' availability highly depends on the running of generators, in case of supply from national grid goes off. Generators work on the fuel i.e. diesel, therefore, each cellular site contains a fuel tank attached to a generator. In case of grid supply goes off then the generator is the best hope for site availability, but if fuel is finished, then the last line of defense for the site is battery backup. The battery bank runs for short time, and as result, sites go offline if power from a generator or national grid is OFF for longer period.

The cellular operators heavily invest on fueling of sites. Most of the time they filled the site in advance to avoid any scenario in which fuel is finished, and filling team reached after the site goes offline. But due to bad law and order condition in under developed countries like Pakistan, fuel is theft from cellular sites and in such case when grid supply also goes off then generator doesn't work due to unavailability of the fuel tank. In some part of the countries like Pakistan, almost half of the fuel is theft from sites. A survey was done to estimate fuel consumption and theft ratio. At 500 cell sites in peak seasons of May, June, July and August, when National Grid supply less electricity to sites, the cellular operators consume approximately 250,000 litres of diesel. The report

occurred and is reported to server and site engineer. There is also the possibility to decrease in fuel level when the generator is also running, in such case, the fuel consumption rate will be high as compared to normal consumption of generator, as result fuel theft is also reported. During fuel level check, if there is no change in fuel level then the system re-checks fuel level after a set period and no fuel theft is reported. In this way, fuel theft is detected, as shown in flow chart of Fig. 3.

B. Benefits of System

This system has high commercial viability as described in section 3. The system can avoid fuel theft and can save large OPEX of the cellular operator. But there are other benefits of this system. The system will reduce the malfunctioning of sensitive electronic equipment, which mostly malfunctions due to electrical shutdown. There is another benefit of an increase in site availability, which is one of major benefit from the proposed system, as with proper fueling there is less chance of site down. This system also reduces the security cost, which companies mostly spend on fuel theft investigation. This system will also improve the mobility of filling teams in case if fuel level reaches any threshold value and team reach site before the site goes down. The fuel audits will be easier and accurate with such system deployment.

VI. RESULTS OF THE SYSTEM

The prototype of the Context Aware Fuel Monitoring System was deployed on a running generator of 12KVA for more than five days. The system working was monitored and readings were taken successfully.

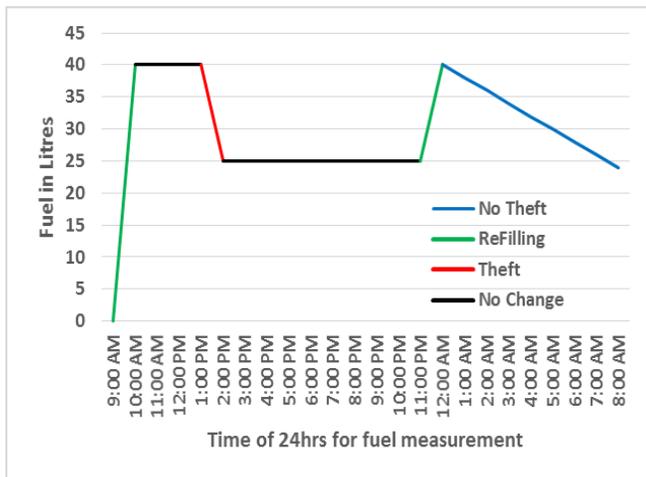


Fig. 4. Fuel theft detection with proposed system.

The system successfully detected the fuel theft, which was intentionally and randomly done in various ways. The system reported and logged various activities of fuel filling, running of generator and fuel theft.

The system at the site reported fuel level to the server after every minute. In Fig. 4 the graph shows the reading of the 3rd day. The red line show theft detection when generators were running but fuel theft was done to avoid abrupt changes in fuel level. The graph show generator was running from 1:00 pm to 4:00 pm, but fuel consumption was very high, instead of

standard 6 liters (with a margin of error), the consumption is 15 liters. Hence fuel theft was reported to server and site engineer. The red line is where fuel was theft. The black line shows no change in fuel level, means neither fuel is theft nor generator is running. The blue line appeared when the generator is running and fuel is consumed according to standard values as shown between 11:00 pm to 4:00 am.

Fig. 5 shows the results taken on the 5th day of testing when filling was done also fuel was theft with an abrupt change in fuel level, which is a very common way of fuel theft. At 09:00 am generator had a filling with 40 liters of fuel which was reported and shown with the green line. The theft occurred when someone suddenly took 15 liters of fuel from tank and generator was not running. This was reported as theft and shown with a red line. Again, filling of 15 liters was at 11:00 pm. The blue line show generator running and no theft was done, as consumption rate is matching with standard values.

All the other hardware and software parameters of the system were tested and found good during the period of testing. All the messages were properly logged on the server, and anomalies reported to server and site engineer.

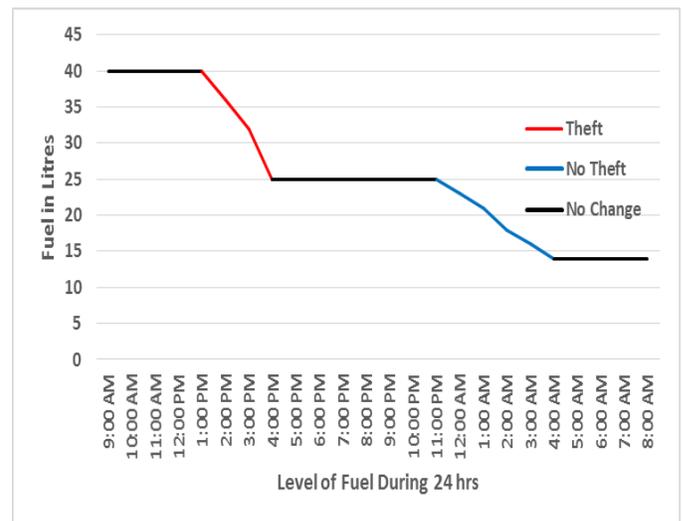


Fig. 5. Common way of fuel theft detection.

VII. CONCLUSION AND FUTURE WORK

This paper discussed the design, implementation, and result of Context Aware Fuel Monitoring System which coped with the issue of fuel theft occurred at Cellular Sites and a large portion of Operator's revenue is lost. The proposed system was tested and results shows, all possible ways of fuel theft were detected, reported and logged for any future usage like an audit. This system has high commercial viability and will be deployed in industry. This system can also be used in other places where fuel theft is reported.

This system has very good directions in future. The research team would work and include other modules in a system which can report the electrical values (voltage, load, battery charge, fluctuation of power, etc.). There will be an android application for this system which can be used for a client (Engineer's Phone) and server (GUI access) activities.

REFERENCES

- [1] Rashid Aziz and Munawar Baseer Ahmad, "Pakistan's Power Crisis The Way Forward", United States Institute of Peace: Special Report, 375, pp 2-3, June 2015.
- [2] Karen Hampton, "Ringing The Changes", Energy Storage Journal, Nov 2013 Issue.
- [3] Wissam Balshe, "Power system considerations for cell tower applications", Technical information from Cummins Power Generation, Power topic #9019, 2011
- [4] Kunal Dhande1, Sarang R. "Fuel Level Measurement Techniques: A Systematic Survey", International Journal of Research in Advent Technology, Vol.2, No.4, April 2014 E-ISSN: 2321-9637.
- [5] R.Surendra, B.Karunaiah, Murali Mohan, "Power Management Of Cell Sites", International Journal of Computer Technology and Applications ,Vol 3 (1), pp 5-8.
- [6] Gbenga Daniel Obikoya, "Design, Construction, and Implementation of a Remote Fuel-level Monitoring System", EURASIP Journal on Wireless Communications and Networking,,December 2014.
- [7] Riny Sulityowati , Bayu Bhahttra Kurnia Rafik, "Prototype Design of a Realtime Monitoring System of a Fuel Tank at a Gas Station Using an Android-Based Mobile Application", Proceedings of Second International Conference on Electrical Systems, Technology and Information, 2015
- [8] Nandini Hiremath, et al., "Smart Fuel Theft Detection using Microcontroller ARM7", International Journal of Engineering Research & Technology. Volume. 4 - Issue. 09, September - 2015
- [9] Deo Ashwini et al. , "Wsn Based Fuel Level Detection and Dispensary Monitoring System for Moving Vehicles", Journal of Emerging Technologies and Innovative Research., Volume 3, Issue 4 April 2016
- [10] Alwyn J. Hoffman et al. , "An Investigation into the Economics of Fuel Management in the Freight Logistics Industry", 17th International Conference on Intelligent Transportation Systems (ITSC)., October, 2014, China
- [11] Benjamin Stump, "Hybrid power management system and method for unmanned remote cell sites", Westell Inch, Patent issued Aug, 2014
- [12] Philippe Gagnon, Denis Pomerleau, Roger Paradis, "Back-up power system and monitoring system therefore", Avestor Limited Partnership, Patent issued May

Text Steganography using Extensions Kashida based on the Moon and Sun Letters Concept

Anes.A.Shaker, Farida Ridzuan, Sakinah Ali Pitchay
Faculty of Science and Technology
Universiti Sains Islam Malaysia
Nilai, Negeri Sembilan, Malaysia

Abstract—Existing steganography methods are still lacking in terms of capacity. Hence, a new steganography method for Arabic text is proposed. The method hides secret information bits within Arabic letters using two features, which are the moon and sun letters and the redundant Arabic extension character “-” known as Kashida. The Arabic alphabet contains 28 letters, which are classified into 14 sun letters and 14 moon letters. This classification is based on the way these letters affect the pronunciation of the definite article (ال) at the beginning of words. This method uses the sun letters with one extension to hold the secret bits ‘01’, the sun letters with two extensions to hold the secret bits ‘10’, the moon letters with one extension to hold the secret bits ‘00’ and the moon letters with two extensions to hold the secret bits ‘11’. The capacity performance of the proposed method is then compared to three popular text steganographic methods. Capacity is measured based on two factors which are Embedding Ratio (ER) and The Efficiency Ratio (TER). The results show that the Letter Points and Extensions Method produces 24.91% and 21.56% as the average embedding ratio and the average efficiency ratio correspondingly. For the Two Extensions ‘Kashida’ Character Method, the results for the average embedding ratio and the efficiency ratio are 56.76% and 41.81%. For the Text Using Kashida Variation Algorithm method, the average embedding ratio and the average efficiency ratio are 31.61% and 27.82% respectively. Meanwhile, the average embedding ratio and the efficiency ratio for the Proposed Method are 61.16% and 55.70%. Hence, it is concluded that the Proposed Method outweighs the other three methods in terms of their embedding ratio and efficiency ratio which leads to the conclusion that the Proposed Method could provide higher capacity than the other methods.

Keywords—Text steganography; Arabic text; extension Kashida; capacity

I. INTRODUCTION

Steganography can be described as the concealment of confidential messages implanted within other apparently regular messages, graphics or sounds [1]. Steganography is defined as the study of these invisible communications. Steganography deals with ways of hiding the existence of the communicated data in such a way that it remains confidential [2]. It maintains secrecy between two communicating parties. In text steganography, secrecy is achieved by embedding data into cover text and generating a stego-text. There are different types of steganography techniques and each has its own strengths and weaknesses. According to the medium used for

the steganography, carrier files can be termed as cover text, cover images, cover audio, cover video or cover network [2]. The main drawback for all the methods is their low capacity wherein only a small amount of bits are allowed to be hidden. The lower the capacity of a method the bigger the carrier file must be to hide the secret message. Hence, a new method is proposed which use the concept of moon and sun letters with extension Kashida. Both the sun letters and moon letters are at the beginning of a word preceded by (ال) [3]. The Arabic language comes in two groups with each group consisting of 14 letters.

The objectives of this paper are: 1) to present a proposed work using the concept of moon and sun letters with extension Kashida, and 2) to present an evaluation study of the proposed work compared to three existing methods. The rest of this paper is organized as follows: related works are presented in Section II followed by an explanation of the proposed work in Section III. Section IV explains how the evaluation was carried out. Section V presents the results and a discussion of the results. Finally, the conclusion and future research suggestions are presented in Section VI.

II. RELATED WORKS

For the past few years, a lot of research has focused on the development and potential applications of Arabic script steganography.

A. Steganography Using Multiple Diacritics

An entire message can be hidden in a single diacritic mark by generating a number of extra-diacritic keystrokes equal to the binary number representing the message. For this scenario, consider this example of (110001)b as a secret message, the first diacritic is repeated 3 extra times (3 = (11)b); the second one, 0 extra times (0 = (00)b); and the third one, 1 extra time (1=(01)b) [4].

B. Word Spelling Method

The author presents a new text steganography method for hiding data in English texts. This method is based on substituting US and UK spellings of words. In English some words have different spelling in US and UK. For example “program” has a different spelling, in UK (program), and US (program). By using this feature, the author proposes his method for hiding data in an English text. In this method, the data is hidden in the text by substituting such words [5].

C. Vertical Displacement of the Points

This method makes use of dotted letters. Some language texts, which include Arabic and Persian, come with a substantial number of dotted letters. The Arabic text has 26 characters, of which 13 have dotted letters, while the Persian text has 32 characters, of which, 22 have dotted letters. With this method, '1' is encoded to move up the point, or else '0' is encoded. This process is replicated for the following dotted characters in the text as well as the following bits of information [6].

D. Mixed-case Font

The concept for this method was formed during an Internet search for popular fonts used for chatting and presentations. The author came across an innovative kind of font that can type capital and small letters in sequence. For instance, if one typed the word 'software', this word would appear as 'SoFtWaRe'. Sometimes the size of the letters would differ, and at other times they would be of similar size. The authors developed an innovative text steganography technique for the transmission of confidential information using this newly-discovered font [7].

E. Move the Diacritic Up

The Arabic language comes with diacritics. The inclusion of these diacritics in most Arabic texts is usually non-obligatory and this study emphasized the employment of this characteristic. The vertical shifting of the diacritic is in accordance to the character. 'Zero' denotes no change, and 'one' denotes the increased distance between the letter and its diacritics [8].

F. Using "La" Word

This method is proposed based on the feature code using the "La" word. This word is obtained by connecting "Lam" and "Alef" letters into a single word. The hiding process is based on the existence of two forms of this word which are the special form "La" ("ﻻ") which has a unique code and the normal form "La" ("ﻻ") by inserting Arabic extension character between the "Lam" and "Alef" letters. The normal form "La" is used to hide bit zero while bit one is hidden using the special word [9].

G. Improved 'La' word

The authors proposed an enhanced method for the utilization of the "La" word which involved the use of a different Unicode of 'Lam' and 'Alef' to exploit the 'La' word into both special and normal forms. This recommendation takes into account the fact that each letter comes with four dissimilar outlines depending on its location in the word [10].

H. Sharp-edges Method

This method exploits the sharp-edged Arabic characters for the concealment of confidential information. Keys are introduced to facilitate the positioning of the secret bit. The diverse number of sharp edges in Arabic characters enhances the concealment effectiveness of bits '1' and '0'. The character with one sharp edge can conceal either secret bit '1' or '2'. Concurrently, if the number of sharp edges is two, the possible bit location is 11, 10, 00 or 01 [11].

I. Using Letter Points and Extensions method

This process takes advantage of the fact that more than half the text letters in the Arabic language come with dots. While these dotted letters were loaded with the confidential bit "one", the letters without dots were loaded with the confidential bit "zero". As the confidential information needs to conform to the cover-text letters, not every letter is loaded with confidential bits. Other than the letters used to indicate the particular letters containing the confidential bits, redundant Arabic extension characters are also included in the system. The advantage that comes with letter extensions is the fact that their utilization does not affect the writing content in any way [12].

J. Two-extension 'Kashida' Character

This paper presents a novel steganography method useful for Arabic and other similar languages. This method benefits from the feature of having the Kashida character, "ـ" in Arabic script. An extension character is inserted after a letter in the cover object if the secret bit is 'zero'. Instead, if the secret bit is 'one', two consecutive extension characters will be inserted [13].

K. Enhanced Kashida

The author utilized the Kashida by encoding the original text document with Kashida according to a specific key which was produced before the encoding process. Kashida are inserted before a specific list of characters {ذ د و ا} until the end of the key is reached where the kashida is inserted for a bit 1 and omitted for a bit 0. This process is repeated until the end of the document is reached in a round robin fashion [14].

L. Text Using Kashida Variation Algorithm (KVA)

Most of the previous methods apply the same procedure for the whole text which may allow steganalysis to study the text format, hence, to breaking the code or, in other words, find the hidden message. However, this study proposed a method to apply four scenarios randomly to improve data privacy.

The method presents four scenarios. The first scenario is by adding Kashida after pointed letters to be encoded as one, otherwise, it is encoded as zero. The second scenario is by adding Kashida after nonpointed letters to be encoded as one, otherwise, it is encoded as zero. The third scenario is by adding Kashida after letters to be encoded as one. Otherwise, it is encoded as zero. The fourth scenario is by adding Kashida after letters to encode as zero. Otherwise, it is encoded as one. This method provides a high embedding ratio as it allows bits to be encoded in four different scenarios [15].

In summary, most of diacritics-based methods are simple to implement and provide higher capacity and robustness than others [10]. However, these methods cannot be used in text, in which, the appearance of all diacritics is important, like the Holy Quran. Conversely, Kashida-based methods provide good capacity [10] and could be used in printed documents with different font formats. However, they are easily detected or observed. Thus, many researchers add more security features to decrease the number of Kashidas and enhance the capacity [10]. On the other hand, shifting line, word or points methods are simple to implement however, their drawback is the high

probability of destroying the watermark when retyping or printing [10]. Another weakness is that they are also noticeable by Optical Character Recognition (OCR) programs [10].

III. PROPOSED WORK

A new method that could provide higher capacity is needed to improve the implementation of steganography generally. The proposed work hides the message in Arabic text using the characteristics of Arabic language. In Arabic language, there are two groups of letters, namely sun letter (solar letters) and moon letter (lunar letters). The secret text is hidden in the form of zeros and ones represented by the 16-bit Unicode for each character (the UTF-8 encoding scheme uses 16 bits to represent one Arabic character). Table 1 consists of the letters classifications.

TABLE I. MOON AND SUN LETTERS

Moon letter				Sun letter			
1	أ	8	خ	1	ت	8	ش
2	ب	9	ف	2	ث	9	ص
3	غ	10	ع	3	د	10	ض
4	ح	11	ق	4	ذ	11	ط
5	ج	12	ي	5	ر	12	ظ
6	ك	13	م	6	ز	13	ن
7	و	14	هـ	7	س	14	ل

The proposed method presents four scenarios. The first scenario is implemented by adding a Kashida after a sun letter to represent (00). The second scenario is implemented by adding two kashidas after a sun letter to represent (11). In the third scenario, a kashida is added to represent (01) after a moon letter. The fourth scenario is implemented by adding two kashidas after a lunar letter to represent (10). The pseudo code of the new proposed method is presented in Fig. 2.

IV. EVALUATION

The proposed work is evaluated and compared with three other methods [12]-[14]. The first method uses the letter points and extensions method [12]. The second method is the two extension “Kashida” character and the third method is the frequency recurrence of characters. The main aim of the research is to improve the steganography in terms of capacity. Capacity is determined by the embedding ratio and the efficiency ratio features.

Equation (1) is used to calculate embedding ratio (ER).

$$ER = \frac{\text{Total letters of cover text} - \text{Letters of embedded message}}{\text{Total letters of cover text}} \quad (1)$$

The efficiency ratio (TER) is computed following (2).

$$TER = \frac{\text{Total letters of cover text} - \text{Letters of embedded message}}{\text{Total letters of cover text}} \quad (2)$$

Algorithm Encode
 Input: Cover text denoted by C_t
 Binary secret message denoted by m_1, m_2, \dots, m_n
 where $n =$ the message size
 Output: Stego_Object denoted by S
 Step 1: $S = C_t$
 Step 2: for $i=1$ to n step 2
 if ($m_i = 0$ and $m_{i+1} = 0$) or ($m_i = 1$ and $m_{i+1} = 1$)
 Search S to get first Sun letter location k
 if ($m_i = 0$ and $m_{i+1} = 0$)
 Insert one Kashida in location $k+1$
 else
 Insert two Kashida in location $k+1$
 end if
 else
 Search S to get first Moon letter location k
 If ($m_i = 0$ and $m_{i+1} = 1$)
 Insert one Kashida in location $k+1$
 else
 Insert two Kashida in location $k+1$
 end if
 End if
 End for
 Step 3: output S

Algorithm Decode
 Input: Stego_Object denoted by S
 Output: Binary secret message denoted by m_1, m_2, \dots, m_n where
 n equal the message size
 $i=1$
 $k=1$
 while $k \leq S_length$
 Begin
 $j=1$
 if the S_k letter is Sun letter
 Begin
 if ($S_{k+j} = \text{Kashida}$ and $S_{k+1+j} = \text{Kashida}$)
 Begin
 $m_i=1$ and $m_{i+1}=1$
 else $m_i=0$ and $m_{i+1}=0$
 end if
 else
 if ($S_{k+j} = \text{Kashida}$ and $S_{k+1+j} = \text{Kashida}$)
 Begin
 $m_i=1$ and $m_{i+1}=0$
 else $m_i=0$ and $m_{i+1}=1$
 end if
 end if
 $i=i+1$
 $j=j+1$
 End while
 Output Binary secret message m

Fig. 1. Pseudo code and decode of the proposed method.

The evaluation is carried out similarly to [6]. Ten cover texts are selected from highly circulated Iraqi newspapers [6]. The word “GOOD” is used to be embedded as secret text in the cover texts. The source of the cover text is presented in Table 2.

TABLE II. COVER TEXT SOURCES

No	Cover text
1	www.almadapaper.net/3784
2	www.almadapaper.net/2440
3	www.almadapaper.net/3000
4	www.almadapaper.net/3001
5	www.almadapaper.net/3010
6	www.almadapaper.net/3100
7	www.almadapaper.net/2001
8	www.almadapaper.net/2010
9	www.almadapaper.net/2111
10	www.almadapaper.net/2054

V. RESULTS AND DISCUSSION

An example of embedding the word ‘GOOD’ in Cover Text 1 using the four related methods for comparison are presented in Table 3.

TABLE III. EMBEDDING RESULTS

Methods	The Cover texts after imbedding
A Novel Arabic Text Steganography Method Using Extensions [12]	ولكنَّ اللهجة العامية طغت على ألسن الناس، حتَّى أننا وجدنا البعض ممن هو من بني جلدتنا قد دعا البعض إلى إلغاء التكلم باللغة الفصحى.
Steganography in Arabic text using Kashida variation algorithm method [14]	ولكنَّ اللهجة العامية طغت على ألسن الناس، حتَّى أننا وجدنا البعض ممن هو من بني جلدتنا قد دعا البعض إلى إلغاء التكلم باللغة الفصحى.
Improved Method of Arabic Text Steganography Using the Extension “Kashida” Character [13]	ولكنَّ اللهجة العامية طغت على ألسن الناس، حتَّى أننا وجدنا البعض ممن هو من بني جلدتنا قد دعا البعض إلى إلغاء التكلم باللغة الفصحى.
Proposed method	ولكنَّ اللهجة العامية طغت على ألسن الناس، حتَّى أننا وجدنا البعض ممن هو من بني جلدتنا قد دعا البعض إلى إلغاء التكلم باللغة الفصحى.

The calculation of Embedding Ratio (ER) and the efficiency ratio (TER) are presented in Table 4.

TABLE IV. ER AND TER RESULTS

Methods	Capacity Evaluation	
	Average ER	Average TER
A Novel Arabic Text Steganography Method Using Extensions [12]	24.91%	21.56%
Steganography in Arabic text using Kashida variation algorithm method [14]	31.61%	27.82%
Improved Method of Arabic Text Steganography Using the Extension “Kashida” Character [13]	56.76%	41.81%
Proposed method	61.16%	55.70%

Table 4 shows the average embedding ratio and the efficiency ratio results for letter points and the extensions method as 24.91% and 21.56%, respectively. The results for this method are pretty low because it is based on the concept of pointed letters. The existence of sentences without pointed letters could have a high impact on the capacity performance of this method. For Kashida Variation Algorithm method, the results for average embedding ratio and the efficiency ratio are 31.61% and 27.82%, respectively. The results are low because this method is also based on the concept of pointed letters. Similarly, as in the previous method, reliance on pointed letters in the sentence could have an effect on its capacity performance. For the two extension “Kashida” character methods, the results for the average embedding ratio and the efficiency ratio are 56.76% and 41.81%, respectively. The results for this method are considered good because it was developed based on the concept of adding Kashida after any letter. Hence, this method does not rely on certain characteristics possessed by any letter in the cover text. The proposed work results for the average embedding ratio, and the efficiency ratio, are 61.16% and 55.70%, respectively. This method produces higher results compared to the others due to its chosen features. The first features which are the moon and sun letters allow secret bits to be hidden in any letter as all Arabic words will contain either moon or sun letters. In addition, the proposed method allows two secret bits to be hidden in a letter. Thus, more secret bits can be hidden in shorter sentences.

VI. CONCLUSION AND FUTURE WORK

This paper presents a novel steganography method useful for Arabic language electronic writing using extension Kashida based on the concept of moon and sun letters. The proposed method uses sun letters with Kashida to represent (01), sun letters with two Kashidas to represent (10), moon letters with Kashida to represent (00) and moon letters with two Kashidas to represent (11). Kashida characters are used beside the Arabic letters to note which specific letter is holding the hidden secret bits. Letter extension is used as it will not affect the writing content. The proposed method outweighs the other three methods because of its capacity performance results. It can be concluded that choosing the right features to hide secret text is critical in determining the capacity performance of a steganography method. The advantage of implementing the moon and sun letters concept is that it is able to increase the probability of hiding the secret bits in any letter. Nonetheless, it is also very important to maintain the imperceptibility aspect while improving capacity. In future, this method will be evaluated in terms of its imperceptibility.

REFERENCES

- [1] A. Siper, R. Farley, and C. Lombardo, “The Rise of Steganography,” Proc. Student/Faculty Res. Day, CSIS, Pace Univ., pp. 1–7, 2005.
- [2] J. Kour and D. Verma, “Steganography Techniques –A Review Paper,” Int. J. Emerg. Res. Manag. &Technology, vol. 9359, no. 35, pp. 2278–9359, 2014.
- [3] عبد السلام محمد هارون. "قواعد الإملاء" مكتبة الانجلومصرية، 1993.
- [4] A. Gutub, Y. Elarian, S. Awaideh, and A. Alvi, “Arabic Text Steganography Using Multiple Diacritics,” no. May 2016, 2008.
- [5] M. Shirali-Shahreza, “Text steganography by changing words spelling,” Int. Conf. Adv. Commun. Technol. ICACT, vol. 3, no. March, pp. 1912–1913, 2008.

- [6] M. H. Shirali-Shahreza and M. Shirali-Shahreza, "A new approach to Persian/Arabic text steganography," Proc. - 5th IEEE/ACIS Int. Conf. Comput. Info. Sci., ICIS 2006. conjunction with 1st IEEE/ACIS, Int. Work. Component-Based Softw. Eng., Softw. Arch. Reuse, COMSAR 2006, vol. 2006, pp. 310–315, 2006.
- [7] A. Ali and A. Saad, "New Text Steganography Technique by using Mixed-Case Font," Online J. Comput. Sci. Inf. Tehcnology, vol. 3, no. 2, pp. 138–141, 2013.
- [8] A. Odeh and K. Elleithy, "Steganography in arabic text using full diacritics text," 25th Int. Conf. Comput. Appl. Ind. Eng. CAINE 2012 4th Int. Symp. Sens. Netw. Appl. SNA 2012, no. November, 2012.
- [9] M. Shirali-shahreza and M. H. Shirali-shahreza, "An Improved Version of Persian / Arabic Text Steganography Using ' La ' Word," no. August, pp. 26–27, 2008.
- [10] R. A. Alotaibi and L. A. Elrefaei, "Arabic Text Watermarking : A Review," Int. J. Artif. Intell. Appl., vol. 6, no. 4, pp. 01–16, 2015.
- [11] N. A. .KBM//Roslan, R. Mahmud, and N. I. Udzir, "Sharp-edges method in Arabic text steganography," J. Theor. Appl. Inf. Technol., vol. 33, no. 1, pp. 32–41, 2011.
- [12] W. Al-Alwani, A. Bin Mahfooz, and A. A. A. Gutub, "A Novel Arabic Text Steganography Method Using Extensions," Proceeding World Acad. Sci. Eng. Technol., vol. 1, no. 3, pp. 483–486, 2007.
- [13] A. Gutub, W. Al-Alwani, and A. Mahfoodh, "Improved Method of Arabic Text Steganography Using the Extension „Kashida“ Character," Bahria Univ. J. Inf. , vol. 3, no. 1, pp. 68–72, 2010.
- [14] Y. M. Alginahi, M. N. Kabir, and O. Tayan, "An Enhanced Kashida-Based Watermarking Approach for Increased Protection in Arabic Text-Documents Based on Frequency Recurrence of Characters," no. 5, pp. 381–392, 2014.
- [15] A. Odeh, K. Elleithy, and M. Faezipour, "Steganography in Arabic text using Kashida variation algorithm (KVA)," 9th Annu. Conf. Long Isl. Syst. Appl. Technol. LISAT 2013, no. September 2013, 2013.

Studying the Influence of Static Converters' Current Harmonics on a PEM Fuel Cell using Bond Graph Modeling Technique

Wafa BEN SALEM¹, Housseem CHAOUALI², Dhia MZOUGHI and Abdelkader MAMI
UR-LAPER, UR17ES11, Faculty of Sciences of Tunis, University of Tunis El Manar, 2092 Tunis, Tunisia

Abstract—This paper shows the results of adding static converters (Boost, Buck and Buck-Boost converters) as an adaptation solution between a PEM Fuel Cell generator and a resistive load in order to study different effects of the converter on the generator performances in terms of voltage and current behavior. The presented results are obtained by simulating the Bond Graph developed model under 20-Sim software and show current and voltage behaviors with each converter under different scenarios of working conditions.

Keywords—Static converters; PEM Fuel Cell; boost; buck and buck-boost converters; bond graph; 20-Sim software

I. INTRODUCTION

In order to find a solution for the problem of energy production, new types of generators have been invented during these last years and are witnessing an important development and continuous evolution [1], [2].

Fuel Cells (FC) are basically used to convert hydrogen and oxygen into electricity. We can find various types of FCs nowadays such as Proton Exchange Membrane Fuel Cell (PEMFC) which are widely used in mobile applications such as electrical means of transportation [3], [4] but using these DC generators imposes the need of an adaptation method between the source and the load in order to get the best operating conditions of the system especially in the case of a variable power demand [5].

In order to study different electrical systems, several methods have been invented in order to model them. Among these approaches, we chose to use a graphical technique which is the Bond Graph (BG) modeling technique in order to simulate the dynamic behavior of the system [6], [7]. The benefit of this approach is that a complex electrical system for example can be designed as different subsystems, where each one presents one part of the real system, and then make the connection between the ports of these different subsystems in a form of bonds which describe the transfer of energy between the real parts of the system [8], [9].

In this context, this work proposes a study on current harmonics effects on a PEMFC when a static converter is used as an adaptation stage between it and the load. The most used static converters which are the Boost, Buck and Buck-Boost converters are studied. The Bond Graph approach is used to model the system and energy exchange between its different parts under the 20-Sim software which makes it possible to

use the developed model to carry out different investigative simulations.

This paper is organized as follows:

- In Section II, Bond Graph approach is firstly presented along with PEMFC and the main three topologies of static converters. BG model of each component is also presented.
- Section III presents the obtained results of studying the interactions of the three static converters and the PEMFC. The influence of current harmonics, caused by addition static converters, on a PEM Fuel Cell is investigated.

II. PRINCIPAL OF BOND GRAPH MODELING METHOD

A. General Over-view of Bond Graph Modeling Technique

The bond graph tool is a graphical language confirmed as a structured approach to the modeling and simulation of multidisciplinary systems. This language allows building models of dynamic physical systems [10].

The energy exchanges in the bond graphs are represented by an arrow describing the direction of the transfer between two elements. This transfer is characterized by two variables: an effort and a flow whose product is equal to the power exchanged [11].

Table 1 summarizes the nine elements involved in the BG representation which are: the active elements (sources of effort and flow Se and Sf) which provide power, the passive elements which transform the power supplied to them into dissipated energy in the form of heat Element R) or stored (element C and I), the junction elements (0 , 1 , TF , GY) which are conservative of power [12], [13].

TABLE. I. ELEMENTARY BRICKS OF A BOND GRAPH

Symbol	Elements	Equations without causality
R : r	Resistance, friction	$e \cdot rf = 0$
I : i	Inductance, inertia	$e \cdot idf/dt = 0$
C : e	Capacity	$f \cdot cde/dt = 0$
GY	Gyrator, MCC	$e1 = rf2, e2 = rf1$
TF	Transformer	$e1 = ne2, f2 = nf1$
Se	Effort Source	$e = \text{constant}$
Mse	Controlled Effort Source	$e = e(\text{input})$
Sf	Flow Source	$f = \text{constant}$
Msf	Controlled Flow Source	$f = f(\text{input})$

B. BG Model of PEMFC

1) Working Principle of the PEMFC

The principle of the PEM stack consists of the oxidation of a fuel at the anode in order to release protons. These protons are conveyed via a proton conductor (electrolyte) to the cathode where the synthesis of the water. Fig. 1 shows the block diagram of a PEMFC type generator [14].

The operation of a fuel cell is described by a chemical relationship. This chemical reaction is an electrochemical oxidoreduction.

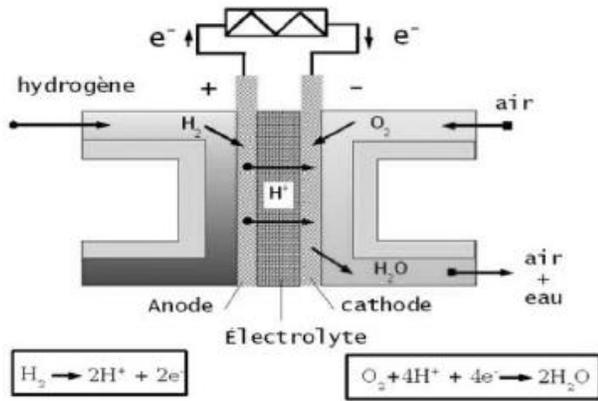


Fig. 1. Electrical presentation of a PV generator.

It reacts to hydrogen and oxygen to produce electricity, water and heat, depending on the overall chemical reaction given by (1).



This reaction takes place between the anode and the cathode which will be the electronic conductors, separated by a solid electrolyte which will be the proton conductor. More precisely, the reactions given by (2) and (3) occur at both electrodes.

At the anode: couple H⁺/H₂ (Acid electrolyte):



At the cathode: couple O₂/H₂O:



2) Bond Graph Model of PEMFC

Fig. 3 presents the BG model of the PEMFC which is composed of its three main parts: Anode, Electrolyte and Cathode.

Several information may be extracted from the V-I characteristic curve given in Fig. 2. First of all, the no-load voltage between is the voltage measured when the current is

zero, so this first analysis is carried out without power generation.

We can talk then about three distinct areas of operation as detailed in Fig. 5 [15].

- The first zone is for the low intensities, it represents the anode and cathode activation surge.
- The second zone, which is linear and has a large range of current variation, characterizes the Ohmic behavior of the cell, it is the zone most used in operation.
- The third area related to diffusion limitation. This last zone must not be used in operation because the flooding deteriorates the performance of the battery very strongly.

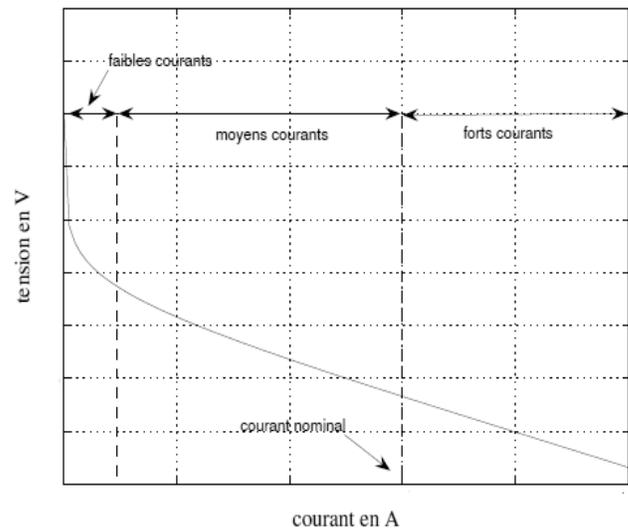


Fig. 2. Detailed V-I characteristic curve of the PEM fuel cell.

C. BG Model of static converters

1) Boost Converter

A boost converter allows increasing the output voltage supplied by the fuel cell by applying control signals to its parallel switch which is the basic component of the static converter [16].

For simulation in the 20-Sim environment, MTF represents a component that allows interrupting or authorizing the passage of a stream without obligation to distinguish what type it is.

Fig. 4 presents the BG model of the boost converter developed for simulation and Table 2 shows the parameters used in the developed BG model of the Boost Converter with causality assignment.

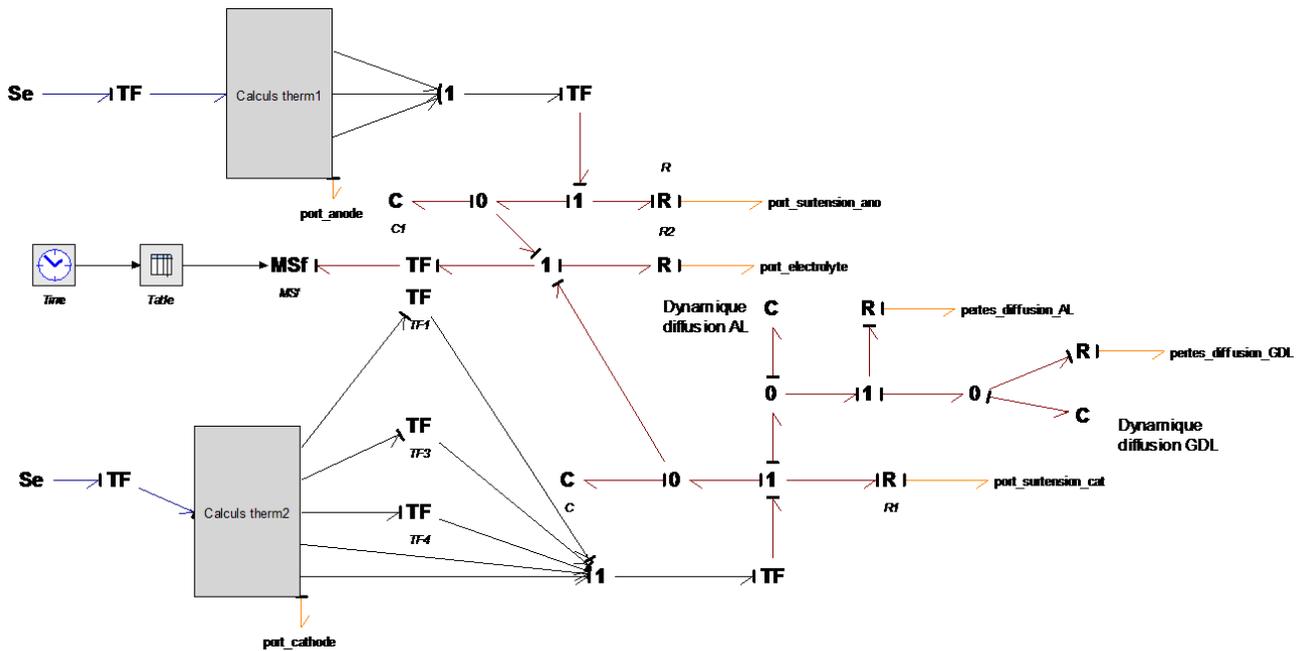


Fig. 3. BG model of the PEM fuel cell.

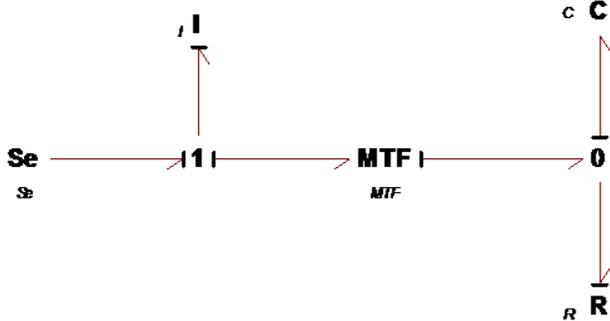


Fig. 4. BG model of the boost converter.

TABLE. II. PARAMETERS USED FOR THE BG MODEL OF BOOST CONVERTER

Parameters	Values
Resistance R	10 ohm
Capacitance C	0.002 F
Inductance L	0.000002 H
Voltage	1 V

The boost operation realized by the boost converter is given by (4).

$$\frac{V_0}{V_i} = \frac{1}{1-\alpha} \quad (4)$$

Such That:

V_o and V_i : output and input voltages

α being the duty cycle. It represents the duration of the period T during which the switch conducts. α is between 0 and 1.

In 20-Sim simulation environment, we fixed the duty

cycle $\alpha = 0.5$ and $Se = 1V$. The obtained curves of the input and the output voltages of the converter in Fig. 5 show that steady state is reached at $t=0.002s$ and the output voltage is stable around 2V.

In (4), we can see that the output voltage is always higher than the input voltage (the output voltage increases with α , and that theoretically it can be infinite when α approaches the value '1'. That's why we talk about boosters.

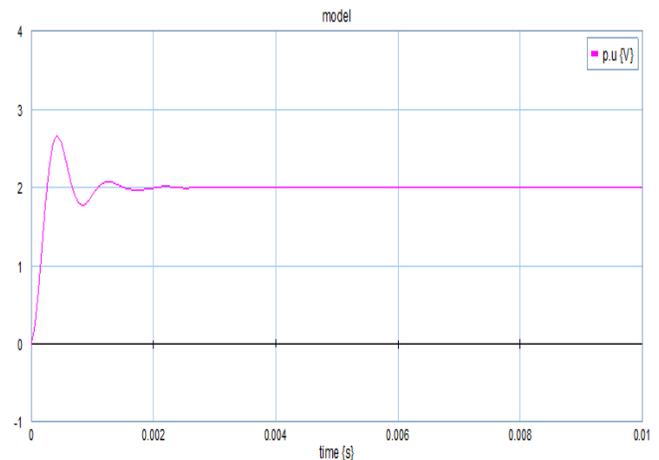


Fig. 5. Voltage at the terminals of the boost converter.

When a Boost converter operates in continuous conduction mode, the current flowing through the inductance never vanishes.

The characteristic describes the same evolution and the same behavior as that obtained by [17] as shown in Fig. 6.

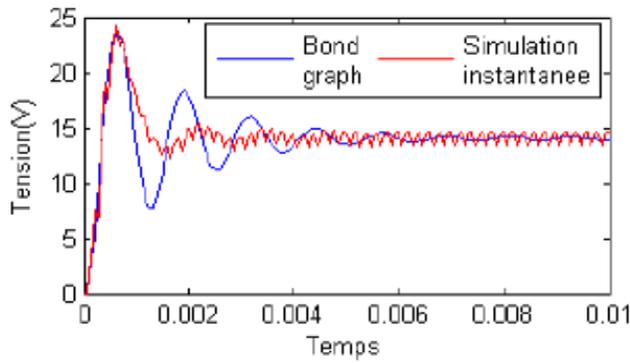


Fig. 6. Comparing BG simulation with real time simulation.

2) Buck Converter

A buck converter is used to reduce the voltage supplied by the fuel cell by applying control signals to its series electronic switch [16].

Its BG model with causality assignment that we developed is given in Fig. 7 where the different parameters that we used are presented in Table 3.

TABLE. III. PARAMETERS USED FOR THE BG MODEL OF BUCK CONVERTER

Parametres	Values
Resistance R	1 ohm
Resistance R1	1 ohm
Resistance R2	0.01 ohm
Resistance R3	1 ohm
Capacitance C	0.00010 F
Inductance L	0.000432 H
Voltage	10 V

In continuous conduction, we the output voltage and current of a Buck converter are given by (5).

$$\frac{V_o}{V_i} = \frac{I_i}{I_o} = \alpha \quad (5)$$

V_o et V_i : output and input Voltages of the converter.

I_o and I_i : output and input currents of the converter.

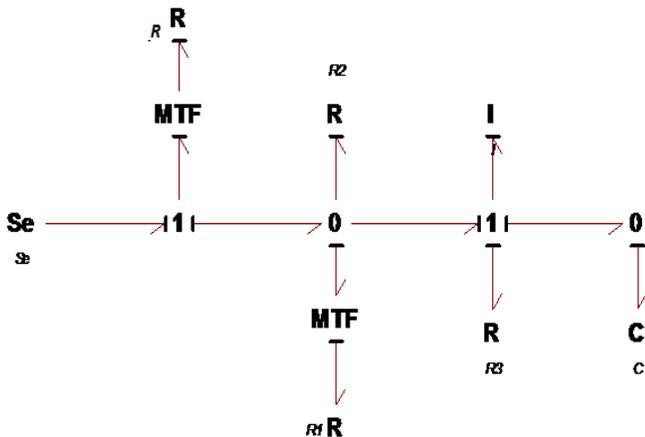


Fig. 7. Bond graph of the buck converter.

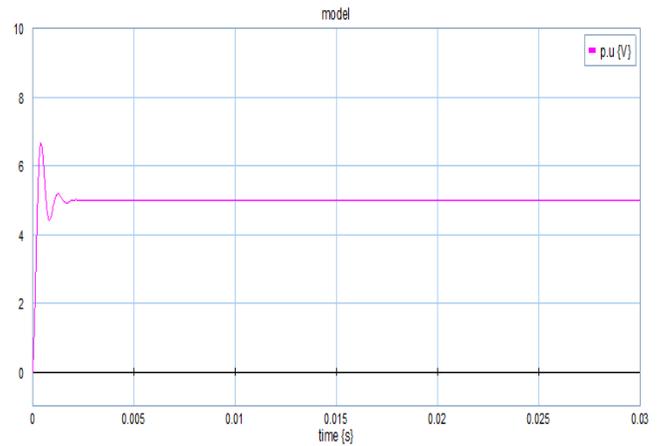


Fig. 8. Tension curve as a function of time of the buck converter.

Fig. 8 shows a simulation with $Se = 10V$ while fixing a cyclic ratio $\alpha = 0.5$. The voltage curve is at the input and the output of the Buck Converter as a function of time. It appears that the steady state is reached after 0.002s, the voltage stabilizes around the value of 5V at the output of the converter.

Back to (5), we can see that the output voltage varies linearly with the duty cycle. Since the duty cycle is between 0 and 1, the output voltage V_o is always lower than the input voltage. It is for this reason that we speak of a Buck Converter.

The results describe the same evolution and the same behavior as obtained in [17] as shown in Fig. 9.

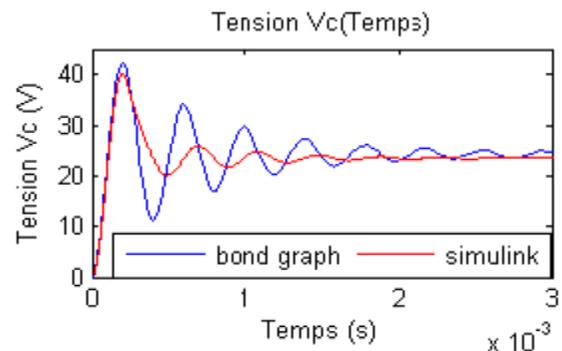


Fig. 9. Comparing BG simulation with real time simulation.

3) Buck-Boost Converter

A Buck-Boost converter is a switched-mode power supply that converts a DC voltage to a DC voltage of lower or higher value but of reverse polarity [18].

Compared to the Buck and Boost converters, the main differences are:

- The output voltage is reverse polarity to the input voltage.
- The output voltage can vary from 0 to $-\infty$ (for an ideal converter).

Fig. 10 presents the BG model used in our simulation.

The output voltage is described by (6):

$$\frac{V_o}{V_i} = -\frac{\alpha}{1-\alpha} \quad (6)$$

V_i and V_o : input and output Voltages.

Different types of conversion can be made by the Buck-Boost converter based on the value of α :

- For $0 < \alpha < 0.5$: The static converter is Buck and the output voltage is presented in Fig. 11.
- For $\alpha = 0.5$: The input voltage equal to the output voltage and is presented in Fig. 12.
- For $0.5 < \alpha < 1$: The static converter is Boost and the output voltage is presented in Fig. 13.

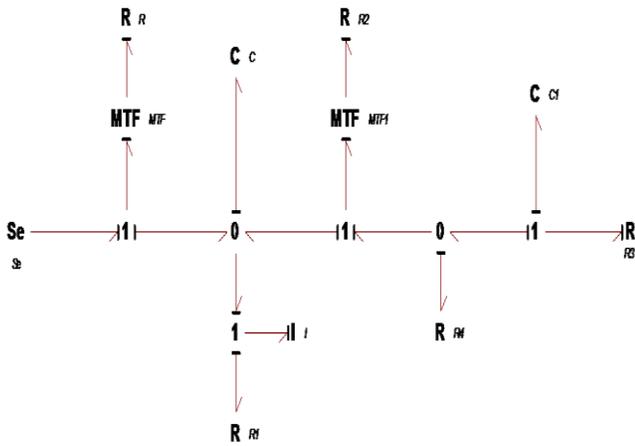


Fig. 10. BG model of a buck-boost converter.

Based on (6), we can see that the output voltage of a buck-boost converter is always negative and that its absolute value increases with α to ∞ when α approaches 1. The sign (-) allows the chopper function as a buck-booster. A disadvantage of this converter is that its switch does not have a terminal connected to zero, thus complicating its control.

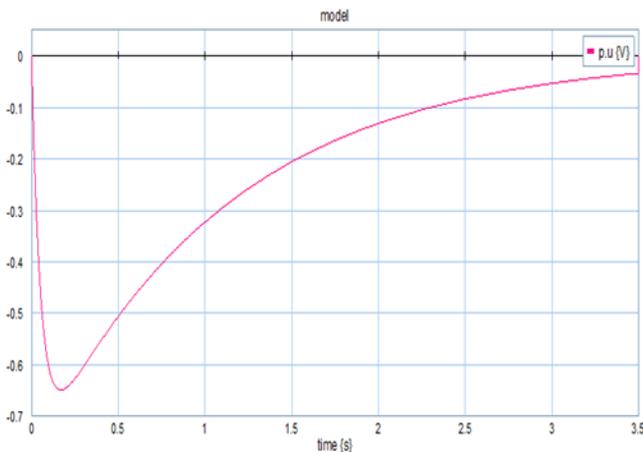


Fig. 11. Output voltage in Buck status ($0 < \alpha < 0.5$).

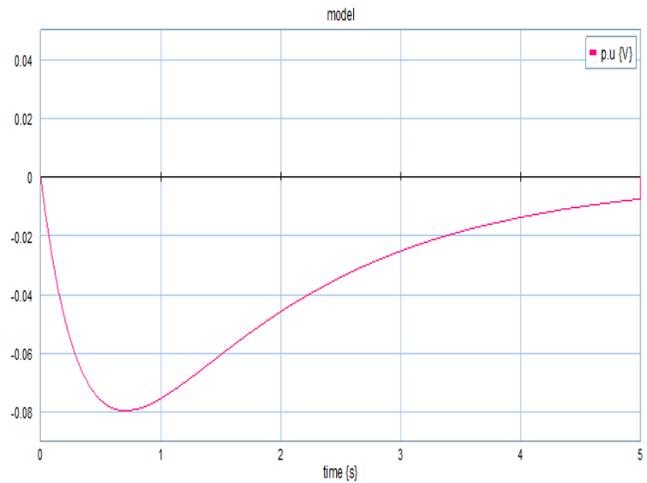


Fig. 12. Output voltage for $\alpha = 0.5\#$.

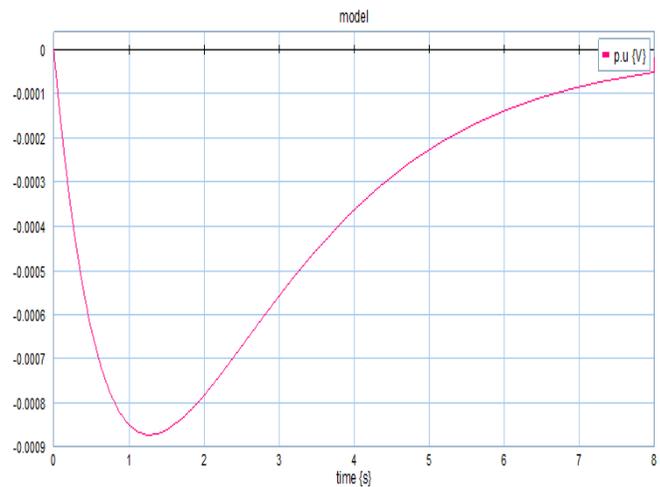


Fig. 13. Output voltage in boost operation ($0.5 < \alpha < 1$).

III. INTERACTION BETWEEN PEMFC AND STATIC CONVERTERS

The tests were performed at a cutting frequency of 20 kHz and any increase in frequency leads to a decrease in the speed of the simulation response, even though the curve pattern remains the same (delayed system).

A. Interaction Between PEMFC and Boost Converter

1) BG Model used for Simulation

In order to study the direct connection of a static converter to a PEMFC we have carried out perturbation studies by varying the value of α to keep the output voltage constant and to eliminate current ripples which are parasites which have a direct influence on the efficiency and electrical performance of the fuel cell, which can cause aging (limited life).

Fig. 14 presents the global BG model of the PEMFC and the boost converter used for the next simulations.

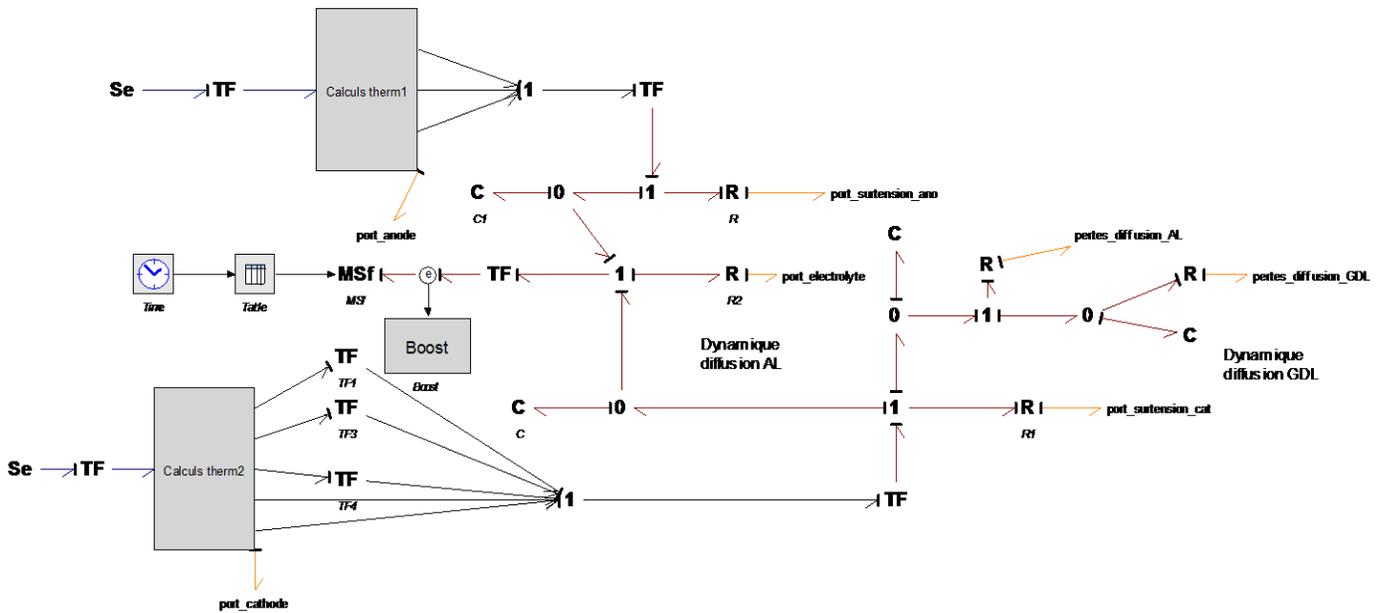


Fig. 14. BG Model of the boost converter connected to the PEMFC.

2) Simulation Results and Discussion

Fig. 15 and 16 present consecutively the output voltage and current obtained by simulating the PEMFC with the boost converter for duty cycle α at 0.5 and Se at 40V.

We notice that, at start, the voltage increases up to 110V before reaching its steady state around 80V at $t=0.002s$.

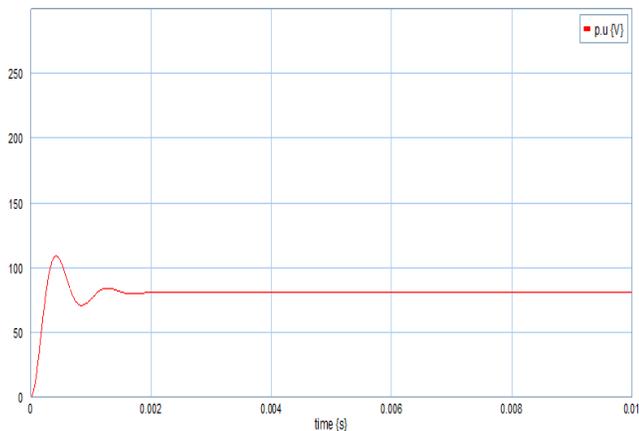


Fig. 15. Tension curve as a function of time of the PEMFC unit and boost converter ($\alpha = 0.5$).

By studying the current curve in Fig. 16, we note that the current is higher with respect to the voltage with fluctuations which negatively influences the efficiency of the cell.

This increase in current causes a local increase in temperature which can degrade the membrane [19], [20].

Fig. 17 and 18 present consecutively the output voltage and current obtained by simulating the PEMFC with the boost converter for duty cycle α at 0.8 and Se at 40V.

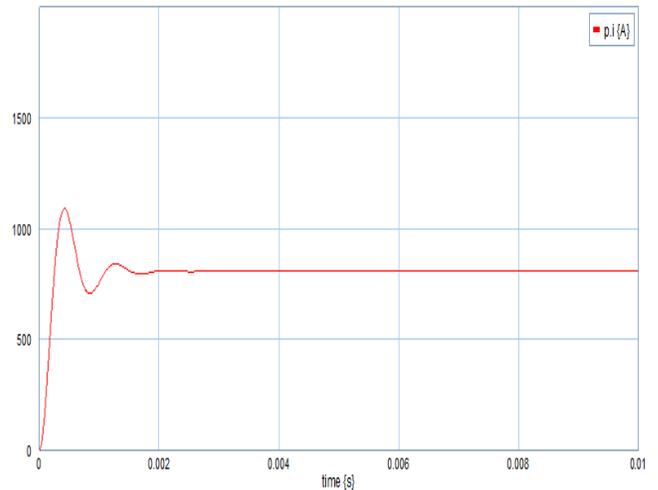


Fig. 16. Current curve as a function of time of the PEMFC unit and boost converter ($\alpha = 0.5$).

Fig. 23 shows that the obtained current when $\alpha = 0.8$ is higher than the obtained current with $\alpha = 0.5$. We also notice the disappear of fluctuations in both voltage and current curves (a deterioration of the disturbances) which means that increasing the value of α procures the elimination of current undulations.

Based on the works presented in [19]-[21], this important increase in current, in order to reach stabilization at high values, causes an increase in water generation at the cathode which is called the Diffusion Phenomenon. This phenomenon might cause a high level of water at both anode and cathode sides if the evacuation process of the water is not properly managed, this will eventually cause the so-called flooding.

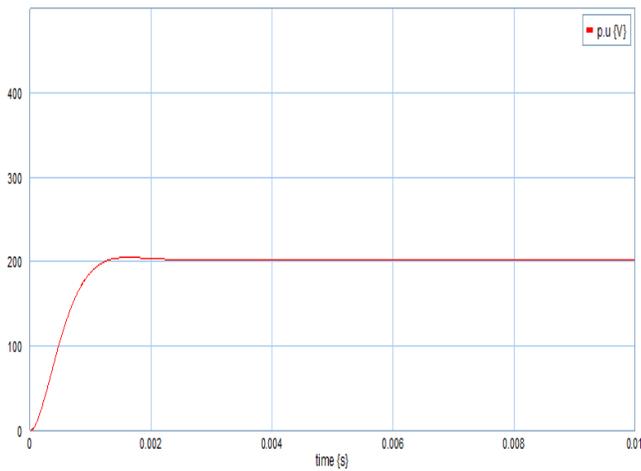


Fig. 17. Output voltage curve $\alpha = 0.8$.

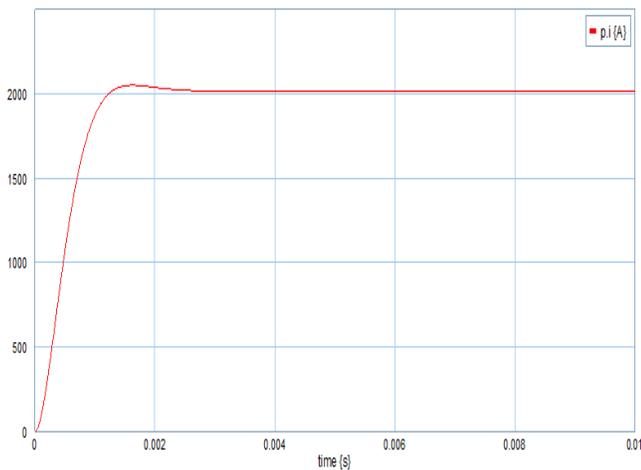


Fig. 18. Output current curve for $\alpha = 0.8$.

We conclude that the stack's operating temperature has an important influence on flooding conditions. Thus, permanent degradation of the stack performance can be resulted by the influence of bad water management.

Note: It is the double layer capacitor which fixes the dynamics of the diffusion phenomenon in the activation layer [22].

In the case of a boost Converter, according to fonts [22] and as shown in Fig. 19, the losses of the membrane are proportional to the current ripple Δi . The ripple of 200% of the average value increases by 33% of the losses in the membrane. On the other hand, a 20% ripple increases only 0.33% of the losses in the membrane.

B. Interaction Between PEMFC and Buck Converter

1) BG Model used for Simulation

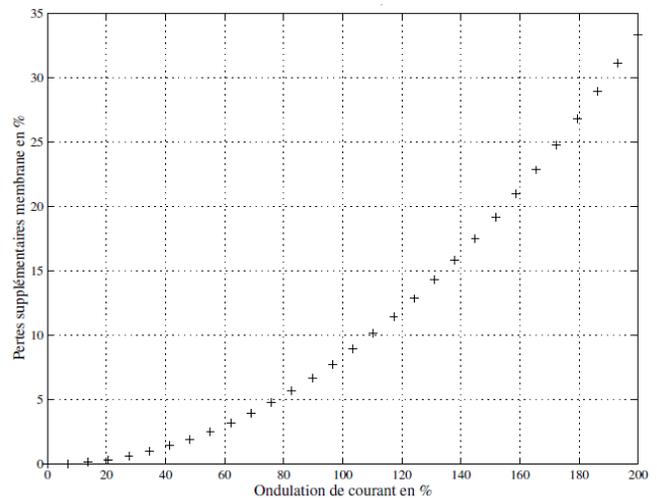


Fig. 19. Additional losses in the membrane as a function of the amplitude of the current ripple.

Fig. 20 presents the global BG model of the PEMFC connected to the Buck converter used for the next simulations to investigate the interconnections between them for different values of α .

2) Simulation results and discussion

Fig. 21 and 22 present consecutively the output voltage and current obtained by simulating the PEMFC with the Buck converter for duty cycle α at 0.2 and S_e at 40V.

The simulation of the PEMFC system with the Buck Converter shows an increase, without any fluctuations or disturbances, in voltage value to reach the 20v steady state.

The role of the diode can be demonstrated in this simulation by the sudden decrease of the current (drop of current) after the important increase in the beginning of the simulation.

- From 0 to αT : the current increases through the switch S by closing it.
- The voltage source is the origin of energy during this phase.
- From αT to T: the current decreases through the diode D by opening the switch S.

This is the “freewheel” phase due to the freewheeling diode. The purpose of these diodes is to prevent the occurrence of surges and sudden variations in intensity (essentially at break) [23].

Note: Current ripple Δi decreases when L increases (smoothing).

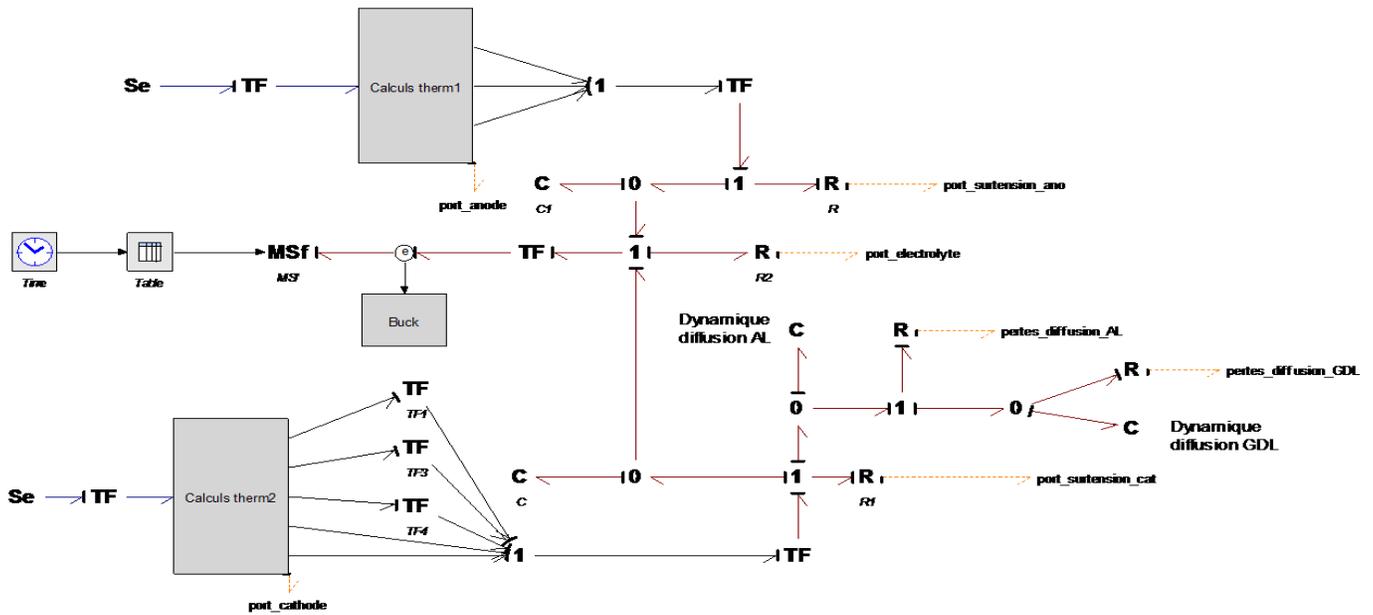


Fig. 20. BG model used for simulation of PEMFC connected to Buck converter.

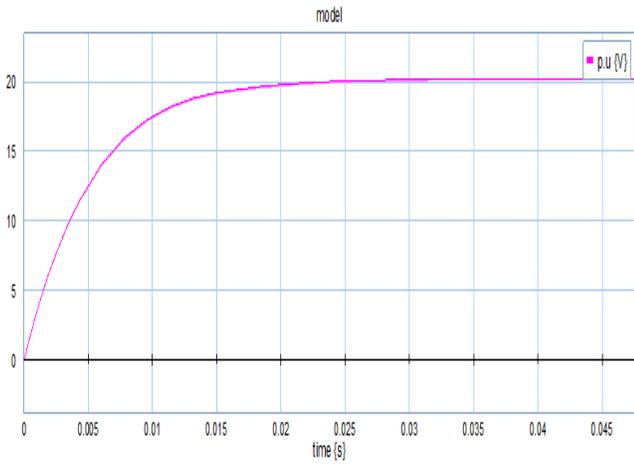


Fig. 21. Output voltage curve for $\alpha = 0.2$.

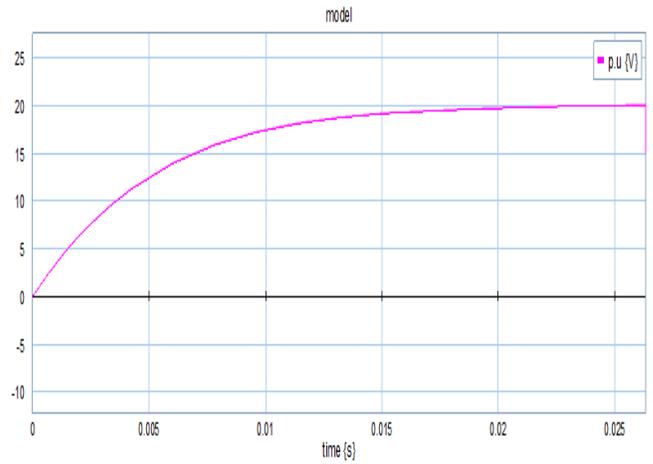


Fig. 23. Output voltage curve for $\alpha = 0.8$.

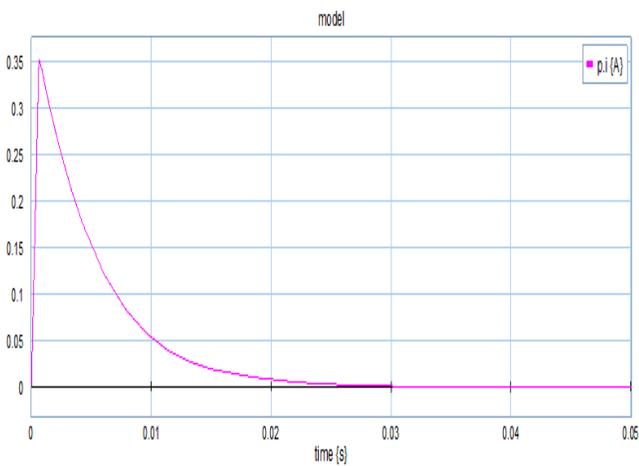


Fig. 22. Output current curve for $\alpha = 0.2$.

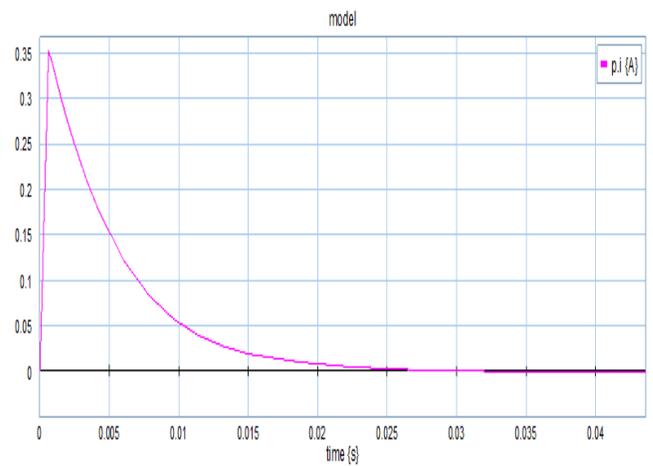


Fig. 24. Output current curve for $\alpha = 0.8$.

Fig. 23 and 24 present consecutively the output voltage and current obtained by simulating the PEMFC with the Buck converter for duty cycle α at 0.8 and Se at 40V.

The obtained results at both values of α , 0.2 and 0.8, are similar which proves the filtering of high frequency harmonics has been efficiently done by the double layer capacitor.

The generated undulations by the buck converter are more restrictive than those generated by the boost one.

Among the high frequency corrugations, those generated by The Buck converters are more constraining than those generated by the boost converters. This proves that the Buck converter generates more losses, shown in Fig. 25, in the PEMFC performance more than the Boost converter as a consequence of the additional generated ripples [22].

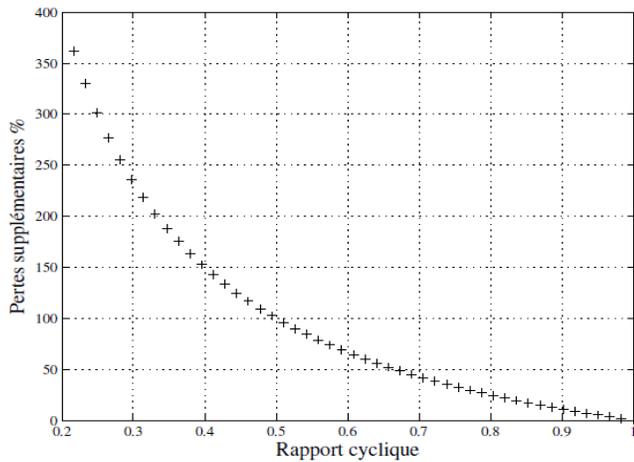


Fig. 25. Additional losses in the membrane as a function of duty cycle.

We can ask the question about the aging of the battery and the membrane which results in a reduction in electrical performance.

C. Interaction Between PEMFC and Buck-Boost Converter

1) Operating as Buck Converter

Fig. 26 and 27 present the curves of voltage and current at the output of the buck-boost converter obtained by imposing a value of $\alpha = 0.1$.

We notice a decrease in current to -35A and in voltage to -35V, the steady state is reached after 2s.

Fig. 28 and 29 present the curves of voltage and current at the output of the buck-boost converter obtained by imposing a value of $\alpha = 0.4$.

By increasing the value of α , the current and voltage decrease compared to their values for $\alpha = 0.1$. We can see that the undulations of the current and the voltage decrease when α increases.

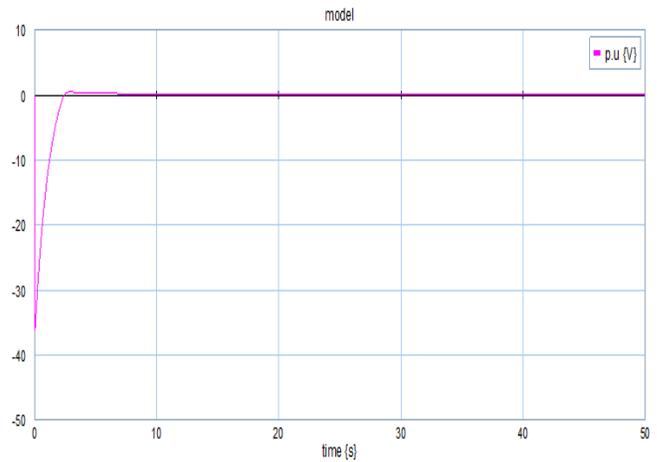


Fig. 26. Voltage curve for $\alpha = 0.1$.

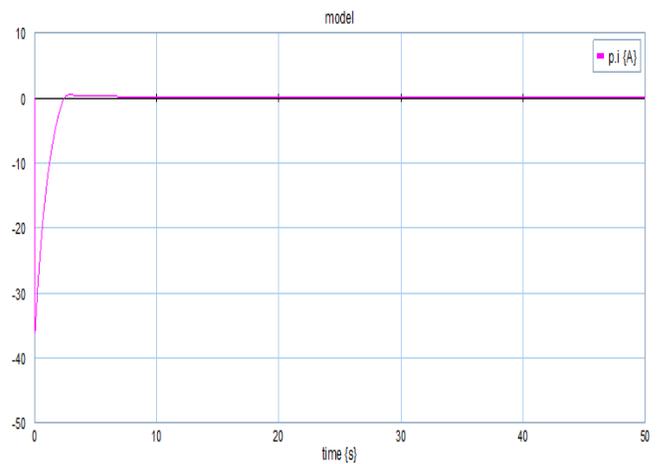


Fig. 27. Current curve for $\alpha = 0.1$.

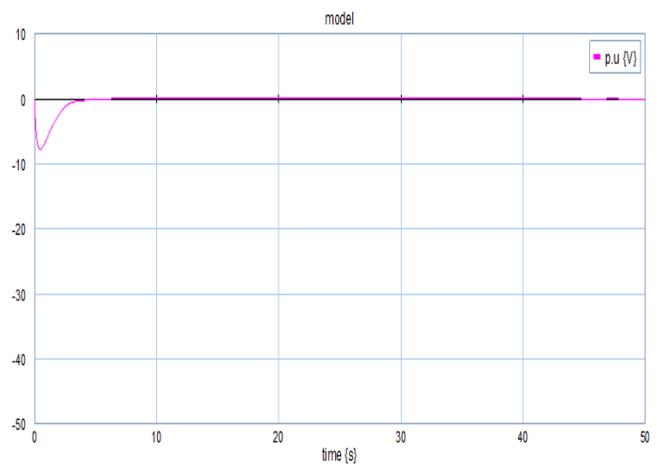


Fig. 28. Voltage curve for $\alpha = 0.4$.

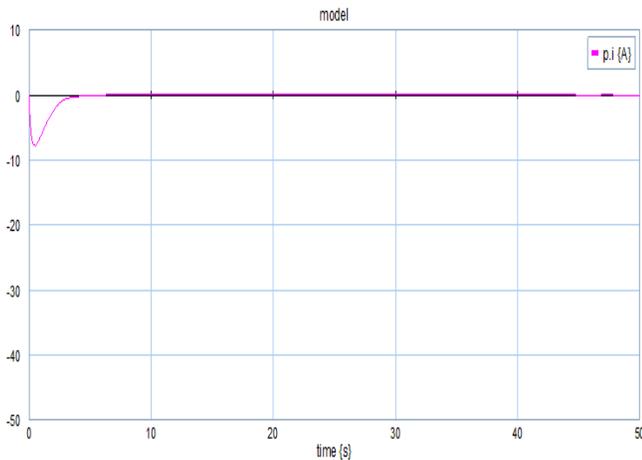


Fig. 29. Current curve for $\alpha = 0.4$.

2) Operating as Boost converter

Fig. 30 present the curve of current at the output of the buck-boost converter obtained by imposing a value of $\alpha = 0.6$ and Fig. 31 and 32 present, respectively the curves of voltage and current at the output of the buck-boost converter obtained by imposing a value of $\alpha = 0.8$.

After the simulation, we note from Fig. 31 and 32 that the output current and the voltage are zero, so the Fuel Cell and the Buck-Boost Converter system is no longer functional from $\alpha = 0.8$.

We have noticed from the preceding figures that each time when we simulate the Fuel Cell system with the Buck-Boost Converter, the voltage and the current abruptly decrease at the instant '0' and then increase and return to its state of "Balance" due to its switch which does not have a terminal connected to the zero, thus complicating its control.

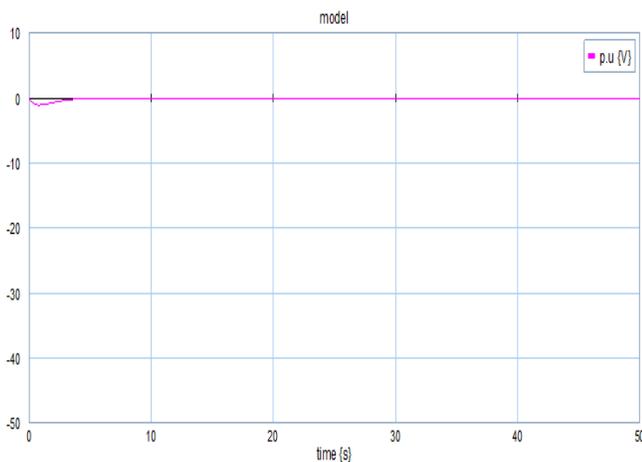


Fig. 30. Current curve for $\alpha = 0.6$.

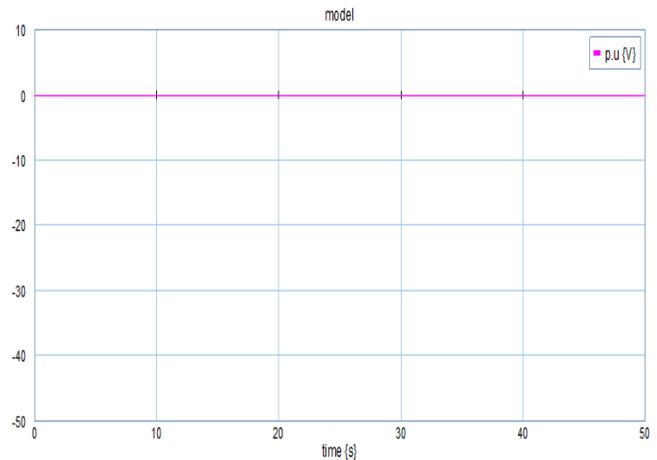


Fig. 31. Voltage curve for $\alpha = 0.8$.

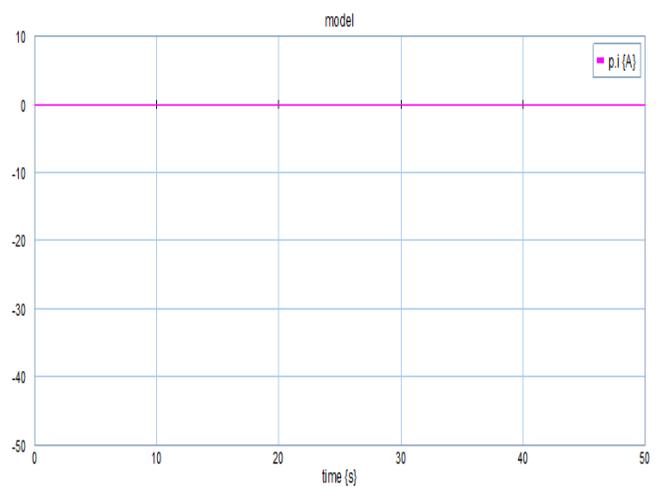


Fig. 32. Current curve for $\alpha = 0.8$.

IV. CONCLUSION

The obtained results showed that using a Buck converter shows important disturbances and undulations in the output voltage and current than the other two converters. Also, a Buck-Boost converter makes the control of the PEMFC more complicated. With using a Boost converter, we are risking to have the diffusion phenomenon if the water evacuation process is not well done.

As future work, we would be interested in investigating different interactions between the PEMFC and more advanced topologies of DC-DC converters such as the Sepic converter which presents the possibility of both buck and boost operations like the Buck-Boost converter but with the advantage of keeping the same polarity in the output.

REFERENCES

- [1] BENCHOUIA Nedjem Eddine, HADJADJ AOUL Elias, LAKHDAR Khochemane and BOUZIANE Mahmah, "Bond graph modeling approach development for fuel cell PEMFC systems", *International Journal of Hydrogen Energy*, 39(27), pp. 15224-15231, 2014. <https://doi.org/10.1016/j.ijhydene.2014.05.034>.
- [2] A.R. Maher, AL-BAGHDADIL Sadiq, "Modelling of Proton Exchange Membrane Fuel Cell Performance Based on Semi-Empirical Equations", *Renewable Energy*, 30(10), pp. 1587 - 1599, 2005. <https://doi.org/10.1016/j.renene.2004.11.015>
- [3] CHAN C. C. , BOUSCAYROL Alain and CHEN Keyu, "Electric, Hybrid, and Fuel-Cell Vehicles: Architectures and Modeling", *IEEE Transactions on Vehicular Technology*, 59(2), pp. 589-598, 2010. <http://dx.doi.org/10.1109/TVT.2009.2033605>
- [4] ANDARI Wahib, GHOZZI Samir, ALLAGUI Hatem and MAMI Abdelkader, "Design, Modeling and Energy Management of a PEM Fuel Cell / Supercapacitor Hybrid Vehicle", *International Journal of Advanced Computer Science and Applications(IJACSA)*, 8(1), 2017. <http://dx.doi.org/10.14569/IJACSA.2017.080135>
- [5] CHAOUALI Houssem ; OTHMANI Hichem; MEZGHANI Dhafer and MAMI Abdelkader , "Enhancing classic IFOC with Fuzzy Logic technique for speed control of a 3~ Ebara Pra-50 moto-pump", *IEEE 17th International Conference on Sciences and Techniques of Automatic Control and Computer Engineering(STA)*, Tunisia 2016. <http://dx.doi.org/10.1109/STA.2016.7951985>
- [6] X. Roboam, and G. Gandanegara, "Causal Bond Graph of Unbalanced Multi-phase Electrical Systems". *International Conference on Integrated Modeling & Analysis in Applied Control & Automation(IMAACA)*, Italy 2004.
- [7] MEZGHANI Dhafer, OTHMANI Hichem, SASSI Fares, MAMI Abdelkader and DAUPHIN-TANGUY Geneviève, "A New Optimum Frequency Controller of Hybrid Pumping System: Bond Graph Modeling-Simulation and Practice with ARDUINO Board" *International Journal of Advanced Computer Science and Applications(IJACSA)*, 8(1), 2017. <http://dx.doi.org/10.14569/IJACSA.2017.080112>
- [8] SUEUR C., "Bond-graph approach for structural analysis of MIMO linear systems", *Journal of the Franklin Institute*, 328(1), pp. 55-70, 1991. [https://doi.org/10.1016/0016-0032\(91\)90006-0](https://doi.org/10.1016/0016-0032(91)90006-0)
- [9] FILIPPA M., MI C., SHEN J. and STEVENSON R., "Modeling of a hybrid electric vehicle test cell using bond graphs", *IEEE Transactions on Vehicular Technology*, 54(3), pp. 837-845, 2005. <http://dx.doi.org/10.1109/TVT.2005.847226>.
- [10] WENZHONG GAO David, MI Chris, EMADI Ali, "Modeling and Simulation of Electric and Hybrid Vehicles", *Proceedings of the IEEE*, 95(4), pp. 729-745, 2007. <http://dx.doi.org/10.1109/JPROC.2006.890127>
- [11] SAISSET Rémi , FONTES Guillaume, TURPIN Christophe, ASTIER Stéphan, "Bond Graph model of a PEM fuel cell", *Journal of Power Sources*, 156(1), pp. 100-107, 2006. <https://doi.org/10.1016/j.jpowsour.2005.08.040>
- [12] J. GRANDA Jose, "The role of bond graph modeling and simulation in mechatronics systems: An integrated software tool: CAMP-G, MATLAB-SIMULINK, *Mechatronics*", 12(9-10), pp. 1271-1295, 2002. [https://doi.org/10.1016/S0957-4158\(02\)00029-6](https://doi.org/10.1016/S0957-4158(02)00029-6)
- [13] BELHADJ J., "Modeling, Control and Analyze of Multi-Machine Drive Systems using Bond Graph Technique", *Journal of Electrical Systems(JES)*, 2(1), pp.29-51, 2006.
- [14] PERAZA César, GREGORIO DIAZ Jose, J. ARTEAGA-BRAVO Francisco, VILLANUEVA Carlos and GONZALEZ-LONGATT Francisco, "Modeling and Simulation of PEM Fuel Cell with Bond Graph and 20sim", *IEEE American Control Conference(ACC)*, USA 2008. <http://dx.doi.org/10.1109/ACC.2008.4587303>
- [15] BEN SALEM W., MZOUGH D., ALLAGUI H. and MAMI A., "The bond graphs to the study of interactions between the PEM fuel cell and static converters", *IEEE 17th International Conference on Sciences and Techniques of Automatic Control and Computer Engineering(STA)*, Tunisia 2016. <http://dx.doi.org/10.1109/STA.2016.7952004>
- [16] Mohamed Akram JABALLAH, Dhafer MEZGHANI and Abdelkader MAMI, "Design and Simulation of Robust Controllers for Power Electronic Converters used in New Energy Architecture for a (PVG)/(WTG) Hybrid System" *International Journal of Advanced Computer Science and Applications(IJACSA)*, 8(5), 2017. <http://dx.doi.org/10.14569/IJACSA.2017.080531>
- [17] BOUDJEMAË Mehimedetsi, "Application du Formalisme Bond Graph à une chaîne de conversion d'Energie Photovoltaïque", Ph.D Thesis, Ecole Doctorale: Sciences pour l'ingénieur et microtechniques, 2006.
- [18] ALAM Aftab, TAO Lei and HABIB Kashif, "Optimal model predictive control for disturbance rejection and stability in buck-boost converter and its comparison with classical technique", *IEEE International Conference on Power System Technology (POWERCON)*, 2016 Australia. <http://dx.doi.org/10.1109/POWERCON.2016.7753946>
- [19] E. FRAPPÉ, "Architecture de convertisseur statique tolérante aux pannes pour générateur pile à combustible modulaire de puissance-traction 30Kw". Ph.D Thesis, L'Institut français des sciences et technologies des transports, de l'aménagement et des réseaux (IFSTTAR-LTN) Et du Laboratoire de génie électrique de Paris (LGEP), 2012.
- [20] N. Yousfi-Steiner, Ph. Mocoteguy, D. Candusso, D. Hissel, A. Hernandez, A. Aslanides, "A review on PEM voltage degradation associated with water management: Impacts, influent factors and characterization", *Journal of Power Sources*, 183(1), pp. 260-274, 2008. <https://doi.org/10.1016/j.jpowsour.2008.04.037>
- [21] Thibaut Colinart, Dylan Lelièvre, Patrick Glouannec, «Influence de l'hystérésis sur le comportement hygrothermique d'un enduit de finition intérieure biosourcé », *Conférence IBPSA*, France 2014.
- [22] FONTÉS Guillaum, "Modélisation et caractérisation de la pile PEM pour l'étude des interactions avec les convertisseurs statiques", Ph.D Thesis, l'ENSEEIH UMR, 2005.
- [23] Zongqi Hu, Dongsheng Ma, "A pseudo-CCM buck converter with freewheel switching control", *IEEE International Symposium on Circuits and Systems(ISCAS)*, Japan 2005. <http://dx.doi.org/10.1109/ISCAS.2005.1465279>

The Effect of Religious Beliefs, Participation and Values on Corruption: Survey Evidence from Iraq

Marwah Abdulkareem Mahmood Zuhaira
School of Management, Harbin Institute of Technology
Harbin, China

Tian Ye-zhuang
School of Management, Harbin Institute of Technology
Harbin, China

Abstract—This research tests the role that religious beliefs, rituals and values plays on the corruption in Iraq. Furthermore, the research assesses ethical and moral ideals pertinent to religion, in the Iraqi educational sector. Correlation analysis and linear regression help assess the relations among the study's constructs and variables. The hypotheses tested by multiple regression technique with the help of SPSS software. Grounded in the data collected from 600 employees, the results affirm that religious beliefs have negative association with levels of corruption. Prayers in religious institution are influenced by the clergy, which serves as a set of life instructions to avoid corrupt practices. The generalizability of our results might be limited because we surveyed workers from a single sector; this calls for future studies to verify the stability of our findings across another sectors and firms.

Keywords—Beliefs; participation; values; corruption

I. INTRODUCTION

Corruption is seen as one of the bad behaviors, in view of this, international organizations, governments and donors are looking for mechanisms to fight it. Corruption as an immoral behavior includes a deviation from the rules, laws and moral values. Corrupt actors benefit from the power entrusted to them for personal benefits [1]. The most famous definition of corruption is the use of public power for private gain [2]. Several attempts to fight corruption has always let to disappointing results, and this has prompted the United Nations Development Program (UNDP) to highlight on the social changes and behavior in this issue [3].

It has been discussed that in countries where religion plays role in the lives of most humans, many individuals, including workers, are likely to derive their ethical framework in part from their religion. Religion supplied many with the language of moral, ethics and often an actual 'list' of rules to live by, some of which can be explained as being of particular importance to fighting corruption. The basis for the growing attention given to the religion–corruption relation generally stems from the argument that fairness and sincerity form the basis of many religions, and as such, religious leaders can be utilized in the war against corruption [4], [5]. Developing countries are experiencing increased cases of using religion to curb corruption. Systemic studies on religious factors for individuals, Mutascu (2010) confirmed that religion significantly have affected corruption [6]. Lopez (2014) Proved religion employees negatively related to corruption levels [68]. Porta et al. (1999) [7] and Treisman (2000) [8] found religious mores of employees have been using cultural

attitudes towards social hierarchy. Wherever more hierarchical religions such as Islam dominate Catholicism, Eastern Orthodoxy, defies to office-holders might be scarce than in cultures shaped by more equality and individualistic religions, such as Protestantism. In addition, religions may explain how individuals seeing their loyalties to them organization contrary to other organizations. Through the historical pattern of influence that sophisticated in different settings between religious institutions and state, religion could affect corruption levels. In religious practice such as Protestantism, the religious institutions may play an important role in denouncing and monitoring abuses by state employees.

Faith in the afterlife reduces corruption. As a matter of fact, religious leaders speak publicly about suspicious practices within religious establishments when individuals participates in religious rituals [5]. Religious participation is working to support and strengthen the cohesion of confidence within the religious groups [9], as well as enable the individual to deal with negative and stressful events and turn the individual to pray to ease the psychological pressure [10]. Religious leaders often interact with individuals belonging to the same community, to provide strict guidance on behavior and lifestyles [9] as well as teaching them to prevent believers from participating in corrupt activities [11]. However, this can be influenced by cults. And as it turns out to be, cult is one of the factors that decreases corruption [5]. Also Marshall and Van Saanen (2007) added that, religious institutions, leaders and networks offer a powerful potential force in raising governance standards in the work of development. Because they have special "expertise" in values and integrity, and because of their extensive presence and reach [12]. From here we can see that the fundamental point is the possibility of the impact of religious participation on corruption [10].

Religious beliefs are linked to ethical abuses, especially in the area of rewards and punishment [9]. Religion promotes equality, condemns deception and rejects corruption [13]. Faith in afterlife is negatively associated with death anxiety, and small experience of life-threatening events is enough to push up an individual's faith [14]. While atheists believe that the individual goes to heaven after death and heaven differs from hell [15], individuals with moderate religious status have a higher death anxiety than the very religious individuals [16]. Although we are not certain about the existence of afterlife, religions working exceedingly to provide a set of beliefs about what happens after death and focus on individuals who are not committed to such beliefs, as such individuals will go to hell in the afterlife [15]. Thus we can ask the question: "Does

religious belief have a real impact on the levels of corruption”?

Values are a set of principles in the life of individuals [17]. Adherence to religious values require strength, organization, and control to achieve the goals [18]. Through religious values religions provide a set of guidelines on the violations that are forgiven and the ones that should not be forgiven [19]. Here, we must pause for a moment to ask if the values that can be forgiven are those values that encourage individuals to engage in corrupt practices. But [20] found that Muslims are less forgiving than Christians. And by referring to the Corruption Perceptions Index, we can confirm that Muslim countries are the most corrupt than Christians. Therefore, we can temporarily say that the values of forgiveness reduce the levels of corruption.

The study aims to investigate three points for organizational behavior literature. First, whether the increase in religious beliefs reduces corruption levels. Secondly, it aims at the investigation into the subject of an individual's participation in religious rituals and their effect on the transparency and if they discourage corruption level. Finally the values of forgiveness may have an unknown role in changing the levels of corruption. In this study, findings show that the increase in religious beliefs of the individual has increased the fear and provided a good deterrent to corrupt behavior. As long as the individual increases the prayer level, the probability of avoiding corrupt practices is high through the instructions and directives from the clergy group on the basis of the sacred texts.

II. METHODOLOGY

A. Main hypotheses

Several researches deal with the investigation into the issue of religion and corruption but reached contradictory results. Religion provides an internal barrier to move away from corrupt acts [21]. It also provides the basis for moral corruption that weakens [22]. The religious obligation to be more laws and more respected them [23]. In [24], most recent turning point is when the issue of corruption when he stressed the issue to focus on the religious beliefs and especially life after death. In addition, religious participation provides a range of life teachings that enable individuals stay away from corrupt behaviors [25]. Religious values (forgiveness) encourages individuals to stay away from corrupt practices [20]. Therefore, it is hypothesized that:

- 1) H1: Religious beliefs can decrease corruption levels negatively.
- 2) H2: Religious participants can decrease corruption levels negatively.
- 3) H3: Religious values can decrease corruption levels negatively.

B. Participant

In September 2016, we distributed questionnaires to workers by random-sampling in Iraqi universities, which are the religious areas of Islamic religion. The composition of the

participants was more than 18 years. The proportion of males was 65% while that of females was 35%. Using the Likert scale Quintet gradient of (1) strongly agree, and (5) strongly disagree.

C. Measures

1) Independent variables

Current study relies on the survey method to reach the maximum information in a limited time frame. The study used a Likert scale Quintet gradient of (1) strongly agree and (5) strongly disagree.

Religious beliefs were measured through three main dimensions and we set answer choices in a scale of 1-5, where: 1) always, 2) often, 3) sometimes, 4) rarely, 5) never. In the first dimension, faith in God, respondents were asked to determine the level of their faith, “I believe in God”. The second dimension, believe in afterlife scale [11], [9], included three items, such as, “I think that there is life after death”. The third dimension is death anxiety, which included 11 items derived from the scale [26] [11], for instance, “I am not sure what will be after death and I am very worried”.

Religious Participation measure in the fifth set of answers were ranked as follows: 1) Daily, 2) weekly, 3) monthly, 4) a few times a year, 5) never. It has been classified into public ritual consists of two items [27] focused on public religious participation in religious establishments. An example of such question is: “Except at funerals what is your presence in mosques and Shiite mosques rate these days?” Current study also have worked to put items in an attempt to measure the Individual rituals target denomination (Muslims) is composed of four items. Such as “I gave alms to the poor out of my own money”. The value scale was based on the scales of [28] and [14], an example is, “I think that corruption actions can be forgiven”.

2) Dependent variable

Four items in the scale of T. Stepurko et al. was adopted to examine corruption [29] to determine the amount of cash or in-kind gifts contained. For example, “Have you ever (or one of your family) paid cash amount on an informal basis to employees in government organizations?” The answer was determined by five choices 1) always, 2) often, 3) sometimes, 4) rarely, 5) never.

III. DATA ANALYSES AND DISCUSSION

We conducted multiple regression analysis to estimate the linear hypotheses to find out the role of religious beliefs, participation and values on corruption by SPSS 20 statistics.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar.

A. Descriptive Statistics

Descriptive statistics includes mean, standard deviation and bivariate correlation. (Beliefs, rituals and values) had a negative relationship with corruption (-0.64**, -0.61**, -0.47**) and ($p < 0.01$) (Table 1).

TABLE. I. MEANS, STANDARD DEVIATIONS, AND CORRELATIONS FOR ALL VARIABLES

Variable	M	SD	1	2	3	4	5
1. Age	0.60	0.48					
2. Gender	2.57	1.11	-0.07				
3. Belief	3.91	1.04	0.03	0.02			
4. Rituals	4.01	1.14	0.01	0.02	0.88**		
5. Values	3.98	0.94	0.01	0.01	0.77**	0.74**	
6. Corruption	1.96	1.09	0.01	0.06	0.64**	0.61**	0.47**

Note: N = 600. *p < 0.05; **p < 0.01

B. Multiple Regression Analysis: Direct Effects of Religious Beliefs, Participation and Values on the Corruption Level

The multiple regressions examined three models during the effect of religious variables on corruption. As expected, this study models have significantly negative effects on the corruption level, where F-value= 139.74, 122.11 and 59.19, p-value < 0.001, respectively. Model 1, 2 and 3, recorded an adjusted R2 of 41, 38 and 23 percent, respectively. In the first model, we estimated the effects of age, gender and employees beliefs on corruption. Religious Beliefs for employees had a negative and significant influence on corruption. It is equivalent to (-0.67**) for one per cent increase in corruption level. Age and gender had no influence on corruption. The impact of religious rituals for employees on corruption is negative and significant. It is equivalent to (-0.59**). Moreover, for one per cent increase in corruption level, we do not see any impact of gender and age in the second model. Finally, religious value for employees (forgiveness) is negatively associated with corruption (-0.56**) for one per cent increase in corruption level.

TABLE. II. SIGNIFICANT TESTING RESULTS OF THE MULTIPLE REGRESSION FOR MAIN-HYPOTHESES

Model				R ²	Adjusted R ²	F-value
1. Age, Gender and Beliefs (H1)(model 1)	0.05-0.905	0.04 0.155	-0.67** 0.000	.410	0.41	139.74
2. Age, Gender, and Rituals (H2)(model 2)	0.00-0.998	0.04 0.162	-0.59** 0.000	.380	0.38	122.11
3. Age, Gender, and Values (H3)(model 3)	0.006-0.945	0.06 0.085	-0.56** 0.000	.230	0.23	59.19

Note: *p < 0.05; **p < 0.01 standard error in parentheses, n = 600.

In the summary of results, religious beliefs, participation and values were negatively and significantly associated with corruption levels, but the effect of age and gender on the corruption level is insignificant (see Table 2).

The beliefs, rituals and values have beta β (-0.64, -0.61 and -0.48) respectively on corruption, and standard error (0.033, 0.063, and 0.059). The adjusted R2 is 0.41, which means that the beliefs explains 41 percent of the variation in corruption, where F- calculated is 139.74 higher than scheduled value. Rituals clarify 38% and values 23% of the variation in corruption. Where, F- calculated is 122.11 and 59.19, respectively higher than the scheduled value.

C. Discussion

The findings from current study show that beliefs, rituals and values have a relationship with corruption. In order to attain anti-corruption, one must resort to religious beliefs because it has a positive effect on corruption more than rituals. Religious values had a weaker influence on corruption in comparison with beliefs and rituals, but cannot be ignored to this effect.

The participation of individuals in religious rituals may contribute to the withdrawal of the individual from corrupt deals. The increase in religious beliefs may pose a barrier in the fear of doing suspicious business. [25] estimated the coefficient value of religion Sectarian participation on corruption equal -0.38%. [30] indicated that religion has a positive effect on happiness. They also argued that religiosity affected the sense of injustice as well as the individual's attitudes. Again, there are many reasons lurking behind corruption and injustice. [31] considered a culture of mistrust to reduce the level of transparency and increase corruption in institutions.

The current study ignored the feeling of happiness and its impact on corrupt deals, not to mention the differences in religious affiliation. This study reflects the perceptions of Muslims only and does not include other religions such as Christianity, Judaism, Buddhism and many others, so we deem it necessary to focus on a comparison between the study of religions and among different cultures in future studies. As well, the relationship between religious leaders and corruption is more problematic, and we encourage consideration of this question in future research.

IV. CONCLUSION

Corruption is far back in history and widespread in all countries, and all countries try as much as possible to avoid it. Through three independent variables beliefs, participation in rituals and values of religion we tested the possibility of reducing this type of behavior (corruption). We also found out that beliefs in afterlife is possible to contribute to the withdrawal of the individual from corrupt deals. Besides that religious participation would contribute to providing the foundations of guidelines as well acting as a mentor for individuals in daily life, especially when exposed to moral dilemma. However, the current study does not provide any clarification on the feeling of happiness, and the effect of an individual's sense of injustice in the field of corruption. The

results will be the best if it tested in different environments within different sectors.

ACKNOWLEDGMENT

I graciously thank Portia Opoku boad to help me in English editing. I also thank Karrar Abdulelah Azeez helped sustain my enthusiasm and energy throughout the project.

REFERENCES

- [1] T. Rabl and T. M. Kühlmann, "Understanding corruption in organizations—development and empirical assessment of an action model," *Journal of business ethics*, vol. 82, pp. 477-495, 2008.
- [2] J. C. Andvig, O.-H. Fjeldstad, I. Amundsen, T. Sissener, and T. Søreide, *Corruption. A review of contemporary research*: Chr. Michelsen Institute, 2001.
- [3] A. C. Gebel, "Human nature and morality in the anti-corruption discourse of transparency international," *Public Administration and Development*, vol. 32, pp. 109-128, 2012.
- [4] J. Luxmoore, "Churches urged to help fight global corruption," *Catholic New Times*, vol. 23, pp. 12-3, 1999.
- [5] H. Marquette, "'Finding God' or 'Moral Disengagement' in the Fight against Corruption in Developing Countries? Evidence from India and Nigeria," *Public Administration and Development*, vol. 32, pp. 11-26, 2012.
- [6] M. I. Mutascu, "Corruption, Social Welfare, Culture and Religion in European Union 27," *Transition Studies Review*, vol. 16, pp. 908-917, 2010.
- [7] R. La Porta, F. Lopez-de-Silanes, A. Shleifer, and R. Vishny, "The quality of government," *Journal of Law, Economics, and organization*, vol. 15, pp. 222-279, 1999.
- [8] D. Treisman, "The causes of corruption: a cross-national study," *Journal of public economics*, vol. 76, pp. 399-457, 2000.
- [9] Q. D. Atkinson and P. Bourrat, "Beliefs about God, the afterlife and morality support the role of supernatural policing in human cooperation," *Evolution and Human Behavior*, vol. 32, pp. 41-49, 2011.
- [10] A. Fenelon and S. Danielsen, "Leaving my religion: Understanding the relationship between religious disaffiliation, health, and well-being," *Social science research*, vol. 57, pp. 49-62, 2016.
- [11] J. Dezutter, B. Soenens, K. Luyckx, S. Bruyneel, M. Vansteenkiste, B. Duriez, et al., "The role of religion in death attitudes: Distinguishing between religious belief and style of processing religious contents," *Death studies*, vol. 33, pp. 73-92, 2008.
- [12] K. Marshall and M. B. Van Saanen, *Development and faith: where mind, heart, and soul work together*: World Bank Publications, 2007.
- [13] S. L. Adams, "The Justice Imperative in Scripture," *Interpretation*, vol. 69, pp. 399-414, 2015.
- [14] A. B. Cohen, J. D. Pierce, J. Chambers, R. Meade, B. J. Gorvine, and H. G. Koenig, "Intrinsic and extrinsic religiosity, belief in the afterlife, death anxiety, and life satisfaction in young Catholics and Protestants," *Journal of Research in Personality*, vol. 39, pp. 307-324, 2005.
- [15] D. Pyne, "An afterlife capital model of religious choice," *Journal of Economic Behavior & Organization*, vol. 92, pp. 32-44, 2013.
- [16] M. Ardel and C. S. Koenig, "The role of religion for hospice patients and relatively healthy older adults," *Research on Aging*, vol. 28, pp. 184-215, 2006.
- [17] S. Roccas and S. H. Schwartz, "Church-state relations and the association of religiosity with values: A study of Catholics in six countries," *Cross-Cultural Research*, vol. 31, pp. 356-375, 1997.
- [18] D. Mathras, A. H. Cohen, N. Mandel, and D. G. Mick, "The effects of religion on consumer behavior: A conceptual framework and research agenda," *Journal of Consumer Psychology*, April, 2016.
- [19] N. C. Scull, "Forgiveness, revenge, and adherence to Islam as moderators for psychological wellbeing and depression among survivors of the 1990 Iraqi invasion of Kuwait," *Journal of Muslim Mental Health*, vol. 9, 2015.
- [20] E. Mullet and F. Azar, "Apologies, repentance, and forgiveness: A Muslim-Christian comparison," *The International Journal for the Psychology of Religion*, vol. 19, pp. 275-285, 2009.
- [21] L. Shadabi, "The impact of religion on corruption," *The Journal of Business Inquiry*, vol. 12, pp. 102-117, 2013.
- [22] P. B.-N. Bloom and G. Arikan, "A two-edged sword: The differential effect of religious belief and religious social context on attitudes towards democracy," *Political Behavior*, vol. 34, pp. 249-276, 2012.
- [23] R. Gatti, S. Paternostro, and J. Rigolini, "Individual attitudes toward corruption: do social effects matter?," *World Bank Policy Research Working Paper*, 2003.
- [24] G. Gorer, "The pornography of death," *Encounter*, vol. 5, pp. 49-52, 1955.
- [25] M. Kalin and N. Siddiqui, "Islam's Political Disadvantage: Corruption and Religiosity in Quetta, Pakistan," *Politics and Religion*, vol. 9, pp. 456-480, 2016.
- [26] P. T. Wong, G. T. Reker, and G. Gesser, "Death Attitude Profile-Revised: A multidimensional measure of attitudes toward death," *Death anxiety handbook: Research, instrumentation, and application*, vol. 121, 1994.
- [27] B. Torgler, "The importance of faith: Tax morale and religiosity," *Journal of Economic Behavior & Organization*, vol. 61, pp. 81-109, 2006.
- [28] A. Macaskill, "Defining forgiveness: Christian clergy and general population perspectives," *Journal of personality*, vol. 73, pp. 1237-1266, 2005.
- [29] T. Stepurko, M. Pavlova, I. Gryga, and W. Groot, "Informal payments for health care services—Corruption or gratitude? A study on public attitudes, perceptions and opinions in six Central and Eastern European countries," *Communist and Post-Communist Studies*, vol. 46, pp. 419-431, 2013.
- [30] M. Joshanloo and D. Weijers, "Religiosity reduces the negative influence of injustice on subjective well-being: A study in 121 nations," *Applied Research in Quality of Life*, pp. 1-12, 2015.
- [31] J. Rowbottom, "CORRUPTION, TRANSPARENCY, AND REPUTATION: THE ROLE OF PUBLICITY IN REGULATING POLITICAL DONATIONS," *The Cambridge Law Journal*, vol. 75, pp. 398-425, 2016.

Detection of Distributed Denial of Service Attacks Using Artificial Neural Networks

Abdullah Aljumah

College of Computer Engineering and Sciences
Prince Sattam Bin Abdulaziz University, KSA

Abstract—Distributed Denial of Services (DDoS) is a ruthless attack that targets a node or a medium with its false packets to decline the network performance and its resources. Neural networks is a powerful tool to defend a network from this attack as in our proposed solution a mitigation process is invoked when an attack is detected by the detection system using the known patters which separate the legitimate traffic from malicious traffic that were given to artificial neural networks during its training process. In this research article, we have proposed a DDoS detection system using artificial neural networks that will flag (mark) malicious and genuine data traffic and will save network from losing performance. We have compared and evaluated our proposed system on the basis of precision, sensitivity and accuracy with the existing models of the related work.

Keywords— Distributed Denial of Services (DDoS); ANN; IDS

I. INTRODUCTION

The modern network world suffer due to security and threat vulnerabilities despite being from different origin or manufacturer or for different purpose and on the ground level, it is truly difficult technically and economically not feasible as far as both creating and maintaining such systems and to ensure that both the network and the associated systems are not susceptible to threats and attacks [1]. IDS is a special security tool that is being used by the network experts to keep the network safe and secure from network attacks which can

come from many different sources [2]. It has emerged as one of the basic and powerful tool in order to deal with data security and availability issues over the communication networks.

These attacks have a major influence of the networks and the systems as they include network performance, data security, loss of intellectual property [3] and a real liability for the compromised notes or networks data and that is why need a powerful IDS? Fig. 1 illustrates the architecture of IDS. The data packets received from the internet is forwarded to the processing unit where the format of the data is changed in order to make it compatible with the associated IDS and eventually the data packets are categorized as an attack or normal [4]. The normal data packet re allowed to pass through but the attack data packets as identified as attack type and are kept in the attack table and the alarm is raised and the defense procedure is invoked [5].

Large amounts of research have been conducted to improve IDS using artificial neural networks. The research proved that the network data traffic can be filtered and modeled more efficiently using artificial neural networks. Using artificial neural network proved itself as more advantageous as it take a thorough conscientious, perfect and accurate training, validation and top level testing phases before it is applied to the networks to detect malicious data and network attacks[6].

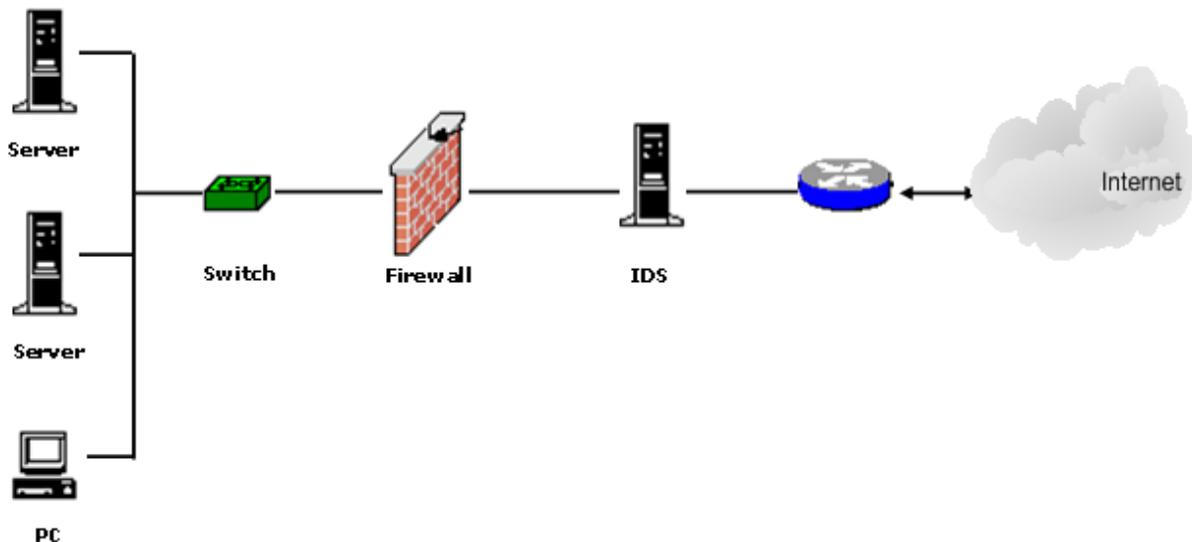


Fig. 1. Intrusion detection system.

II. ARTIFICIAL NEURAL NETWORK

Neural network (also known as artificial neural network) is an information processing model that is based and inspired from the human nervous system like the human brain does for humans [7]. The most important characteristic feature of this model is its unique structure of the system that processes the information. It consists of numerous exceptionally interconnected processing nodes (neurons) that work simultaneously to solve the specified problems [8]. Fig. 2 shows the real mathematical form of a neural network neuron. Neural networks, like humans do, learn by examples. Neural network is configured for a particular application, such as data classification or recognizing patterns through a learning process [9]. The learning process in humans requires synaptic connections adjustments between the neurons and same is the case with neural networks as well.

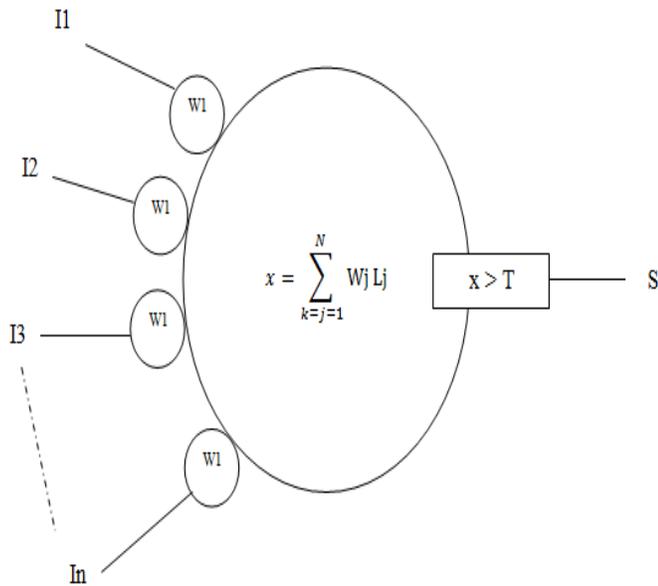


Fig. 2. Block diagram of an artificial neuron.

With the extra ordinary character of deriving meaning from complex and indefinite data, neural networks can be used to recognize and detect the patterns that are exceptionally complicated to be even observed or detected by humans and even by computer techniques [10]. After training process, a neural network can be treated as an expert one in the class or group information that has been given for analysis. This expert system can answer “what if” questions. There are other advantages of neural networks which include Adaptive learning, Self organization, Real time operation, redundant information coding, etc. [11]. Neural networks learn by examples and cannot be programmed to accomplish any specific job [12]. These examples need to be selected correctly and delicately otherwise the precious time of the system will get wasted or the network might work improperly.

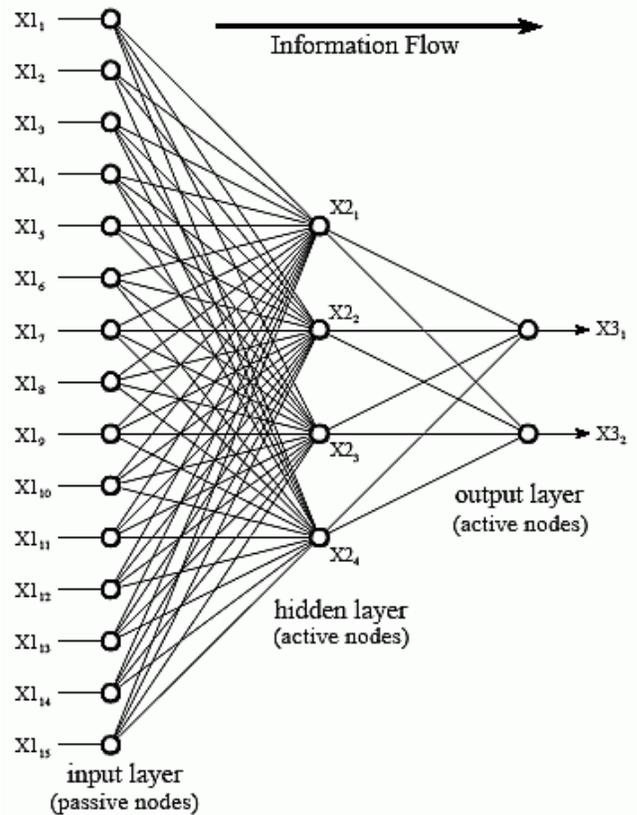


Fig. 3. Architecture of the neural network.

Neural network mainly have three categories of layers which include Input layer, Hidden Layers and output layers. Fig. 3 illustrates the basic architecture of the neural network. This is the most common architecture of neural networks. The input nodes are input nodes and rest of the nodes are active nodes. The input layer nodes are connected to hidden layer nodes and the hidden layer nodes are connected to output units. The action of this neural network is decided by the weight that is put on hidden layer nodes. The main job of the input nodes is to represent the raw information that is received by the network. This input and the weight on the connections between hidden nodes and input nodes decide the action of the hidden layer units. This action or activity of the hidden layer nodes and the weight between output layer nodes and the hidden layer nodes decide the performance and the behavior of the output layer nodes.

III. DDoS

Denial of Service (DoS) attacks is a deliberate, malicious, criminal attempt to deprive legitimate network users from using their network resources. DoS affect service providers in many aspects, most notably crippling availability of services provided by them. DDoS themselves are not powerful enough to bring down any web service in present computational resources scenario. A more sophisticated scalable and distributed attack evolved out of DoS is DDoS or Distributed

Denial of Services. It was first reported by Computer Incident Advisory Capability (CIAC) in somewhere around summers of 1999 [20]. Since then almost all DoS attacks were somehow of distributed characteristics.

To sabotage any website by DDoS there are broadly two methods, first and primitive one is to send packet with morped packed to confuse routing protocols also known as vulnerability attack [21]. Second and somewhat advance and more sophisticated mechanism involve attempts of either one or both of following (a) at network/transport layer attack flooding web server to exhaust bandwidth, router processing capability and hence paralyzing connectivity to the legitimate user [21]; (b) attack at application layer for depriving legitimate user with services by consuming server resources of provider website, e.g. sockets, memory, disk I/O, etc. [22].

Usually attacker seldom acts directly, rather a series of pre compromised nodes are chosen by him to launch attack on

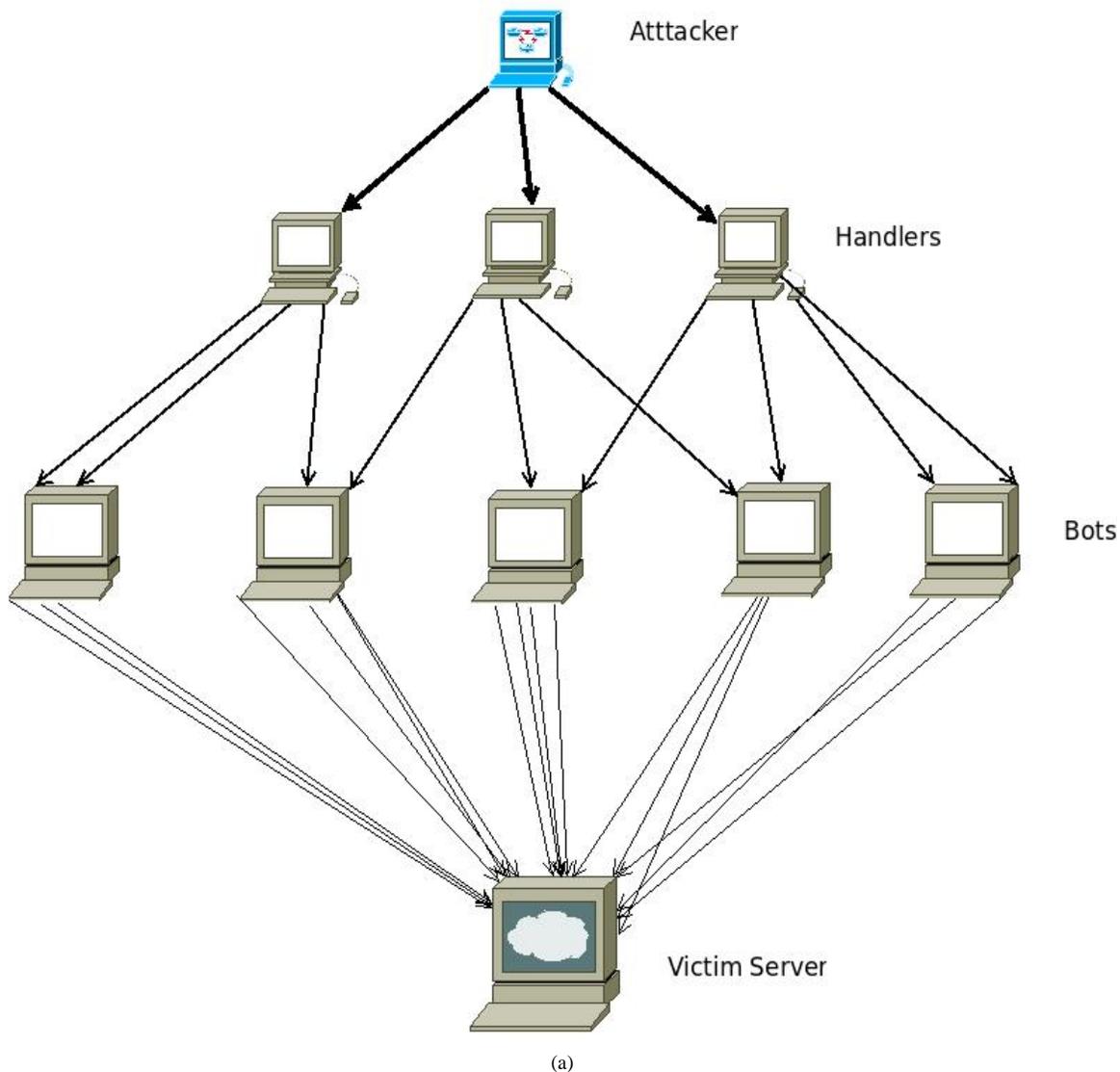
behalf of him, known as Botnets or simply Bots or Zombies (Fig. 4(a)). Attacker may have gain access to these computers by any means of infection [19].

A more recent trend is to magnify the amplitude of attack so as overwhelm victim even with enormous amount of resources, a way to get it is “DNS Amplification” (Fig. 4(b)).

A. Role of Amplifiers/Reflectors

DNS amplification is a phenomenon where a small query is amplified several folds as this amplified query with much larger payload than original one is then directed to victim server. Amplification of usually 70 folds is achieved easily [18].

DNS amplification a kind of reflective attack where spoofed IP of victim server is used for DNS query, in return victim server is flooded with large number of UDP packets.



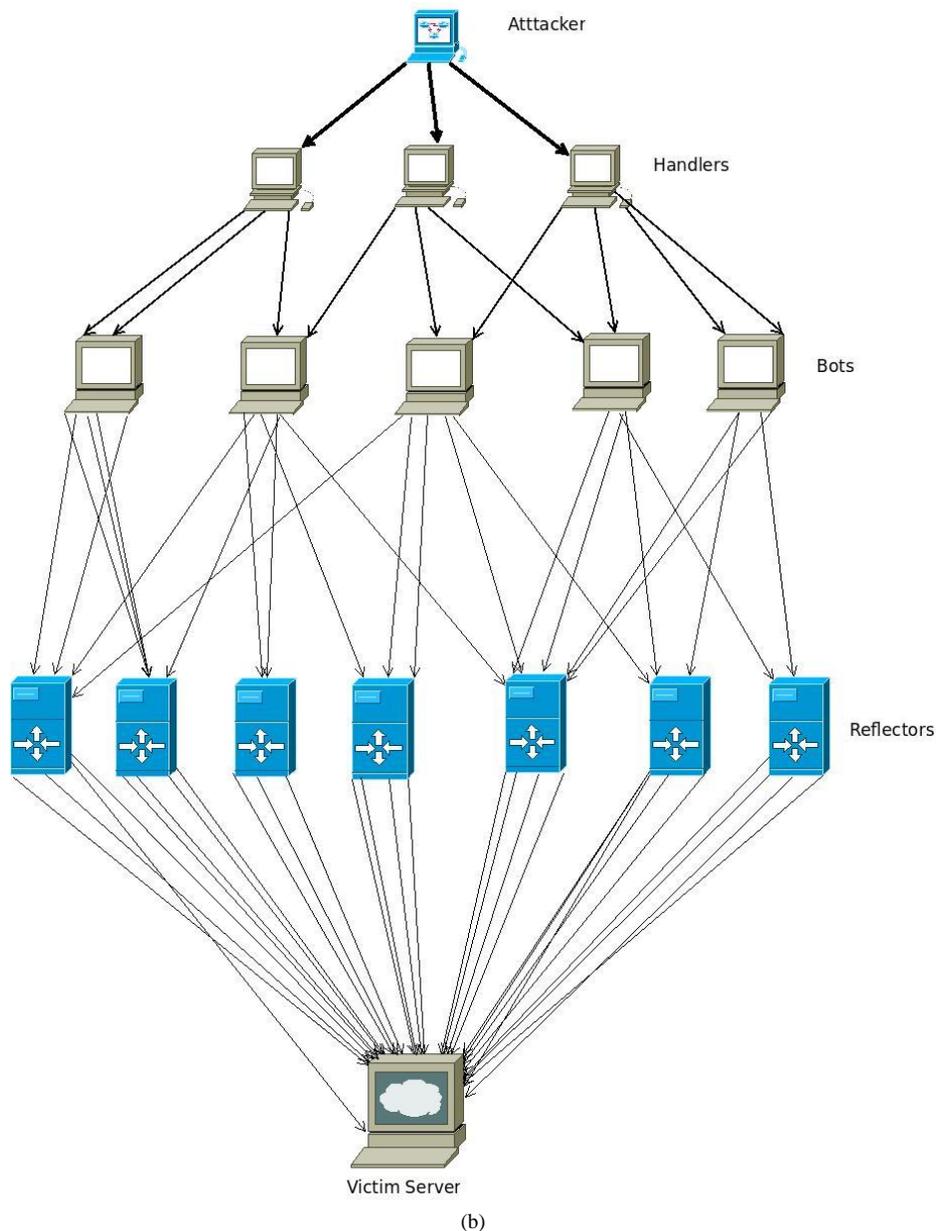


Fig. 4. (a) Direct DDoS attack; (b) Reflexive DDoS attack.

IV. CONSEQUENCES OF DDOS

Effects of DDoS attacks on business installation are immediately reflected as Revenue Losses, with loss rate going as high as \$ 300K/hour for service outage hours [13]. With advent of time, cost to mitigate DDoS attacks kept ever rising, in a survey by Forrester Research survey of Canadian decision-makers, DDoS attacks were declared most expensive with average cost associated with a typical DDoS reaching well beyond \$ 100,000 per security incident [14].

Besides being attacked is direct blow onto market reputation of any e-commerce website. In their findings, Bell Canada mentioned, 67% corporate say DDoS cause negative impact to customers, 56% say it critically impacts the brand name while 55% are concerned with negative effects on customer relations [15].

Al though, DDoS attacks are not meant for theft, but recently there has been shift in DDoS activities with stealing of user data, customers information, intellectual properties, etc. while enterprise resources were busy in mitigation of DDoS and related effects, known as Smoke-Screen effect. In the transitional time when IT experts of target organization are busy to bring back critical application. On line, attacker try to bypass security checks and get away with crucial business data, e.g. during DDoS attack on Carphone Warehouse, while internal team was busy with DDoS mitigation, hackers stole personal and banking details of 2.4 million people [16]. In their security report, Kaspersky Lab has published, 26% of DDoS attacks end up with Data Loss [17].

V. RELATED WORK

With the use of ANN for the detection of DDOS attacks by Jie-Hao and Ming [24] in which the results were compared with output and the decision tree, ANN, Bayesian and entropy. The researchers recognize the user demands for any particular resource on the involved system and their control data. Moreover, the samples of such identifications were sent to the attack detection system for any vulnerabilities.

Liu, Gu and et al. established a system called Learning Vector Quantization (LVQ) neural networks to identify attacks [25]. The technique is supervision type of quantization, which can be used for further procedures such as pattern recognition, data compression and multi-class classifications. Furthermore, the inputs were supplied to neural networks as data sets in the form of numerical calculations.

Akilandeswari and Shalinie [26], proposed a Probabilistic Neural Network Based Attack Traffic taxonomy in order to detect various DDOS attacks. In contrast, the authors mainly focused on distinguished between Flash Crowd Event from Denial of Service Attacks. Moreover, their work also involved the use of Bayes decision rule for Bayes interference coupled with Radial Basis Function Neural Network (RBFNN) for precisely classifying the DDOS attack traffic and the legitimate traffic.

Siaterlis & Maglaris [27] came up with a procedure of single network characteristics to mitigate the attacks. With the use of data fusion algorithm with Multi-layer Perceptron (MLP) in which the inputs were initialized from various non-active measurement which were available on the network, and hence the data combined with the traffic which were generated from the experimenters itself.

Joshi, Gupta and Misra [28] used a design consideration of neural network in order to detect zombie systems which were fueling the DDOS attacks. The main motive to their initiative was to figure out the connection between the zombie computer and sample entropy. The entire process workflow comprises on the predictions with the help of feed-forward neural network. Another objective for their research is to utilize the current infra for detecting and mitigating such attacks.

Badishi, Yachin & Keidar [29] used an approach of cryptography and authentication to defend DDOS attacks from affecting network resources and services. A very close approach proposed by Shi, Stoica and Anderson [30], However, DDOS attacks are detected using a different technique called puzzling mechanism.

Hwang and Ku [31] proposed a distributed technique to mitigate DDOS attacks. The mitigation system called Distributed Change-point Detection (DCD), which primarily reduces the risk of such attacks. The researcher suggests using non-parametric CUSUM (Cumulative Sum) algorithm to identify any major or minor variations in the network traffic. The team also focused on the initial source of the attack for detection.

A group of author [32]-[34] proposed a system of packet-marking and entropy in which each packet is marked on every router involved in communication in order to track the source

of the packet. However, a number of techniques proposed by some authors used ANN or infrastructure to defend against DDOS attacks, where as a couple of them identified the source of the attack. In contrast, none of them describes any unknown or zero day attacks labeled as high or low risk attacks. Hence, our main objective is to detect and mitigate unknown DDOS attacks and differentiate our proposed solution from the authors of [25]-[28].

VI. CONCEPTUAL FRAMEWORK

If deployed properly the DDOS detectors can minimize the strength of an attack. The DDOS detectors prevent the malicious packet from reaching the target after detection by analyzing the network for abnormal behavior or the abnormalities in the network. It is important for DDOS detectors to allow legitimate packets to pass through and reach the destination. So, it is extremely important for the detection system to be explicitly precise and checked against every possible and imaginable patterns and cases. Most commonly TCP, ICMP and UDP are used because of ease in practicality, implementation and documentation. The yearly report of Proplexic explained that these protocols are used by most attackers to launch most of the DDOS attacks. Since we have used ANN (artificial neural networks) for our detection mechanism where its precision predominantly depend on the quality of the algorithm training and the associated datasets and patterns used. The patterns include packet source address, sequence numbers and ID along with port numbers of source and destination, all these entities of packets are used for training the ANN. Based on our analysis and experimental verification, maximum number of zombies installed to oppose the operating system libraries in order to generate genuine packets that the installed zombie agents use their integrated built-in libraries. This is just to help the attackers in manipulation and forging the message throughout the attack.

Hence, it is easily possible to study the main properties of authentic packets that are created by authentic applications and can be easily compared with fake packets that are created by the attack tools and feed them as input patterns to train the artificial neural networks. We launched different kinds of DDOS attacks at distinct levels in order to select the different patterns for input to the artificial neural networks by creating an elite network infrastructure in unanimous and solitary environments. We studied the results very carefully and compared them with authentic traffic in order to verify the characteristic patterns that distinguish authentic traffic from the attack traffic. This segment of the process demanded thorough comprehension of how distinctive protocol interchange data or do the communications. The java neural network simulator accepts the authentic and malicious pattern in a specified format because the data sets are designed and assembled to accommodate both types of patterns. However 79% of the datasets are used in training the algo and 21% are used to ratify the process of learning. The input entities are normalized in order to increase the capability in delicate applications like the one we have where exact detection is extremely important otherwise if applied directly will lead to vanquish the impact of smaller values because normalization has positive effect on artificial neural network's training and performers.

A normal artificial neural network is made up of three layers i.e. input layer, an output layer, and a hidden layer, the datasets and patterns are given through input nodes for the learning process. These input attributes indicate the main pattern that distinguishes the genuine traffic from the attack traffic. Then we selected three different structures of topological artificial neural networks having three layers each i.e., input layer, output layer and hidden layer. But every topological artificial neural network structure will have different number of nodes as shown in Table 1.

TABLE I. NO. OF INPUT AND OUTPUT NODES FOR ICMP, TCP AND UDP

Topological ANN structure	No. of input nodes	No. of hidden nodes
ICMP	3	4
TCP	5	4
UDP	4	3

However the computation process deals with hidden nodes regarding input and output nodes. A single node is used as output layer to represent 1 or 0 for attack and normal traffic,

respectively. Fig. 5 displays the TCP topological artificial neural network structure, Fig. 6 displays ICMP topological artificial neural network structure and Fig. 7 displays the UDP topological artificial neural network structure. Selecting an appropriate learning algo, invoking function and number of hidden nodes where chosen on the early experiments where the accurate results were provided by Back Propagation and Sigmoid. Bidirectional associative memory, Elliot, Sigmoid and Softmax are used as functions while the comparison was between Quick-Prop, Back Propagation, Bidirectional Associative Memory, Back Prop Weight Decay, Back Prop thru time (16, 17, 18).

Our experiment shows 98.5% accuracy in selected topological structures when sigmoid invoking function is paired with Back Propagation as shown in Table 2.

TCP topological structure's input layers as shown in Fig. 4 is composed of five nodes with TCP sequence, source IP address, source port number, destination port number and flags.

ICMP topological structure is shown in Fig. 5 where ICMP ID and sequence number, source IP address are the input nodes.

TABLE II. COLLECTIVE RESULTS OF LEARNING ALGO, INVOKING FUNCTION, HL

Protocol	Learning Algorithm	Invoking Function	No. of Hidden Nodes	Detection Accuracy and CPU Usage	Best Results
TCP	Back Propagation	Sigmoid, Elliot, BAM, Softmax	One or more Hidden Nodes	98.6% and 66%-CPU Utilization	Best Recorded With 4 hidden nodes using Sigmoid.
UDP	Back Propagation	Sigmoid, Elliot, BAM, Softmax	One or more Hidden Nodes	98.6% and 69%-CPU Utilization	Best Recorded With 3 hidden nodes using Sigmoid.
ICMP	Back Propagation	Sigmoid, Elliot, BAM, Softmax	One or more Hidden Nodes	98.5% and 70%-CPU Utilization	Best Recorded With 4 hidden nodes using Sigmoid.

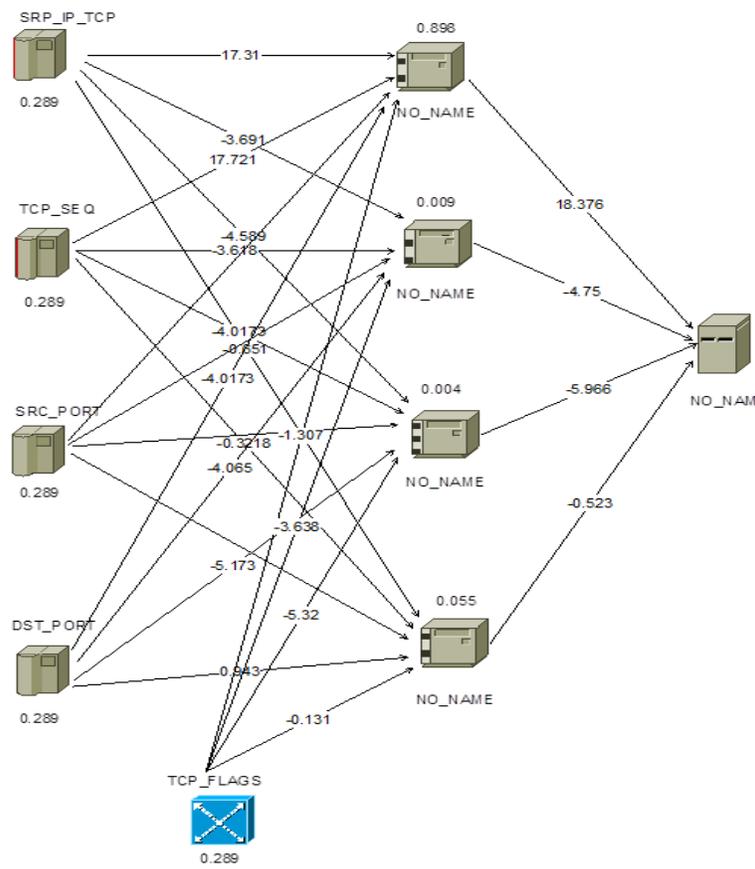


Fig. 5. ANN TCP topological structure.

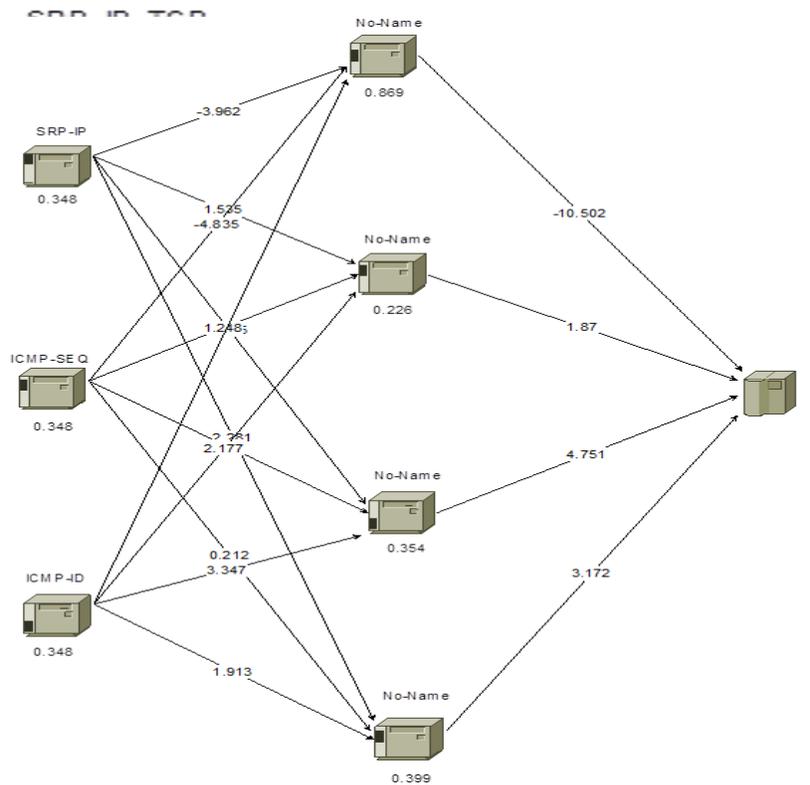


Fig. 6. ANN ICMP topological structure.

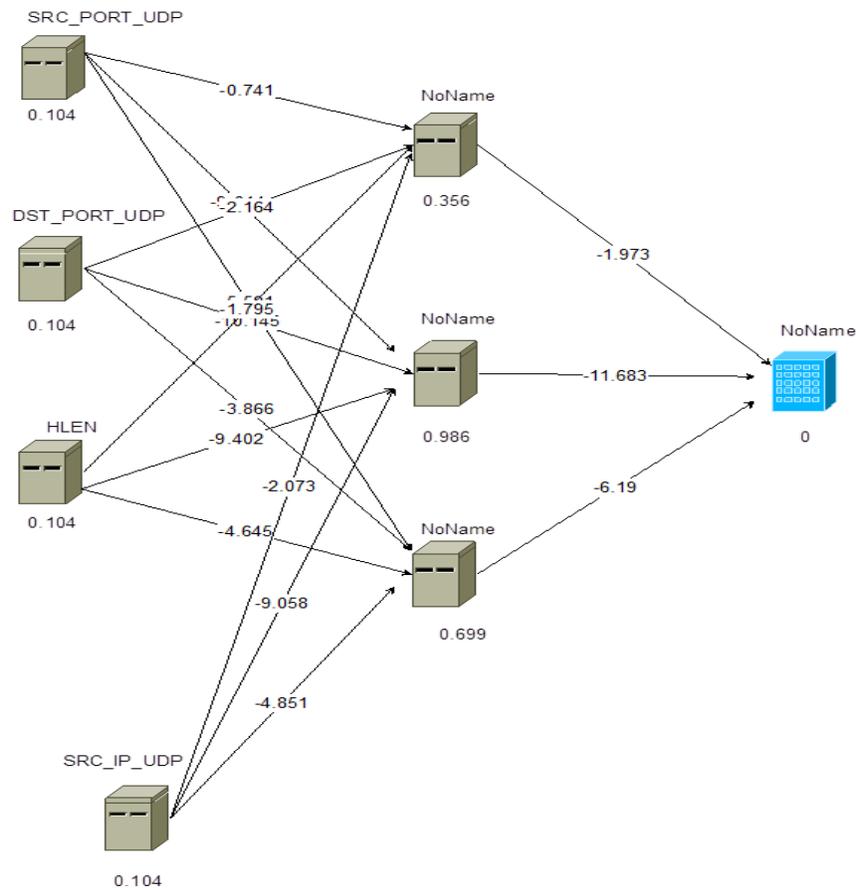


Fig. 7. ANN UDP topological structure.

UDP topological structure is shown in Fig. 6 where UDP source port, UDP destination port, Packet size and source IP address are the input nodes.

The supervised Back Propagation uses the weight that is represented by the numbers between the nodes to calibrate and learn by the patterns (examples). So if we provide more new pattern then it would be better in detecting the attacks. The algorithm keeps on changing the numbers between the nodes (Weight) till the desired result is obtained (having flag either 1 or 0). Fusing all the artificial neural network's as single application against instances can be deficient in availability if the system breaks down technically. Thus, if one instance is technically unavailable or down (for example an instance that detects TCP attack), the other two still will be present to detect TCP and ICMP attacks.

In the meantime, instigating artificial neural network instances separately for every protocol bestows improved maintenance, more control to analyze and to train the algo. The moment detection system detects the forged packets, the defense mechanism is invoked to allow the legitimate traffic go through and drop the forged traffic and as soon as the system flags the traffic as normal the system unblocks the flagged traffic. The legitimate traffic floating through the network and the system will not be interrupted because of being already flagged as legitimate traffic by our proposed system.

Besides the detection system provide the consciousness about attacks through communications via encrypted messages. This kind of information exchange between the detectors enhance the security system by identifying the malicious behavior and if required deploy countermeasures.

VII. DESIGN

We designed our solution to monitor the network continuously for malicious behavior by analyzing the header information of retrieved packets of the networks using trained artificial neural networks. Since retrieving a large amount of data in a network needs higher processing rate and is very expensive. Therefore, to overcome this for every protocol we used an individual packet threshold. If the amount of data packets in specific network is higher than the specified threshold of the protocol then the redeemed packets have to go through investigation. Based on our experiments, we selected the best threshold per protocol by counting the maximum number of data packets per unit time in selected distinctive environment where the true values of threshold are configurable. The amount of data packets are segregated and devised for examination, our proposed mechanism feeds those patterns into artificial neural network to decide the genuineness of the retrieved packets. One DDoS detection system is installed in every network to communicate through encrypted message with other DDoS detectors as shown in Fig. 8.

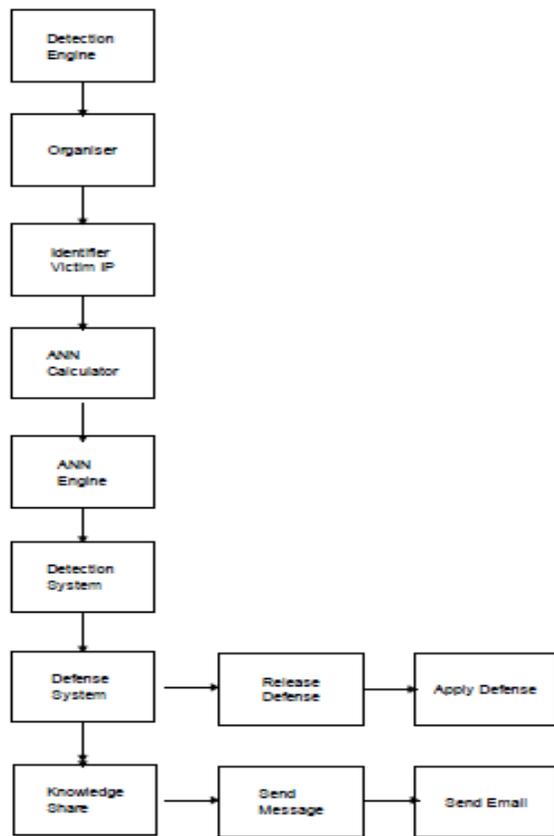


Fig. 8. Detection, defense and cooperative mechanism.

Following are the details of Fig. 7.

- 1) Install DDoS detectors on different networks.
- 2) Each DDoS detector will maintain the registered IP address of each hop DDoS detector in order to communicate through encrypted message whenever an attack is detected.
- 3) There should be continuous monitoring by DDoS detector for abnormal behavior or data.
- 4) Every passing packet is flagged as abnormal in case the value of passing packets is higher than the threshold.
- 5) If the value of passing traffic is higher than the threshold then:
 - a) The organizer removes the undesired characters and arranges the packets accordingly.
 - b) The victim IP addresses are identified by IP identifier.
 - c) The retried patterns are calculated by artificial neural network calculator and device them for artificial neural network engine.
 - d) The patterns are taken as input by artificial neural network engine and produce a single output i.e. 0 for normal and 1 for attack.
 - e) Step D is repeated three times to produce three outputs before the defense system is invoked.
 - 6) Then the detection system sends the output to the defense system and:

A.

Output	Action	Status
000	0	Traffic clear and allow traffic

B.

111	1	Traffic malicious allow only genuine traffic to pass through
110	1	Traffic malicious allow only genuine traffic to pass through
101	1	Traffic malicious allow only genuine traffic to pass through
011	1	Traffic malicious allow only genuine traffic to pass through

C.

100	0	Repeat point 5
010	0	Repeat point 5
001	0	Repeat point 5

If outcome from C is:

	Output	Action	status
A	111	1	Attack
	110	1	Attack
	101	1	Attack
	011	1	Attack
B	100	1	Low rate attack
	010	1	Low rate attack
	001	1	Low rate attack
C	000	0	No attack

D. However, if the outcome matches none of the above combination then a value 2 is generated by the system that means the traffic is unknown and is not used in the process of training artificial neural networks. In this scenario the system scans its local database to check if some data is received or detected by other hop DDoS detectors. If the neighbor DDoS detection systems respond with 0 or 1 then the algo is obsolete and outmoded as the algo detection was too. Thus proving that the local detector's algo needs and offline retraining with up to date patterns else no action is executed.

7) The knowledge share block communicates with all enrolled neighbor DDoS detectors by sending them encrypted message in cooperating protocol used, destination IP and type of attack. This information is also forwarded to security

offices by emails to let them know about these attacks for logistics purpose.

When we train the algo with old datasets the outcome of the detection system is two and artificial neural network has the special characteristics to detect the unknown pattern if the type of attack or attack itself is similar to the pattern that the algo was trained with. However the experimental results proved that if we train the system with old datasets then the algo fails to detect the unknown patterns. The experiments also proved the fact that the system can detect the known and the unknown attacks if we train the system with up to date patterns while the algo that is trained with old datasets failed in such scenarios. In this situation the artificial neural network of DDoS detection system (detector) that failed to detect attack while other neighboring DDoS detectors detect the same attack that was trained with old datasets previously must be trained with latest up to date datasets but offline because training process is supervised process and different patterns must be instigated or re-instigated whenever required. Thus, when the algo training is not up to date the extra assistance can be acquired from the share knowledge between the detectors to make further decisions. In the meantime every detector sends a complete email including full report of DDoS attacks acquired during that period to the security officers. One deployed detector may collect all the attacks and forward it as a single email to the security officer. However, no information will be sent to the security officer in case the deployed central point is down by any reason and consequently no more countermeasures are deployed if needed. All the DDoS detectors are devised to work and process as a standalone element or distributed detectors which communicates with other registered detectors through encrypted message within the networks or that are deployed in different networks.

Our solution is not confined to a least number of detectors to communicate through encrypted messages. Thus in case one DDoS detector stops functioning the other detectors deployed in the system can still send and receive messages therefore making the solution durable, reliable and resistant to DDoS detector collapse or crash.

To implement our designed solution, we have devised our detection module as plug-in and amalgamated it with Snort-AI (19). Snort AI is devised on Snort signature IDS project (20) and authors of this project are active in providing Snort AI plug-in and other amalgamation processes. The outcome of the IDS is combined with destination IP address to request iptables (21) to elevate malicious or fake packets while allowing legitimate data to pass through. In addition to this, we have also used RSA encryption technique for message

encryption over TCP connection while the deployed detectors act as sender and receiver both.

VIII. EVALUATION

We used precision, susceptibility – expertise to recognize positive results and specificity – expertise to recognize malicious results, to evaluate our solution. Table 2 represents the comparison of our results with other four approaches and a signature based solution for which quantitative assessments are recorded. We used legitimate and attack data traffic (high and low rate) to test our solution in an isolated and controlled network environment. During our experiments we launched 60 rounds of genuine traffic and 60 rounds of DDoS attacks (ICMP, UDP, TCP) involving 80 to 90 zombies to target the destination. We used VMware boxes to install the zombies and attack from the virtual platform where the boxes were connected to the target devices using virtual routers. We deployed the DDoS detectors between the victims and the virtual router where they examined the data traffic for irregularity and deformity.

Based on the results obtained from our experiments our solution provided a better result in terms of detection, precision, susceptibility and specificity as compared to other solutions including Snort as shown in Table 3 and Fig. 10 to 12, when all the tools were placed in the same manner and same DDoS attacks were launched in the same environment at the same time.

The author (Author Name) used probabilistic neural network over two periods and the accuracy was calculated up to 92% and 97% for attack and normal traffic, respectively. Author name (6) compared back propagation and learning vector quantization. Since our solution is based on back propagation we compared our solution to back propagation that stipulates better precision and performance. In [22] Leu and Pai used as statistical method while [23] Xu, Wei and Zang used KPCA and PSO-SUM to detect DDoS Attacks. KPCA (Kernel Principle Component Analysis) is used to eliminate unnecessary features and PSO (Particle Swarm Optimization) to optimize SVM (Support Vector Machine). During the experiments our solution provided 98% detection accuracy while the percentage of known and unknown attacks was 50% and 48%, respectively. We further evaluated our approach and during the evaluation against low and high rate DDoS attacks the detection results for low and high rates DDoS attacks were 98% and 97.4%, respectively as compared to 93% and 92% of Snort results. We also trained our solution with existing and latest dataset and deployed various known and unknown DDoS attacks. Table 4 and Fig. 9 represent the experimental results.

TABLE III. COMPARISON OF DIFFERENT APPROACHES WITH OUR APPROACH

Approach/Result %	Our Approach	Snort	PNN	BP	Chi-Square	K-PCA-PSO-SVM
Precision	98	93	92:97	90	94	NA
Susceptibility	96	90	NA	NA	92	96
Specificity	100	97	NA	NA	NA	NA

TABLE IV. RESULTS USING OLD AND UP-TO-DATE DATASETS

Our Solution	Accuracy	Susceptibility	Specificity	Precision
Old Datasets	92	88	96	92
Up-to-date Datasets	98	96	100	98

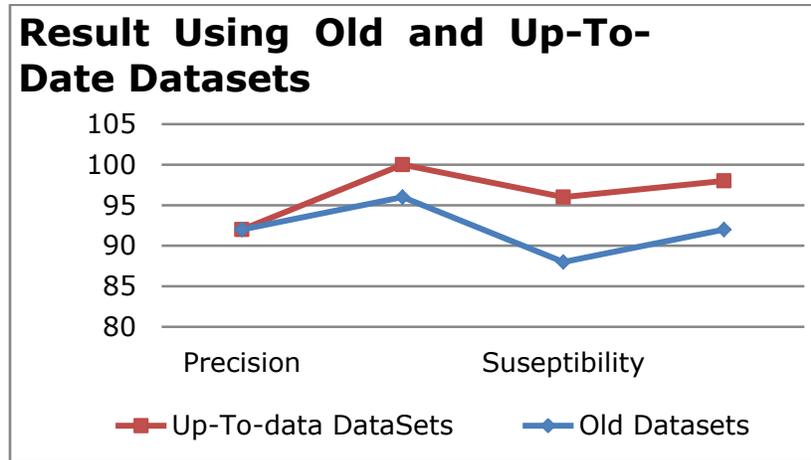


Fig. 9. Result using old and up-to-date datasets.

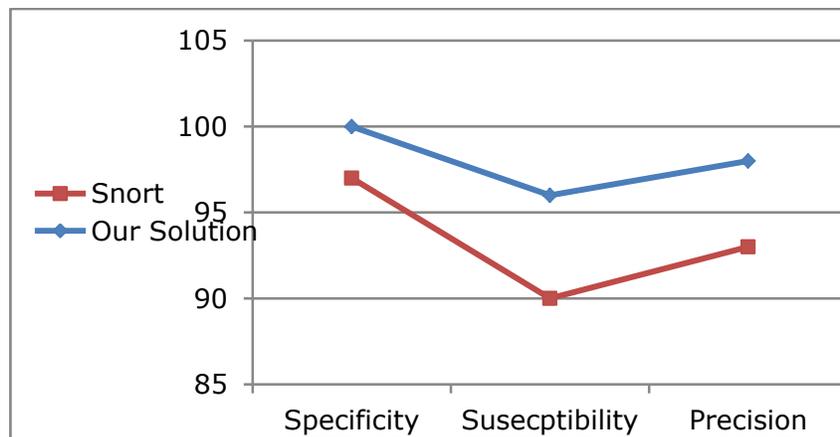


Fig. 10. Comparison result of our solution with Snort.

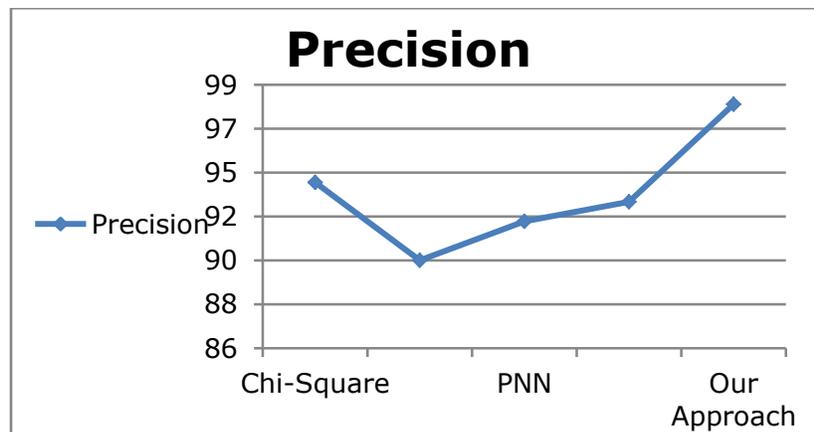


Fig. 11. Comparing our solution with others on Precision results.

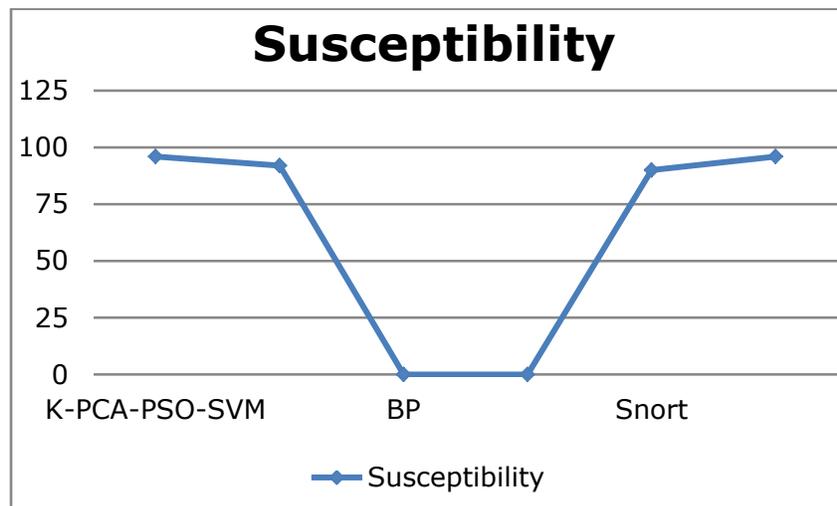


Fig. 12. Comparing our solution with others on susceptibility results.

The results in Table 4 shows that after training our solutions with old datasets, the system responded poorly with 92% of detection accuracy where the detection accuracy is 60% and 32% for known and unknown DDoS attacks respectively. After training our solutions with latest datasets the solution's detection accuracy was 98% with 50% and 48% for known and unknown attacks, respectively. This proved the fact that if we train artificial neural networks with latest and updated datasets the solution can provide better results with greater detection accuracy.

IX. CONCLUSION

We used trained ANN algo to identify TCP and UDP attacks using the basic key patterns that distinguish between authentic traffic from DDoS attacks. A mirror image of real network environment is used to start the learning process. We launched different DDoS attacks during the flow of the legitimate traffic through the network. JNNS were used to train the algorithm with prepared and pre-processed data sets and Snort AI was integrated with detection technique and got tested against different attacks. We evaluated our designed solution with other related research on signature based. We designed our solution to prevent malicious and fake data packets from reaching the target while letting go the legitimate traffic to pass through. We also evaluated our solution by training it with old existing and recently updated datasets and our designed solution provided better results and detected DDoS attacks that were almost indistinguishable with latest patterns it was trained with. Some DDoS attacks were not detected because the ANN was trained with old data patterns and thus proving that old datasets or improper training can display poor results but different DDoS cases can display better result in detecting DDoS attacks.

ACKNOWLEDGEMENT

This research was funded and conducted at Prince Sattam bin Abdulaziz University, Alkharj, Saudi Arabia during the academic year 2017 under research number 2017/01/7091.

REFERENCES

1. Tariq Ahamad, Abdullah Aljumah, "Detection and Defense Mechanism against DDoS in MANET", Indian Journal of Science and Technology, Vol 8, No. 33, Dec 2015.
2. Abdulaziz Aldaej and Tariq Ahamad, "AAODV (Aggrandized Ad Hoc on Demand Vector): A Detection and Prevention Technique for Manets" International Journal of Advanced Computer Science and Applications(IJACSA), 7(10), 2016. <http://dx.doi.org/10.14569/IJACSA.2016.071018> .
3. K. K Gupta, B. Nath, and R. Kotagiri, "Layered Approach Using Conditional Random Fields for Intrusion Detection," IEEE Transactions on Dependable and Secure Computing, Vol. 7, No. 1, pp. 35-49, Jan 2010.
4. Tariq Ahamed Ahanger, " An Effective Approach of Detecting DDoS Using Artificial Neural Networks", IEEE international Conference on Wireless Communications, Signal Processing and Networking March 2017.
5. M. A. Pérez del Pino, P. García Báez, P. Fernández López, and C. P. Suárez Araújo, "Towards Self-Organizing Maps based Computational Intelligent System for Denial of Service Attacks Detection", INES2010, 14th International Conference on Intelligent Engineering Systems, pp. 151-157, Spain, May 5-7, 2010.
6. Tariq Ahamad, Abdullah Aljumah, " Hybrid Approach Using Intrusion Detection System", International Journal of Engineering Research & Technology, Vol. 3 Issue 2, February - 2014.
7. Z. F. Chen, P. D. Qian and Z. F. Chen, "Application of PSO-RBF Neural Network in Network Intrusion Detection", 2009 3rd International Symposium on Intelligent Information Technology Application, pp.362-364, 2009
8. I. F. Akyildiz and I. H. Kasimoglu, "Wireless sensor and actor networks: research challenges," to be published Ad Hoc Networks, 2004.
9. N. Bulusu, D. Estrin, L. Girod and J. Heidemann, "Scalable coordination for wireless sensor networks: self-configuring localization systems," International Symposium on Communication Theory and Applications (ISCTA), Ambleside, UK, July 2001.
10. E. Shih, S. Cho, N. Ickes, R. Min, A. Sinha, A. Wang and A. Chandrakasan, "Physical layer driven protocol and algorithm design for energy-efficient wireless sensor networks," Proceedings of ACM MobiCom, Italy, pp:272-286, July 2001.
11. A. D. Wood and J. A. Stankovic, "Denial of Service in Sensor Networks," IEEE Computer, pp:54-62, 2002.
12. Eugene Y. Vasserman and Nicholas Hopper, "DOS flooding: Draining Life from Wireless Ad Hoc Sensor Networks", IEEE Transactions on mobile computing, Vol. 12, No. 2, February 2013.

13. P. J. Criscuolo, Distributed Denial of Service, Tribe Flood Network 2000, and Stacheldraht CIAC-2319, Department of Energy Computer Incident Advisory Capability (CIAC), UCRL-ID-136939, Rev. 1., Lawrence Livermore National Laboratory, February 14, 2000.
14. J. Mirkovic and P. Reiher, A taxonomy of DDoS attack and DDoS defense mechanisms, ACM SIGCOMM Computer Communications Review, vol. 34, no. 2, pp. 39-53, April 2004.
15. S. Ranjan, R. Swaminathan, M. Uysal, and E. Knightly, DDoS-Resilient Scheduling to Counter Application Layer Attacks under Imperfect Detection, IEEE INFOCOM'06, 2006.
16. <http://ddosattackprotection.org/blog/large-scale-ddos-attacks/>
17. <https://www.incapsula.com/ddos/attack-glossary/dns-amplification.html>
18. https://www.verisign.com/en_US/security-services/ddos-protection/what-is-a-ddos-attack/index.xhtml
19. https://business.bell.ca/web/Shop/resources/pdf/Voice/White-paper-DDoS-Forrester_Final%20EN.pdf
20. <http://blog.bell.ca/costs-and-consequences-of-a-distributed-denial-of-service-ddos-attack/>
21. <https://www.arbornetworks.com/blog/insight/ddos-as-a-smokescreen-for-fraud-and-theft/>
22. <http://usa.kaspersky.com/about-us/press-center/press-releases/2015/collateral-damage-26-ddos-attacks-lead-data-loss>
23. Mitchell T. M. (1997). Machine Learning, 1st ed. New York, McGraw-Hill Science/Engineering/Math, ch. 3,4,6,7 pp. 52-78, 81-117, 128-145, 157-198.
24. Pino, M. (September, 2005) "A Theoretical & Practical Introduction to Self Organization using JNNS". University of Applied Sciences Brandenburg.
25. Jayalakshmi, T.; Santhakumaran, A. (2011) "Statistical Normalization and Back Propagation for Classification," International Journal of Computer Theory and Engineering VOL. 3, NO. 1, pp. 89-93.
26. Bedón, C.; Saied, A. (January, 2009) Snort-AI (Version 2.4.3) "Open Source project". Available from: <http://snort-ai.sourceforge.net/index.php>
27. Roesch, M. (1998) Snort (Version 2.9) "Open Source Project". Available from: <http://www.snort.org>
28. Russell, R (1998) iptables (Version 1.4.21) "Open Source project". Available from: <http://ipset.netfilter.org/iptables.man.html>
29. Leu F.; Pai C. (2010) "Detecting DoS and DDoS Attacks Using Chi-Square", Fifth International Conference on Information Assurance and Security (IAS-09), 18-20 August 2009, Xian, pp.225-258.
30. Aljumah, Abdullah, and Tariq Ahamed Ahanger. "Futuristic Method to Detect and Prevent Blackhole Attack in Wireless Sensor Networks." *International Journal of Computer Science and Network Security (IJCSNS)* 17.2 (2017): 194.
31. Xu, X. ;Wei, D. ; Zhang, Y. (2011) "Improved Detection Approach for Distributed Denial of Service Attack Based on SVM". 2011 Third Pacific-Asia Conference on Circuits, Communications and Systems (PACCS) 17-18 July 2011, Wuhan, pp.1-3
32. Li, J.; Liu, Y.; Gu, L. (2009) "DDoS attack detection based on neural network" 2nd International Symposium on Aware Computing (ISAC), 1-4 Nov. 2010, Tainan, pp. 196 – 199
33. Akilandeswari, V. ; Shalinie, S.M. (2012) "Probabilistic Neural Network based attack traffic classification". Fourth International Conference on Advanced Computing (ICoAC), 13-15 Dec. 2012, Chennai, pp. 1-8
34. Siaterlis, C. ; Maglaris, V. (2005) "Detecting incoming and outgoing DDoS attacks at the edge using a single set of network characteristics". Proceedings of the 10th IEEE Symposium. on Computers and Communications, (ISCC) , 27-30 June 2005, pp. 469 – 475.
35. Gupta, B.B.; Joshi, C.; Misra, M. (2011) "ANN Based Scheme to Predict Number of Zombies in a DDoS Attack". International Journal of Network Security, VOL.13, No 3, pp.216–225
36. Badishi G.; Keidar I.; Romanov O.; Yachin A. (2006) Denial of Service? Leave it to Bea-ver, Project supported by Israeli Ministry of Science pp. 3-14.
37. Shi E.; Stoica I.; Andersen D. ; Perrig D. (2006) "OverDoSe: A Generic DDoS Protection Service Using an Overlay Network", Technical report CMU-CS-06-114, pp. 2-12 [Online] Available from: www.cs.umd.edu/~elaine/docs/overdose.ps
38. Chen Y.; Hwang K.; Ku W. (2007) "Collaborative Detection of DDoS Attacks over Multiple Network Domains", IEEE Transactions on Parallel and Distributed Systems VOL. 18 NO. 12, pp. 1649 – 1662.
39. Al-Duwairi, B. ; Manimaran, G. (2004) "A novel packet marking scheme for IP trace-back", Proceedings of the tenth International Conference on Parallel and Distributed Systems, 7-9 July 2004. (ICPADS), pp. 195-202
40. Gong, C. ; Sarac, K. (2008) "A More Practical Approach for Single-Packet IP Traceback using Packet Logging and Marking", IEEE Trans on Parallel and Distributed System, VOL. 19 NO. 10, pp.1310-1324.
41. Yu S. ; Zhou, W. ; Doss, R. ; Jia, W (2011) "Traceback of DDoS Attacks Using Entropy Variations", Transactions on Parallel and Distributed Systems VOL. 22, NO 3, pp. 412-425.
42. Jie-Hao, C.; Feng-Jiao, C.; Zhang. (2012) "DDoS defense system with test and neural network". IEEE International Conference on Granular Computing (GrC), 11-13 Aug. 2012, Hangzhou, China, pp. 38 – 43.

Artificial Intelligence in Bio-Medical Domain

An Overview of AI Based Innovations in Medical

Muhammad Salman, Abdul Wahab Ahmed, Omair Ahmad Khan, Basit Raza, Khalid Latif

Department of Computer Science
COMSATS Institute of Information Technology
Islamabad 45550, Pakistan

Abstract—In this era and in the future, artificially intelligent machines are replacing and playing a key role to enhance human capabilities in many areas. It is also making life style better by providing convenience to all including normal human beings and professionals as well. That is why AI is gaining huge attention and popularity in the field of computer science by which it has revolutionized the rapidly growing technology known as expert system. The applications of AI are working in many areas with huge impact and being used widely as well. AI provides quality and efficiency in almost every area, we are evolving it in. The main purpose of this paper is to explore the area of medical and health-care with respect to AI along with ‘Machine Learning’, and ‘Neural Networks’. This work explores the current use of AI in innovations, in the particular field of Bio-Medical and evaluated that how it has improved hospital inpatient care and other sectors related to it i.e. smart medical home, virtual presence of doctors and patients, automation in diagnostic, etc. that has changed the infrastructure of medical domain. Finally, an investigation of some expert systems and applications is made. These systems and applications are widely used throughout the world and a ranking mechanism of their performance has proposed accordingly in an organized manner. We hope, this work will be helpful for the researchers coming to this particular area and to provide a syntactic information that how computer science (i.e. AI, ANN, ML) is revolutionizing the field of bio-medical and healthcare.

Keywords—Artificial intelligence; expert systems; bio-medical; healthcare; innovations

I. INTRODUCTION

Artificial intelligence (AI) [1] came into being as an inspiration whose ultimate goal was to copy and learn like

human brain and to determine the upcoming considerations and real world challenges with a perfect intelligent approach. Scientists and researchers everywhere in the globe were terribly excited regarding advancements in innovations those have arisen from a natural need to form newer and higher technologies. These innovations may facilitate the humanity to increase on the far side of their own physical caliber. The promise of AI thought has continuously been on the horizon from realistic science to the imagination in movies and literature. AI for the most part permits the capability to store and process vast amounts of information in an intelligent manner, and specifically translating that data into information that could be used practical tools. Since its origination, AI has been deployed for extremely selective defense and area exploration applications whereby its success in resolving issues for specific areas similar to risk prediction is concerned. Now, gradual transition of its utility in health care is being widely intimated through AI-based systems those can afford higher diagnosing, cure and treatment of exhausting conditions.

Artificial intelligence has attracted many users over the past, it supports medical sciences, businesses, scientific researches etc. These systems when implemented with great cautions, gave surprisingly accurate results and avoided errors likely happened by humans [2]. These systems never falter because they follow a specific track to achieve a goal by using the information provided [3], [4]. In case if we don't have enough knowledge for designing some system, system is provided with past knowledge base to develop itself and make decisions on that knowledge base.

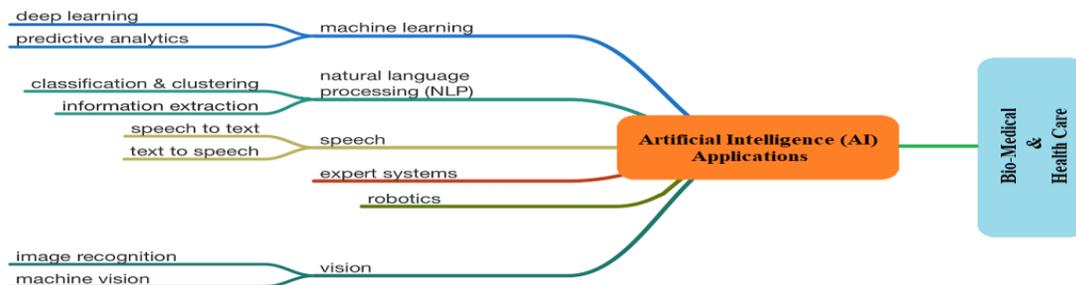


Fig. 1. Branches of AI applied in biomedical domain.

AI in health care and drugs may be a new analysis space that mixes refined realistic and computational approaches with the insights of professional doctors to provide new innovations

for rising health care in the form of tools. AI is the study of concepts, that modifies computers to try and to do the items that create individuals intelligent. The ultimate task or goal of

AI is to form or to evolve computers more and more helpful and to grasp the principles that make intelligence attainable. The branches of AI working in the domain of healthcare and biomedical are shown in Fig. 1.

Medicine could be an area which acquires technology to much more extent. With the boost of our desires and expectations of the very best quality medical diagnosis in health care and the ascension of additional elaborated medical information leave the doctors to pay proper attention and to give proper time to each case. It is also quite tough for a physician to stay up to date with the most recent developments in his field. As a result, due to lack of attention and adequate time, many of the medical recommendations are based on speedy diagnosis of the case, hoping with the physician's unaided experience. Solely, in the very rare cases, there may have a situation to utilize the recent research and methodologies to ensure confidence for both side i.e. doctor and patient and to ensure that, most recent information is delivered to secure any specific patient but not applicable in general.

It is known that, computers are quite intellectual as well as have the capacity to deliver as instrument used for detection, and these aspects could be integrated to improve and to investigate the medical diagnosis and aid tool. Demonstrator to AI research is that physicians and therefore the computer can interact in disrupted communication and dialogue, system ceaselessly being attentive of all the knowledge we are having in laboratory, diagnosis reports, finding history, physical findings, and also informing doctor with foremost appropriate report of diagnosis and by suggesting the suitable and fruitful prescription course for the patient.

Rest of the paper follows as Section I(A) explains the different aspects of medical domain where AI is working whereas Section I(B) explores a few pioneer expert laboratories. Section II shows the glimpses of smart home mechanism with different aspects being evolved in smart healthcare home. Section III demonstrates recognized expert systems and applications which are being used worldwide with exception and follows Section IV that contain the concise information and grading of expert system as an information in tabular form where on the basis of performance some analysis is made by proposing grades. At the end, Section V elaborates importance and future directions of AI in the particular domain of bio-medical.

A. Effectiveness of AI in Medical

Artificial intelligence is what provides computers the power to observe, learn, think, reason, and even perceive human emotions, permitting computers to do quite simply

repetitive tasks. Within the medical field, AI is being designed to help doctors (not replace them) within the effort to cut back the death rate among patients awaiting care from specialists. There is a huge range of aspects in Bio-Medical domain where AI in conjunction with Machine Learning (ML) and natural language processing (NLP) has effects on it. During this section we are going to discuss about the ways that are primarily associated with robotics and machine learning. Table 1 shows the different aspects of AI which are improving the medical field to much extent and providing quick and accurate health facilities by reducing the costs in extra.

B. Expert Laboratories and Clinical Information Systems

Now computer scientists are working on innovative applications which will improve diagnostic techniques and will help to classify diseases without human error. Resources will be managed in a better way by using AI. These systems will also help to avoid procedures which are hazardous to health like X-rays and other electromagnetic waves [5]. The focus of next development of Aml systems should be such, that they are less harmful to humans. All we need is a joint venture of experts from different fields. There are some expert systems those were the very first expert systems for diagnostic purpose.

a) **Puff** is an expert system which is used for automatic clarification of pulmonary function. Most likely Puff was the first AI expert system which has been used in biomedical field in San Francisco by 1977. PUFF can use to test the patient who is suffering from lungs disease also then pulmonary physiologist depicts its presence and generates recommendations and reports for the patient's file.

b) **Germ Watcher** is an expert system that observes microbiological information from different hospitals and laboratory systems. It classifies those microbiology cultures that produce hospital's collected viral and warn the recommended department of U.S. for disease control.

c) **PEIRS** (Pathology Expert Interpretative Reporting System) appends interpretative comments to pathology reports. During its working, the system generated nearly 80-100 patients reports on daily basis, having 95% accuracy in the diagnosis. The major areas of system's reports include thyroid function tests, arterial blood gases, urine and plasma, human chorionic gonadotrophin and alpha fetoprotein, and glucose tolerance tests, etc.

The web references of these expert systems are mentioned below^{1, 2, 3}:

¹ http://www.openclinical.org/aisp_puff.html

² http://www.openclinical.org/aisp_germwatcher.html

³ http://www.openclinical.org/aisp_peirs.html

TABLE I. EFFECTIVENESS OF AI IN MEDICAL

Aspect	Description	Acknowledgment
Robotics	<p>Robotic is based on:</p> <ul style="list-style-type: none"> Hollywood amusement Sci-Fi novels childhood fantasies Still AI is not as evolved as the actor Spielbergian expectations of evolutions 	<ul style="list-style-type: none"> AI has arisen in medication It is to remodel all the fields like education, medical, economics, etc.
Fast and Precise Diagnostics	<p>Through the AI:</p> <ul style="list-style-type: none"> Human brain is imitated with Artificial Neural Network. These Neural Networks has power to learn. These are very promising in the diagnosis Accurate and Quick methods 	<p>Disease Diagnosed:</p> <ul style="list-style-type: none"> Melanoma Optical Issues Huge advancements to cure the different types of cancer
Therapeutic Robots	<ul style="list-style-type: none"> Alzheimer's patients get assistance through therapeutic robots. It came into view to deal with human health impact produced from caressing the animals. 	<p>Robotic pets facilitate:</p> <ul style="list-style-type: none"> Nurture Human Brain Operate Delays cognitive aspect which is responsible to improve quality of life. Decreases the reliance on social services, which helps a human to stay in home with less medical help.
Reduces Errors associated with Human Fatigue	<p>Human Doctors Errors:</p> <ul style="list-style-type: none"> Diagnose almost 80 patients a week Exhausting to pay attention to the needs of each and every patient. Whereas, AI based systems are not limited in the work hours and human fatigues. 	<ul style="list-style-type: none"> Like Spell Checker Helping Physicians to reduce the human like mistakes Providing Relief to overwhelm with different tasks.
Decrease in Medical Cost	<p>How cost can be reduced:</p> <ul style="list-style-type: none"> Reducing the work place visits to almost zero. Online Care is provided. Patient just update his/her medical reports while staying in home and saves the time, energy and cost to travel for a work place 	<ul style="list-style-type: none"> Huge amount of money is getting saved by different technologies evolutions like drug's interactions, precise diagnosis Reduced error is also a way to save money
Movement Assistance	<ul style="list-style-type: none"> Medical personnel are not enough to provide assistance and the field is struggling in it. So, robots seem to be very prominent to utilize as manpower to assist the patients in movements. These are good for repetitive tasks like medical pharmacy and physical therapy. 	<ul style="list-style-type: none"> HAL 5 (Hybrid Helpful Limb) suit has overcome the mobility issues. One can carry the double weight through this A well promising tool for health care professionals.
Improved Radiology	<p>Radio surgery:</p> <ul style="list-style-type: none"> Cyber-Knife, provides the facility to eliminate tumor at any place in the body. Image oriented and guided technologies are use with the computational system to facilitate the patient's movements throughout the process 	<ul style="list-style-type: none"> Without harming other healthy tissue, it targets accurately to the tumor and eliminate it by imposing radiation on it
Virtual Presence	<ul style="list-style-type: none"> You are not supposed to leave your bed once more for any process. Doctors are able to see the matters and to communicate with patients and workers while not being present there. So, they are not bound to one workplace and can deal the matters from far place as being at front. 	<ul style="list-style-type: none"> It is the most helpful because if one is not in a condition to travel then still can get diagnosis while staying in home Doctor can also avail the facility to deal patients at different locations
Invasive Surgery Advances	<ul style="list-style-type: none"> The system which has evolved fantastic achievements in surgery in known as Da Vinci Si HD Surgical System. Da Vinci Si HD provides with clear, accurate and superior visuals in imaging. 	<p>It Delivers:</p> <ul style="list-style-type: none"> Smaller Incisions Cut Back Patient Pain Reduces medication shorten hospital stays

II. SMART HOME/HEALTHCARE FACILITATING TOOLS

In this section, we provide some of directly data taking techniques that, if connected with AI technologies can help out to achieve much more advancements in the field of medical and can provide fast and quick health care services in home office and anywhere i.e. smart-home, smart-hospital, etc.

1) **Activity Recognition:** As human activities contain huge amount of data which can be predicted through psychological and through knowledge based systems. As AI systems are always there, examining the desires and the requirements of the human beings for which these are implemented. Therefore, it requires data about the activities of which it is examining or focusing [6]. And there are huge number of techniques in [7], [8] for the purpose of activity recognition and it is the toughest challenge in the field, and identification to the activities is the ultimate goal of such approaches. As with the passage of time there are huge amount of advancements in computing and sensor networks about which activity recognition techniques are totally dependent and precision cannot be met without these advancements in the certain fields. As sensor networks is becoming more and more advanced in this era, information is being collected from various ways, such as sensors are attached with body [9], or if not possible then stitched in the clothes [10], [11]. For the other movements like sitting, walking, climbing and falling, etc. [12], [13]; for the collection of data from gestures and postures [14]-[18]; and moving to some other activities like sleeping, eating and cooking, it acquires location based sensors which are used to determine activity in indoor environments [19], [20]. So, from the subject of smart city applications, we can also create intelligent application for environment such as smart environment which has been adopted widely for health monitoring [21] and with strong power source these can be affective for collection of data for a long time [22]. And to recognize these activities we are required with some models that can detect the class of an activity and understand the differences between the activities i.e. difference between walking and running, cooking and eating, etc. So, these kinds of models can be labeled as activity models.

2) **Pattern Discovery and Anomaly Detection:** After being messed with the activity detection and recognition, the next step is to find out some patterns related to a human on the basis of activities being performed. It is based on the activities which are used to recognize through supervised learning. Even unsupervised learning can get into it and a system can learn some activities by itself after observing recurrent sequences of some activity. There is a huge research work regarding the methods for the mining with respect to activities, its includes mining of some frequent sequences [23], and in [24] activities are mined in the form of patterns implying the technique of regular expression on

it, and constraint or restriction based mining [25] and mining the patterns which occurs periodically and are frequent [26], [27]. And for the detection of interleaved patterns, [28] implemented a genetic algorithm in a different manner than the previous one i.e. by implying unsupervised learning in it. This discovering of patterns is of extreme wealth because once they are formed, then can be extremely useful in detecting the instance if the same pattern occurs again.

3) **Planning and Scheduling:** Planning and scheduling of any machine is very important. It can be very fruitful in much AmI application when we automate them. In automatic planning, we take an initial state or some initial background and on the basis of these initial knowledge it takes possible actions accordingly. It can be helpful in many AmI application and care related environment. For example, it can be useful in daily scheduling activities in an efficient manner so that many dementia and liver patient can be facilitated. In the past research work many planning techniques are proposed, from which a few are mentioned in Table 2.

4) **Decision Support:** Decision support systems (DSSs) [34]-[36] are mostly used in healthcare environments for assisting and analyzing the data of patients [37]-[42]. In DSSs we have mainly two main stream approaches, one is knowledge based and the other is non-knowledge based.

In knowledge based, we have vocabulary that is stored in database and the inference engine that contains the rules according to different set of information. It also comprised with IF-ELSE rules, where the engine combines the set of rules from the database in order to generate new knowledge and perform set of action accordingly. Different methods are proposed in past those are using this technique [43]. Whereas, in non-knowledge-based DSSs, no direct information or knowledge is provided, but it learns the rules from the past experience. Different algorithms and decision trees are also proposed for learning knowledge.

Both these techniques are frequently used in AmI for enhancing communication skills of doctors and nurses. DSS based Context-aware knowledge is also proposed that can gather data from their environment and take decisions accordingly [44].

5) **Privacy Preserving Techniques:** As AmI systems are getting fame, more information will be gathered through individuals. So, the information is very sensitive and critical. This creates many privacy issues from which many privacy concerns focus on sensitive monitoring [45]. Many AmI system are deployed with internet that can create lots of problems like internal or external attacks so many techniques are also proposed and these techniques are quite mature [46]. Also many approaches are developed to ensure that, critical and sensitive data cannot be gleaned from mined patterns. [47], [48].

TABLE II. METHODS USED IN AI FOR BIOMEDICAL DOMAIN

Methods	Reference	Implemented Technique
Decision-theoretic	[29]	Markov decision processes
Search methods	[30]	Forward and Backward
Graph-based	[31]	Graph plane
Hierarchal	[32]	O-plan
Hierarchal	[33]	Reactive plan

III. APPLICATIONS AND EXPERT SYSTEMS OF ARTIFICIAL INTELLIGENCE GETTING USED IN BIOMEDICAL FIELD

A. Expert Systems of Biomedical and Healthcare

a) *Fuzzy Expert Systems in Medicine*: Fuzzy logic is a technique which is used for data handling purpose that allows ambiguity, and particularly used in medical field. This expert system gets and uses the idea of fuzziness in a computationally efficient way. This technique is used in many medical fields such as multiple logistic regression analysis and also used for the diagnosis of many diseases like lungs cancer, acute leukaemia, breast and pancreatic cancer. Fuzzy expert system can also predict about the survival of patient who is suffering from breast cancer [49].

b) *Automated Fraud Detection in Health care sector*: It is a new medical technology of AI where, the system monitors the employer having sick leave by monitoring its activities on social media. This application presumably analyzes sick employees when they post their status on social media, then investigates that either they are sick or not, but skipping their time and work [50]. The tools of data analytics in this system works automatically and thus, these systems intelligently learn automatically by their own [51].

c) *Medical adherence application for mobile devices (AiCure)*: This application of AI facilitates the patient about the information of disease and its treatment, conforming ingestion and reminds patient for medication doses according to the time table of patient [50]. When patient perform incorrect behavior then system identifies, acknowledges this behavior to doctor and provides data to the doctor for the remedy [52].

d) *Care-O-bot 3 (Fraunhofer IPA)*: In this system, a robot helps and aids a patient in his house. By designing a map, the robot navigates automatically to approach a target by adjusting itself on the map and avoiding the obstacles. The robot can also provide the facilitation to bring and fetch service by learning the object [53] and it works according to the order of user by technique of face recognition.

e) *Evolutionary Computation in Medicine*: Evolutionary computation is a general expression for various computational methods which is based on the process of natural evolution that mimics the procedure of natural selection. Genetic algorithm is the most useful form of evolutionary computation in medical areas [54]. The rule of genetic algorithm is majorly

used to predict the outcome of seriously ill patient. In MRI segmentation of brain tumors, evaluating the adequacy of treatment strategy and it is handled by evolutionary computations [55]. Computerized analysis of mammographic micro categorization is also done by evolutionary computation.

f) *Artificial Intelligence to Improve Hospital Inpatient Care*: Clinical decision support system is one of the most popular methods of AI. This expert system is initially focusing in diagnosing the condition of patient by giving demographic information and his symptoms. Mycin is another expert system developed in 1970 based on rules used for the diagnosis and identification of bacteria and then recommends antibiotics for the treatment of infected patient [56]. Pathfinder is another method used for the identification of lymph-node disease for the support of pathologist. This method used Bayesian network and such technique helps for the diagnosis of varying form of cancer and for unexpected heart diseases [57].

g) *Artificial Intelligence Approaches for Medical Image Classification*: Some applications of AI are used for diagnostic sciences in categorization of different type of biomedical image such as identify tumors in brain etc. Decision-support tools and model-based intelligent system are very useful methods for the medical image classification for analysis and evaluation purpose. CAD support radiologist that uses the result taken from the computerized analysis of those tools [58]. These tools help radiologist to increase the accuracy of results taken from such expert systems and to minimize the rate of errors [59].

h) *Implementation Scheme for Online Medical Diagnosis System Using Multi Agent System with JADE*: In this paper the idea of online medical service is formulated for the users of internet. The working of this multi-agent based system is well but has some challenges:

- i) Communication of Services.
- ii) Data Security.
- iii) Interconnection of User and Agent.
- iv) Synchronization of Different Services.

The main purpose of this system is to build a type of system that could have the ability to run in all environments. The Agent Communication Channel (ACC) is another module which connects the remote and local platforms [60]. This framework is created by JADE and whenever JADE launches, the ACC starts its communication.

B. Usage of Artificial Neural Network based Techniques in Biomedical Domain

a) *MRI Brain Tumor Analysis*: Some ANN techniques used for the classification of images in diagnostic science. A general regression neural network (GRNN) is used as a three dimensional technique of classification for the image of brain tumor [61]. Least Squares Support Vector Machines (LS-SVM) is another proposed method used for the diagnosis of normal and abnormal areas of brain from data of MRI [62].

Because of autonomous way to classify MRI image, it shows result with greater accuracy than other classifiers.

b) *Endoscopic Images:* Advanced fuzzy inference neural network is a technique for classification of endoscopic images. This technique works by merging the methods of fuzzy systems and radial based function. By this idea of mixture of many classifiers it shows particular parameters and features with an accuracy of 94.28%. However, radial based function classifies the fast rate of training than fuzzy systems [63]. These types of techniques show their results by both statistical and texture features [64].

c) *Heart Disease Classification:* Artificial neural network has also proved its ability by working on the classification of heart disease. In this technique for the classification of stroke, the input of sensor is given to the system that uses [65] forward feed network with the rule of back propagation way [66]. Effective result of classification is given by simulation system which is then moved forward to the network for testing purpose.

d) *Decision Support System to Diagnose Nodules:* Through the concept of ANN, a new system is proposed and known as decision support system (DSS). A decision support system that diagnoses nodules into benign, malignant and identify [67] its severity by the analysis the collected data. This known method has delivered the accuracy up to 95% by collecting the dataset of 63 samples [68].

IV. RESULTS AND DISCUSSIONS

So far in the literature review, it is witnessed that the Artificial Intelligence with the concepts or domains of, Machine Learning, Natural Language Processing, Neural Networks and advanced computing, health care facilities are made fast and quick. Not only this, healthcare process is automated as good as it can detect some anomaly from the activities presuming that something severe is going to happen. Through this, it can put alarms on and emergency alert at house level. We can say that after the smart city concept in terms of computing and Internet of Things (IoT), the new concept is smart house or smart environment with respect to healthcare, and from this concept each and every personnel can get benefits.

Table 3 shows some examples of expert systems and their description about the purpose for which they are implemented. We have evaluated the improvement of the systems in medical and clinical with stars. The highest possibility of rank for a system is 5 stars and 5 stars are awarded to those systems which have ideal accuracy i.e. above 99%. There is also a discussion about the type of systems, that what kind of logic they are using. It is of high worth to mention about the expert systems and fuzzy expert system to demonstrate the types of the systems as mentioned in Table 3.

TABLE III. GRADING OF EXPERT SYSTEMS IN CLINICAL

System	Ref #	Purpose	Type	Performance	Improvement (GRADING)
1	[69]	Improving quality of first aids	Expert system	Improved	★ ★ ★
2	[70]	Prediction of low back pain	ANN and adaptive neuro-fuzzy inference system	Few systems detected pain successfully	★ ★
3	[71]	Identifying the type of neuropathy	Fuzzy expert system	93.26 %	★ ★
4	[72]	Diagnosing types of headache	Expert system	98 %	★ ★ ★
5	[73]	Diagnosis of tuberculosis	ANN optimized with genetic algorithm	94.88 %	★ ★
6	[74]	Diagnosing strabismus	Web-based neural network system	100%	★ ★ ★ ★ ★
7	[75], [76]	Diagnosis of hepatitis and its fatality	Multilayer neural network based on Levenberg– Marquardt (LM) algorithm and a probabilistic neural network (PNN)	91.2% 91.87%	★ ★
8	[77]	Suggesting radiotherapy regimen based on anatomy	A combination of an expert system and ANN	96%	★ ★ ★

9	[78]	Analyzing serology laboratory tests and providing expert advice and possible disease	Web-based fuzzy expert system	91%	★ ★
10	[79]	Prescribing the most appropriate Chinese acupuncture treatment	Expert system	--	★
11	[80]	Diagnosis of breast cancer	Support vector machine (SVM), Naïve Bayes classifier (NBC) and ANN based on color wavelet features	98.3%	★ ★ ★ ★
12	[81]	Classification of heartbeats	Wavelet neural network (WNN) based on features extracted from ECG	98.78%	★ ★ ★ ★
13	[82]	Differentiating heartbeats	Fuzzy expert system	90-95 %	★ ★ ★
14	[83]	Identifying diabetic retinopathy	Different ANN classifiers	94%	★ ★ ★

An expert system is a computational based system which emulates the reasoning process of a human expert e.g. it is evolved in such a way that it has the capability to think like intelligent human mind, reasoning like an expert human being and to take decisions like a professional mind. In fact, expert systems are the mimicry of human mind that is expert to some particular field. These expert systems are applied throughout the world for different purposes but some of well demonstrated purposes include consulting, diagnostic, learning, designing, planning and decision support. Whereas, the working of fuzzy expert system involves fuzzy set theory instead of linear algebra or Boolean function, and the knowledge is normally presented in the form of some fuzzy production rules i.e. the most common example is 'IF X THEN Y', where X and Y are fuzzy sets. And these fuzzy sets are said to be 'rulebase' or knowledge base of a fuzzy expert system.

V. CONCLUSION AND FUTURE DIRECTIONS

In this paper, we critically explored the branches of artificial intelligence within the biomedical and healthcare sectors. The information is presented in a very concise way and investigated the performance of some expert systems that are employed in the healthcare domain. We hope, this research work will be helpful for the new researchers of AI to explore this particular domain in an appropriate way and make the field of AI more robust and applicable in sense of performance in the healthcare. Medicine has shaped as an upscale testbed for ML experimental findings within the previous decade, permitting researches and developers to evolve advanced and complicated systems with super power of learning ability. Whereas, we witness abundant sensible use of knowledgeable tools in the clinical recommendations. Now a days, ML based systems appear to be utilized with lot of experimental manner. Therefore, there are many aspects and conditions in medical

where these approaches can be used and can perform a huge contribution the field healthcare.

To make this AI even more powerful, organizations have to implement better hardware architectures like pervasive or ubiquitous hardware approaches. These systems will be able to search for the data that is not in the same place [4]. This will improve data mining techniques and we will be able to search for a solution on a bigger scale. The future of AI is not just limited to this, it will recognize people's expressions, mood, need and will respond to these emotions as they would be preprogrammed to do [59]. As the world is becoming a global village, we are facing the privacy issues more and more. As we add more devices to the system that means we are increasing the privacy issues more and more wide and easy [84]. Ambient Intelligence (Aml) can increase the number of security concerns because it accesses many other devices which increases the breach points. Aml sensors will use wireless protocols that could be intercepted easily. To avoid this situation, every communication should be encrypted, biometric authentication should be used to verify the concerned person [85]. Privacy by Design (PbD) should be taken into account because, it sets privacy on the sensor devices once and privacy limits got set for future as well.

REFERENCES

- [1] "One hundred year study on Artificial Intelligence (AI100)", Stanford University, <https://ai100.stanford.edu>.
- [2] Forestier, Germain, et al. "Automatic matching of surgeries to predict surgeons' next actions." *Artificial Intelligence in Medicine* (2017).
- [3] Jha, Saurabh, and Eric J. Topol. "Adapting to artificial intelligence: radiologists and pathologists as information specialists." *JAMA* 316.22 (2016): 2353-2354.
- [4] Hamet, Pavel, and Johanne Tremblay. "Artificial Intelligence in Medicine." *Metabolism* (2017).

- [5] Hansen, Karl R., et al. "Predictors of pregnancy and live-birth in couples with unexplained infertility after ovarian stimulation–intrauterine insemination." *Fertility and sterility* 105.6 (2016): 1575-1583.
- [6] G. Singla, D. Cook, and M. Schmitter-Edgecombe, "Recognizing independent and joint activities among multiple residents in smart environments," *J. Ambient Intell. Humanized Comput.*, vol. 1, no. 1, pp. 57–63, 2010.
- [7] L. Chen, J. Hoey, C. Nugent, D. Cook, and Z. Hu, "Sensor-based activity recognition," *IEEE Trans. Syst. Man Cybern. C, Appl. Rev.*, vol. 42, no. 6, pp. 790–808, Nov. 2012.
- [8] P. Rashidi, D. Cook, L. Holder, and M. Schmitter-Edgecombe, "Discovering activities to recognize and track in a smart environment," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 4, pp. 527–539, Apr. 2011.
- [9] A. Yang, R. Jafari, S. Sastry, and R. Bajcsy, "Distributed recognition of human actions using wearable motion sensor networks," *J. Ambient Intell. Smart Environ.*, vol. 1, no. 2, pp. 103–115, 2009.
- [10] H. Harms, O. Amft, G. Troster, and D. Roggen, "Smash: A distributed sensing and processing garment for the classification of upper body postures," in *Proc. ICST 3rd Int. Conf. Body Area Netw.*, 2008, p. 22.
- [11] C. Metcalf, S. Collie, A. Cranny, G. Hallett, C. James, J. Adams, P. Chappell, N. White, and J. Burrige, "Fabric-based strain sensors for measuring movement in wearable telemonitoring applications," in *Proc. IET Conf. Assisted Living*, 2009, pp. 1–4.
- [12] U. Maurer, A. Smailagic, D. Siewiorek, and M. Deisher, "Activity recognition and monitoring using multiple sensors on different body positions," in *Proc. Int. IEEE Workshop Wearable Implantable Body Sensor Netw.*, 2006, DOI: 10.1109/BSN.2006.6.
- [13] R. Srinivasan, C. Chen, and D. Cook, "Activity recognition using actigraph sensor," in *Proc. 4th Int. Workshop Knowl. Disc. Sensor Data*, Washington, DC, USA, 2010, pp. 25–28.
- [14] S. Lee and K. Mase, "Activity and location recognition using wearable sensors," *IEEE Perv. Comput.*, vol. 1, no. 3, pp. 24–32, Jul.-Sep. 2002.
- [15] R. Agrawal and R. Srikant, "Mining sequential patterns," in *Proc. 11th Int. IEEE Conf. Data Eng.*, 1995, pp. 3–14.
- [16] H. Junker, O. Amft, P. Lukowicz, and G. Troster, "Gesture spotting with body-worn inertial sensors to detect user activities," *Pattern Recognit.*, vol. 41, no. 6, pp. 2010–2024, 2008.
- [17] N. Krishnan, P. Lade, and S. Panchanathan, "Activity gesture spotting using a threshold model based on adaptive boosting," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2010, pp. 155–160.
- [18] J. Mantyjarvi, J. Himberg, and T. Seppanen, "Recognizing human motion with multiple acceleration sensors," in *Proc. IEEE Int. Conf. Syst. Man Cybern.*, 2001, vol. 2, pp. 747–752.
- [19] D. Cook, "Learning setting-generalized activity models for smart spaces," *IEEE Intell. Syst.*, vol. 27, no. 1, pp. 32–38, Jan.-Feb. 2012.
- [20] C. Nugent, M. Mulvenna, X. Hong, and S. Devlin, "Experiences in the development of a smart lab," *Int. J. Biomed. Eng. Technol.*, vol. 2, no. 4, pp. 319–331, 2009.
- [21] T. van Kasteren, G. Englebienne, and B. Krose, "An activity monitoring system for elderly care using generative and discriminative models," *Pers. Ubiquitous Comput.*, vol. 14, no. 6, pp. 489–498, 2010.
- [22] B. Logan, J. Healey, M. Philipose, E. Tapia, and S. Intille, "A long-term evaluation of sensing modalities for activity recognition," in *Proc. 9th Int. Conf. Ubiquitous Comput.*, 2007, pp. 483–500.
- [23] T. Gao, T. Massey, L. Selavo, D. Crawford, B. rong Chen, K. Lorincz, V. Shnyder, L. Hauenstein, F. Dabiri, J. Jeng, A. Chanmugam, D. White, M. Sarrafzadeh, and M. Welsh, "The advanced health and disaster aid network: A light-weight wireless medical system for triage," *IEEE Trans. Biomed. Circuits Syst.*, vol. 1, no. 3, pp. 203–216, Sep. 2007.
- [24] T. Barger, D. Brown, and M. Alwan, "Health-status monitoring through analysis of behavioral patterns," *IEEE Trans. Syst. Man Cybern. A, Syst. Humans*, vol. 35, no. 1, pp. 22–27, Jan. 2005.
- [25] J. Pei, J. Han, and W. Wang, "Constraint-based sequential pattern mining: The pattern-growth methods," *J. Intell. Inf. Syst.*, vol. 28, no. 2, pp. 133–160, 2007.
- [26] P. Rashidi and D. J. Cook, "Keeping the resident in the loop: Adapting the smart home to the user," *IEEE Trans. Syst. Man Cybern. A, Syst. Humans*, vol. 39, no. 5, pp. 949–959, Sep. 2009.
- [27] E. Heierman, III, and D. Cook, "Improving home automation by discovering regularly occurring device usage patterns," in *Proc. 3rd IEEE Int. Conf. Data Mining*, 2003, pp. 537–540.
- [28] M. Ruotsalainen, T. Ala-Kleemola, and A. Visa, "Gais: A method for detecting interleaved sequential patterns from imperfect data," in *Proc. IEEE Symp. Comput. Intell. Data Mining*, 2007, pp. 530–534.
- [29] Abel, David, James MacGlashan, and Michael L. Littman. "Reinforcement Learning As a Framework for Ethical Decision Making." *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*. 2016.
- [30] Cserna, Bence, et al. "Anytime versus Real-Time Heuristic Search for On-Line Planning." *Ninth Annual Symposium on Combinatorial Search*. 2016.
- [31] Zhou, Xiaolu, Mingshu Wang, and Dongying Li. "From stay to play—A travel planning tool based on crowdsourcing user-generated contents." *Applied Geography* 78 (2017): 1-11.
- [32] Wickler, Gerhard, Lukás Chrpa, and Thomas Leo McCluskey. "KEWI-A Knowledge Engineering Tool for Modelling AI Planning Tasks." *KEOD*. 2014.
- [33] Michaud, François, and Monica Nicolescu. "Behavior-based systems." *Springer Handbook of Robotics*. Springer International Publishing, 2016. 307-328.
- [34] Bonczek, Robert H., Clyde W. Holsapple, and Andrew B. Whinston. *Foundations of decision support systems*. Academic Press, 2014.
- [35] S. Eom and E. Kim, "A survey of decision support system applications (1995–2001)," *J. Oper. Res. Soc.*, vol. 57, no. 11, pp. 1264–1278, 2006.
- [36] Mitchell, Jordan, et al. "Informatics for Health and Social Care Differences in pneumonia treatment between high-minority and low-minority neighborhoods with clinical decision support system implementation between high-minority and low-minority neighborhoods with clinical decision." (2016).
- [37] Massam, Bryan H., and Jacek Malczewski. "The location of health centers in a rural region using a decision support system: a Zambian case study." *Geography Research Forum*. Vol. 11. 2016.
- [38] Weaver, Charlotte A., et al. "Healthcare information management systems." *Cham: Springer International Publishing* (2016).
- [39] Mitchell, Jordan, et al. "Differences in pneumonia treatment between high-minority and low-minority neighborhoods with clinical decision support system implementation." *Informatics for Health and Social Care* 41.2 (2016): 128-142.
- [40] Gray, Carolyn Steele, et al. "Supporting goal-oriented primary health care for seniors with complex care needs using mobile technology: evaluation and implementation of the health system performance research network, Bridgepoint electronic patient reported outcome tool." *JMIR Research Protocols* 5.2 (2016).
- [41] M. Romano and R. Stafford, "Electronic health records and clinical decision support systems: Impact on national ambulatory care quality," *Arch. Internal Med.*, vol. 171, no. 10, pp. 897–903, 2011.
- [42] M. Perwez, N. Ahmad, M. Javaid, and M. Ehsan Ul Haq, "A critical analysis on efficacy of clinical decision support systems in health care domain," *Adv. Mater. Res.*, vol. 383–390, pp. 4043–4050, 2012.
- [43] M. Kaptein, P. Markopoulos, B. de Ruyter, and E. Aarts, "Persuasion in ambient intelligence," *J. Ambient Intell. Humanized Comput.*, vol. 1, no. 1, pp. 43–56, 2010.
- [44] Furmankiewicz, M., A. Sołtysik-Piorunkiewicz, and P. Ziuziański. "Artificial intelligence systems for knowledge management in e-health: the study of intelligent software agents." *Latest Trends on Systems: The Proceedings of 18th International Conference on Systems*, Santorini Island, Greece. 2014.
- [45] G. Demiris, D. Oliver, G. Dickey, M. Skubic, and M. Rantz, "Findings from a participatory evaluation of a smart home application for older adults," *Technol. Health*.
- [46] Huang, Tien-Chi, Chia-Chen Chen, and Yu-Wen Chou. "Animating eco-education: To see, feel, and discover in an augmented reality-based

- experiential learning environment.” *Computers & Education* 96 (2016): 72-82.
- [47] Panagiotakis, Costas, and Georgios Tziritas. “A minimum spanning tree equipartition algorithm for micro aggregation.” *Journal of Applied Statistics* 42.4 (2015): 846-865.
- [48] Zhang Ryoung-tak “Artificial intelligence and medicine.” 춘 추계 학술대회 (KASL) 2016.2 (2016): 65-65.
- [49] Arsene, Octavian, Ioan Dumitrache, and Ioana Miha. “Expert system for medicine diagnosis using software agents.” *Expert Systems with Applications* 42.4 (2015): 1825-1834.
- [50] Hengstler, Monika, Ellen Enkel, and Selina Duelli. “Applied artificial intelligence and trust—The case of autonomous vehicles and medical assistance devices.” *Technological Forecasting and Social Change* 105 (2016): 105-120.
- [51] West, Jarrod, and Maumita Bhattacharya. “Intelligent financial fraud detection: a comprehensive review.” *Computers & Security* 57 (2016): 47-66.
- [52] Gravenhorst, Franz, et al. “Mobile phones as medical devices in mental disorder treatment: an overview.” *Personal and Ubiquitous Computing* 19.2 (2015): 335-353.
- [53] Šabanović, Selma, et al. “A robot of my own: participatory design of socially assistive robots for independently living older adults diagnosed with depression.” *International Conference on Human Aspects of IT for the Aged Population*. Springer International Publishing, 2015.
- [54] Inbarani, H. Hannah, Ahmad Taher Azar, and G. Jothi. “Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis.” *Computer methods and programs in biomedicine* 113.1 (2014): 175-185.
- [55] Eiben, Agoston E., and Jim Smith. “From evolutionary computation to the evolution of things.” *Nature* 521.7553 (2015): 476-482.
- [56] Neill, Daniel B. “Using artificial intelligence to improve hospital inpatient care.” *IEEE Intelligent Systems* 28.2 (2013): 92-95.
- [57] Musen, Mark A., Blackford Middleton, and Robert A. Greenes. “Clinical decision-support systems.” *Biomedical informatics*. Springer London, 2014. 643-674.
- [58] Berner, Eta S., and Tonya J. La Lande. “Overview of clinical decision support systems.” *Clinical decision support systems*. Springer International Publishing, 2016. 1-17.
- [59] S.N. Deepa, B. Aruna Devi, “A survey on artificial intelligence approaches for medical image classification”, *Indian Journal of Science and Technology*, Vol. 4 No. 11 (Nov 2011).
- [60] Gupta, S., Sarkar, A., Pramanik, I. and Mukherjee, B. Implementation Scheme for Online Medical Diagnosis System Using Multi Agent System with JADE. *International Journal of Scientific and Research Publications*, Volume 2, Issue 6, June 2012.
- [61] Londhe, Vaishali. “Brain MR Image Segmentation for Tumor Detection using Artificial Neural.” *Brain* 6.1 (2017).
- [62] Pannu, Avneet. “Artificial intelligence and its application in different areas.” *Artificial Intelligence* 4.10 (2015).
- [63] Vassilis S Kodogiannis and John N Lygouras (2008) Neuro-fuzzy classification system for wireless capsule endoscopic images. *J. World Acad. Sci. Engg. & Technol.*, 45, 620-628.
- [64] Roberts, Kirk, et al. “State-of-the-art in biomedical literature retrieval for clinical cases: a survey of the TREC 2014 CDS track.” *Information Retrieval Journal* 19.1-2 (2016): 113-148.
- [65] Masethe, Hlaudi Daniel, and Mosima Anna Masethe. “Prediction of heart disease using classification algorithms.” *Proceedings of the world congress on engineering and computer science*. Vol. 2. 2014.
- [66] Patel, Ankeeta R., and Mandar M. Joshi. “Heart diseases diagnosis using neural network.” In *Computing, Communications and Networking Technologies (ICCCNT)*, 2013 Fourth International Conference on, pp. 1-5. IEEE, 2013.
- [67] Maltas, Ahmet, Ali Alkan, and Mustafa Karabulut. “Use of artificial neural network algorithm in the immunohistochemical dyeing based diagnosis of thyroid tumor.” In *Signal Processing and Communications Applications Conference (SIU)*, 2014 22nd, pp. 1106-1109. IEEE, 2014.
- [68] Filimon, Delia-Maria, and Adriana Albu. “Skin diseases diagnosis using artificial neural networks.” In *2014 IEEE 9th IEEE International Symposium on Applied Computational Intelligence and Informatics (SACI)*. 2014.
- [69] Ertl, L., and Christ, F., Significant improvement of the quality of bystander first aid using an expert system with a mobile multimedia device. *Resuscitation* 74:286–295, 2007.
- [70] M., Gulbandilar, E., and Cimbiz, A., Prediction of low back pain with two expert systems. *J. Med. Syst.* 36:1523–1527, 2012.
- [71] Kunhimangalam, R., Ovallath, S., and Joseph, P. K., A clinical decision support system with an integrated EMR for diagnosis of peripheral neuropathy. *J. Med. Syst.* 38:38, 2014.
- [72] Maizels, M., and Wolfe, W., An expert system for headache diagnosis: The computerized headache assessment tool (CHAT). *Headache* 48:72–78, 2008.
- [73] Elveren, E., and Yumusak, N., Tuberculosis disease diagnosis using artificial neural network trained with genetic algorithm. *J. Med. Syst.* 35:329–332, 2011.
- [74] Fisher, A. C., Chandna, A., and Cunningham, I. P., The differential diagnosis of vertical strabismus from prism cover test data using an artificially intelligent expert system. *Med. Biol. Eng. Comput.* 45: 689–693, 2007.
- [75] Bascil, M. S., and Temurtas, F., A study on hepatitis disease diagnosis using multilayer neural network with Levenberg Marquardt training algorithm. *J. Med. Syst.* 35:433–436, 2011.
- [76] Bascil, M. S., and Oztekin, H., A study on hepatitis disease diagnosis using probabilistic neural network. *J. Med. Syst.* 36:1603–1606, 2012.
- [77] Wells, D. M., and Niedere, J., A medical expert system approach using artificial neural networks for standardized treatment planning. *Int. J. Radiat. Oncol. Biol. Phys.* 41:173–182, 1998.
- [78] Bascifcti, F., and Incekara, H., Design of web-based fuzzy input expert system for the analysis of serology laboratory tests. *J. Med. Syst.* 36:2187–2191, 2012.
- [79] Lam, C. F. D., Leung, K. S., Heng, P. A., Lim, C. E. D., and Wong, F. W. S., Chinese acupuncture expert system (CAES): A useful tool to practice and learn medical acupuncture. *J. Med. Syst.* 36:1883– 1890, 2012.
- [80] Issac Niwas, S., Palanisamy, P., Chibbar, R., and Zhang, W. J., An expert support system for breast cancer diagnosis using color wavelet features. *J. Med. Syst.* 36:3091–3102, 2012.
- [81] Benali, R., Reguig, F. B., and Slimane, Z. H., Automatic classification of heartbeats using wavelet neural network. *J. Med. Syst.* 36: 883–892, 2012.
- [82] Exarchos, T. P., Tsiouras, M. G., Exarchos, C. P., Papaloukas, C., Fotiadis, D., and Michalis, L. K., A methodology for the automated creation of fuzzy expert systems for ischaemic and arrhythmic beat classification based on a set of rules obtained by a decision tree. *Artif. Intell. Med.* 40:187–200, 2007.
- [83] Kumar, S. J. J., and Madheswaran, M., An improved medical decision support system to identify the diabetic retinopathy using fundus images. *J. Med. Syst.* 36:3573–3581, 2012.
- [84] Ondiege, Brian, and Malcolm Clarke. “Healthcare professionals’ perception of security of Personal Health Devices.” (2017).
- [85] Khan, Iqbal Uddin, and Sadiq ur Rehman. “A Review on Big Data Security and Privacy in Healthcare Applications.” *Big Data Management*. Springer International Publishing, 2017. 71-89.

A Hybrid Curvelet Transform and Genetic Algorithm for Image Steganography

Heba Mostafa Mohamed

Information System
Faculty of Computer and Informatics
Suez Canal University
Ismailia, Egypt

Ahmed Fouad Ali

Computer Science
Faculty of Computer and Informatics
Suez Canal University
Ismailia, Egypt

Ghada Sami Altaweel

Computer Science
Faculty of Computer and Informatics
Suez Canal University
Ismailia, Egypt

Abstract—In this paper, we present a new hybrid image steganography algorithm by combining two famous techniques which are curvelet transform and genetic algorithm GA. The proposed algorithm is called Hybrid Curvelet Transform and Genetic Algorithm for image steganography (HCTGA). Curvelet transform is a multiscale geometric analysis tool, its main advantage is that it can solve the important problems efficiently and other transforms are not that ideal. Genetic algorithm is a famous optimization algorithm with the aim of finding the best solutions to a given computational problem that maximizes or minimizes a particular function. In the proposed algorithm the cover and secret images are passed through a preprocessing process by applying four different filters to them in order to remove the noise and achieve a better quality to both images before the hiding process. Then the curvelet transform is applied to the cover image to find the curvelet frequencies of the image, and the secret image is hidden at the Least Significant Bits (LSB) of the curvelet frequencies of the cover image to reconstruct the stego image. Finally genetic algorithm operations are employed to find different scenarios for the hiding process by rearranging the hiding bits and finally choose the best scenario that can reach a better image quality and a higher Peak Signal to Noise Ratio (PSNR) results.

Keywords—Image steganography; curvelet transform; least significant bits; genetic algorithm; Peak Signal to Noise Ratio (PSNR)

I. INTRODUCTION

Because of the continuous progress in the communication technologies, it becomes necessary to protect the important information sending through any communication facility especially the internet. So, image steganography which is the art of concealing vital data inside a cover which can be image, video or audio, becomes one of the most important fields.

There are many steganographic techniques that are used and developed through times. For example, substitution technique by least significant bits LSB [1], which is the simplest data hiding technique, it works by changing the least significant bits of the cover image by the secret information. This technique is simple and effective and it doesn't influence the quality of the images, however it is good at resisting the steganalysis attacks when it is used alone.

Transform domain technique which consists of different transforms such as: Discrete Cosine Transform (DCT) and Discrete wavelet transform (DWT). In DCT [2], the cover

medium is usually converted to its frequency domain and secret data is hidden into the cover image frequencies without causing any significant changes in the cover image.

Discrete wavelet transforms DWT, is the most famous technique in the steganography field. Wavelet has been the most important technique used in this field, because of its ability to hide the secret images without affecting the quality of the image and because of its robustness against many steganalysis attacks, it works by converting the domains from spatial to frequency domains and it can be used in steganography by isolating the high frequencies from the low frequencies on each pixel, so the image is decomposed into two bands, these bands are detailed and approximation bands which referred to as the sub-bands [3]. Detailed band is the band that contains the high frequency components of images in which the insignificant features of the image exist. Approximation band contains the low frequency components; it contains all the significant features of the image. The important information that defines the image usually exists in the approximation band. So in image steganography, the secret information is usually hidden in the detailed band.

Wavelet transform is the best in capturing zero-order singularities which called point singularities of the image and it cannot deal with the features along edges [4]. Since there are images with higher order singularities, wavelet transform won't have points of interest in such manner. To defeat the shortcoming of wavelet transform in the high dimensions, and to deal with curves better, Candes and Donoho proposed curvelet transform [5].

Spread spectrum technique [6], in which the secret information is distributed over a wide frequency bandwidth. Hence, it is very difficult to completely remove the message without destroying the cover image.

In statistical technique [6], the secret messages are encoded by the change of some numerical properties of the cover image.

Distortion technique [6], in this method the secret messages are stored by the distortion of signals. The sender usually makes a specific change to the cover image and the receiver recovers the secret information by recording the differences between the original and the stego images.

In this paper, we propose a novel image steganography algorithm that merges least significant bit and curvelet

transform to hide the secret messages in the curvelet frequencies of the cover image for a better image quality, and we employ the genetic algorithm technique, to choose the best embedding plan.

The paper is organized as follows: In Section 2; we give a brief description of the related work done on the steganography area. In Section 3, we highlight the principles of the curvelet transform and its implementation techniques. In Section 4, we explain how the genetic algorithm works and we give a brief description of its famous operations. Section 5 contains a detailed description of our proposed algorithm which called Hybrid Curvelet Transform and Genetic Algorithm for image steganography (HCTGA). In Section 6, 50 cover and secret images are tested and analyzed by the new algorithm and the numerical results of the experiments are recorded, then a comparison between our proposed algorithm and other famous techniques is done. Finally, a brief summary and conclusion is given.

II. RELATED WORK

In 2007, Yuan-Yu Tsai and Chung-Ming Wang [7] have applied a data hiding criteria for color images by using the Binary Space Partitioning (BSP tree). It results in high capacity steganography system with low visual distortion. They found that using the tree can enhance the security, making it troublesome for any unapproved user to extract the secret messages without the known of the secret key.

In 2008, L.Y. Por, W.K. Lai, Z. Alireza, T.F.Ang, M.T.Su and B.Delina, [8] have integrated three LSB insertion algorithms in one steganography system. The unauthorized users can't detect the secret data during a communication because this method implements a public key infrastructure that is only the sender and the receiver can know it.

In 2008, M. A. Bani Younes and A. Jantan [9] introduced a steganography technique that depends on hiding secret data inside an encrypted image by using the LSB insertion. In this approach, they randomly overwrite the LSB of the encrypted image by the binary representation of the secret data. The receiver can easily extract the secret data by applying the inverse of the transformation and encryption processes to the stego image.

In 2008, Chang-ChuChen and Chin-Chen Chang [10] have introduced a steganography method that modifies the regular steganography based LSB method by the use of reflected gray code rule. In this method, the hiding criteria and the distortion level are similar to the simple LSB method, but the two methods differs in the change of the secret bits before and after embedding; the LSB substitution keeps them equally, unlike this method where the stego and the secret images are not always the same.

In 2009, Babita Ahuja and Manpreet Kaur [11] have proposed a steganography technique using LSB method with the aim of providing high capacity for the hiding process. In this technique four least significant bits are replaced by the secret data, and a filtering criterion and two level high securities are employed to remove the distortions that can cause suspicions.

In 2010, Ekta Walia, Payal Jain and Navdeep [12] have analyzed two different schemes of steganography the least significant bit (LSB) and the discrete cosine transform (DCT), and they compared the two methods and found that steganography by using DCT method is better than steganography by using LSB, because the DCT method provides a better image quality with little image distortion as compared to LSB. Although LSB can hide a big amount of secret data way more than the amount of secret information that can be hidden using DCT technique.

In 2011, E. Ghasemi, J. Shanbehzadeh and N. Fassihi [13] have applied the wavelets and genetics based mapping function in a new steganographic technique. The frequency domain where the secret message is embedded improves the robustness of the steganography process and genetic algorithm is used to acquire the best mapping function that reduces the errors between the cover and the reproduced images, to enhance the capacity of the hiding process.

In 2012, M. R. Garg and M. T. Gulati [1] focuses on applying steganography by using two methods, Least Significant Bit (LSB) and Most Significant Bit (MSB), and they compared the two methods and found that LSB based steganography is much better than MSB based steganography for hiding the message.

In 2015, D. Babya, J. Thomasa, G. Augustine, E. Georgea and N.R. Michaela [14] have employed the discrete wavelet transform DWT to create a new steganography method, in which the cover is divided into three planes R, G and B planes and the secret messages are inserted into the frequency domain of these planes. They found that the resulted image has high general security and a less recognizable distortion when compared to the original cover image.

In 2015, Z. Yin and B. Luo [15] have proposed a large capacity steganography technique in light of two methods, the pixel pair matching and the modification direction exploitation (MDE). They modified the cover pixel pair implying one or two 9-ary digits as indicated by various payloads. Experimental results of this method show high embedding capacity in addition to great quality and security.

In 2016, M. Saidi, H. Hermassi, R. Rhouma and S. Belghith [16] have introduced a steganography technique depending on chaotic map in DCT domain. In this method the DCT is applied on the original image, the coefficients is scanned in a zigzag order from the least to the most significant bits, as a result of this scan, the embedding positions are determined by the chaotic function in addition to the maximum allowed payload. The results of the experiments demonstrate that this algorithm gives a great imperceptibility and flexibility.

III. CURVELET TRANSFORM

Curvelet transform is a multiscale geometric analysis tool most appropriate for objects with curves. Candes and Donoho created this transform [5] with a goal of representing edges along curves more efficient than the traditional transformations. It differs from other transforms in that aside from location and scale, orientation decomposition is likewise applied for the signals; along these lines the coefficients acquired by the

curvelet show location, scale and angle data of the image features [4].

Curvelet transform is the best in representing the curves in images because it takes image edges as the essential representation elements [4]. The experiments of many researchers proved that the curvelet's small scale coefficients can beat the wavelet's high frequency coefficients, because of the orientation sensitive property of the curvelet coefficients.

Curvelet transform can be applied by two methods [17]:

Using Wrapping method.

Using Fast Fourier Transformation algorithm (FFT).

The two implementations of curvelet basically differ by the spatial grid choice that is utilized to interpret curvelet coefficients at every angle and scale. Both ways give back a table of digital curvelet coefficients listed by spatial location, orientation and scale parameters [17].

In this paper, we used the second method to apply the curvelet, cover image is transformed using the Fast Fourier Transformation algorithm and the secret message was embedded with that transformed image. Then the inverse FFT is applied to propose the stego image.

We choose FFT algorithm because it is very effective and very easy to study its function as it can be expressed as a sum of series of sine and cosines.

The benefits of using FFT algorithm can be listed and summarized as follows [17]:

- Elements of digital contents can be used directly and the secret data can be embedded in them directly.
- Elements that changes over time should be transformed into the frequency domain before the embedding process, this makes processing easier.
- The changes of the quality of digital content before and after embedding are minimized.
- Processing required for steganography and detection is simple.
- The secret information embedded in digital content can be detected as required.

IV. GENETIC ALGORITHM

Genetic Algorithm is a famous optimization algorithm proposed by John Holland in 1975 [18], its fundamental goal is to discover the optimal solutions of a given computational problem that maximizes or minimizes a particular function. Genetic algorithm based steganography developed to generate several stego images or several solutions by inserting the secret message in different blocks of the cover and choose the order that leads to the best PSNR value.

The steps of the GA process are listed below and summarized in Fig. 1:

- 1) Define the initial population with a specific number of individuals.

- 2) Produce the next generations by applying the GA operations such as crossover and mutation.
- 3) Select the best solutions by using the fitness function and repeat Step 2 to produce better solutions, and continue until the best solution is reached.

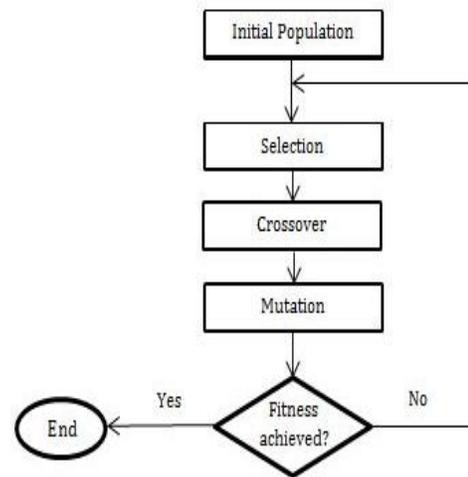


Fig. 1. Genetic algorithm process.

So, the main process of GA can be described as follows:

The initial population, selection process, crossover operation, mutation operation and fitness function.

A. Initial Population

The basic term of the GA process is the chromosome which is the numerical value that represents an individual or a candidate solution to the problem. Chromosomes consist of discrete units referred to as genes. Every gene represents the chromosome features.

A group of chromosomes is referred to as a population. GA usually begins with a randomly chosen arrangement of chromosomes, which serves as the initial population.

B. Selection Process

In this process, two chromosomes are selected to produce two new offsprings. Generally, the fitness function of every individual decides the likelihood of its existence in the next generations. There exist many distinctive selection strategies in genetics, for example, tournament selection, ranking, and proportional selection.

C. Crossover Operation

This operation is the most vital operation in genetic algorithm. In this operation, usually two parent chromosomes are consolidated together and produce new chromosomes, referred to as *offsprings*. This is can be done by swapping the genes of the chosen parents to produce new solutions as shown in Fig. 2.

The parents are usually selected among other chromosomes according to their fitness value so that the produced offsprings are required to acquire the great genes that can make better generations.

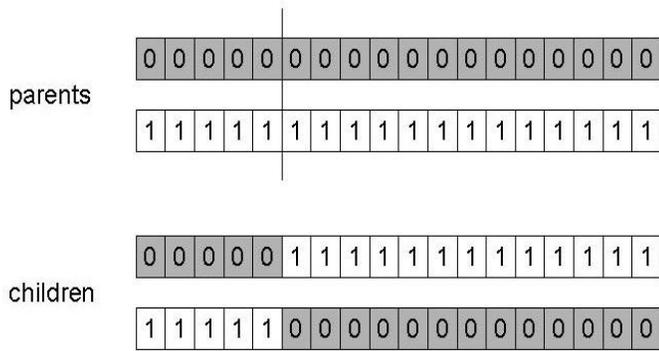


Fig. 2. Example of the crossover operator.

D. Mutation Operation

Mutation operator works by making random changes to the attributes of the chromosome. It is usually applied to the gene level; it works by randomly flips individual bits in the new chromosomes.

The GA algorithm can get stuck at a local optimum before finding the global optimum. The mutation operator helps protect against this problem by maintaining diversity in the population, but it can also make the algorithm converge more slowly.

E. Fitness Function

This function is one of the most vital steps of the GA process as it determines how the chromosomes will change over time and it determines the direction that the population will take. Peak Signal to Noise Ratio (PSNR) is a famous fitness function that is utilized widely in the image processing field. It is generally used to analyze quality of the image by describing the similarity degree between images. When the PSNR function gives higher results, this is indicating a better quality for the resulted image.

V. PROPOSED ALGORITHM (HCTGA)

The proposed algorithm presents a new image steganography algorithm that combines three famous and effective techniques; which are the discrete curvelet transform, the least significant bit (LSB) [19] and uses the genetic algorithm to achieve a better image quality and security.

A. Input Images

The proposed algorithm uses two input images with size 512 x 512. One of them is referred to as the cover image into which the required secret messages are embedded. The other is referred to as the secret image that needs to be hidden inside the cover image; it is also called the payload.

We test our proposed algorithm on 30 cover and secret images.

B. Preprocessing Step

This step is very important step in any image processing technique; its main goal is to enhance the image data to get rid of the unwanted distortions or enhance the features that is important for the next operations.

In our method, a preprocessing technique is done by applying four famous filters to the cover and the stego images, these filters are *gaussian*, *average*, *motion*, *unsharp*, each filter plays a great role in removing the noise from images while keeping its features and enhancing the images quality before applying the steganography process.

C. Embedding Process

In this step, the steganography process which is the most important step in our algorithm is applied as follows:

First the cover is divided into 16 equally blocks, each block consists of 128 x 128 pixels.

Then the discrete curvelet transform DCT is performed into the cover image. The DCT can be performed by the steps shown in Fig. 3 as produced by J. Starck, E. Candes, and D. Donoho [20]:

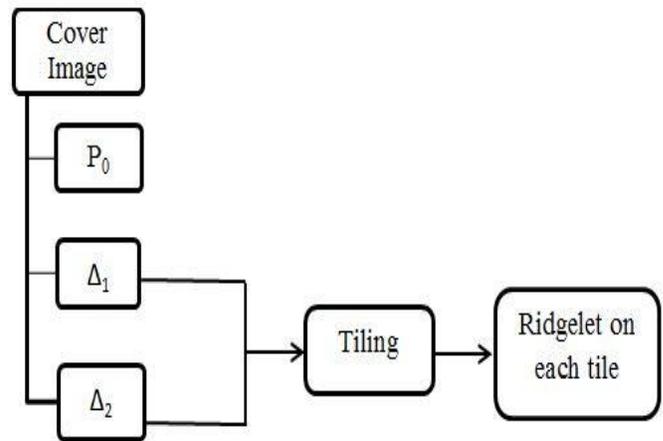


Fig. 3. Curvelet decomposition.

1) Subband filtering

The cover is divided into three subbands P_0 , Δ_1 and Δ_2 (see Fig. 4).

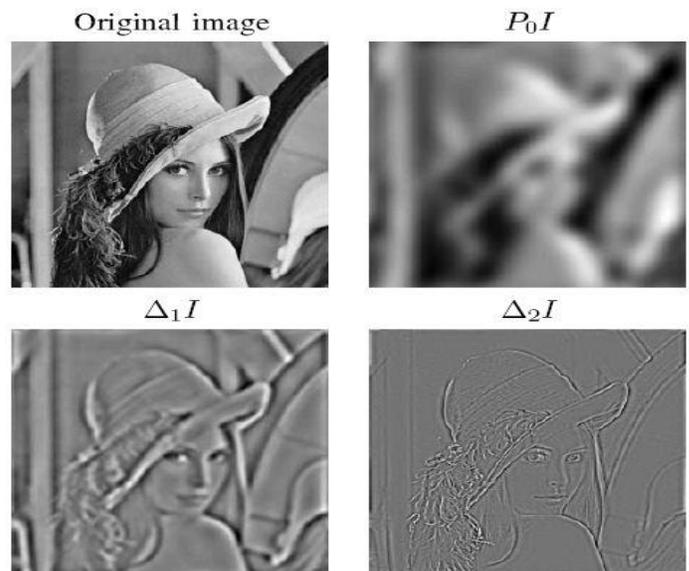


Fig. 4. Curvelet subbands.

2) Tiling

The subbands $\Delta 1$ and $\Delta 2$ are tiled to make the curved edges become linear singular, so that the Ridgelet transform can handle it well. Fig. 5 illustrates the tiling process.

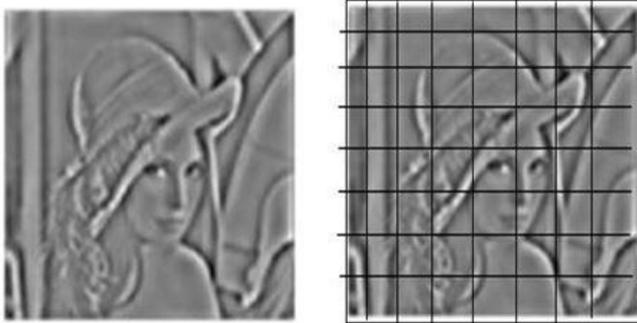


Fig. 5. Smooth partitioning (tiling).

3) Ridgelet transform

Applying the Discrete Ridgelet transform is the final step of the process.

After applying the curvelet to the cover, the secret message should be embedded into the LSB of the curvelet subbands.

D. Applying Genetic Algorithm GA

The genetic algorithm is utilized to insert the secret image by different ways and scenarios and test the quality of the image at every scenario to choose the one that result in a better image quality. The cover image is divided into blocks and the secret data are embedded at a different arrangement of these blocks every time to create different embedding scenarios.

Two GA operators are used: Crossover and Mutation operators. We used the crossover operation to merge two scenarios together to create a new better one. The mutation operation is utilized to change some data when the resulted scenarios become similar to each other.

Finally, the stego image is generated by applying the Inverse Curvelet Transform on the modified coefficients.

To test the quality of the reconstructed image, PSNR is used as the GA fitness function. It works by computing the difference between the original image and the resulted stego image by applying the following equation:

$$PSNR = 20 \log_{10} \frac{255}{\sqrt{MSE}} \quad (1)$$

Where the MSE is the Mean Square Error that computes error difference between two images by the following equation:

$$MSE = \frac{\sum_{i=1}^{255} \sum_{j=1}^{255} (Cover(i,j) - Stego(i,j))^2}{255 \times 255} \quad (2)$$

The steps of our proposed algorithm are summarized in Algorithms 1 and 2 and Fig. 6.

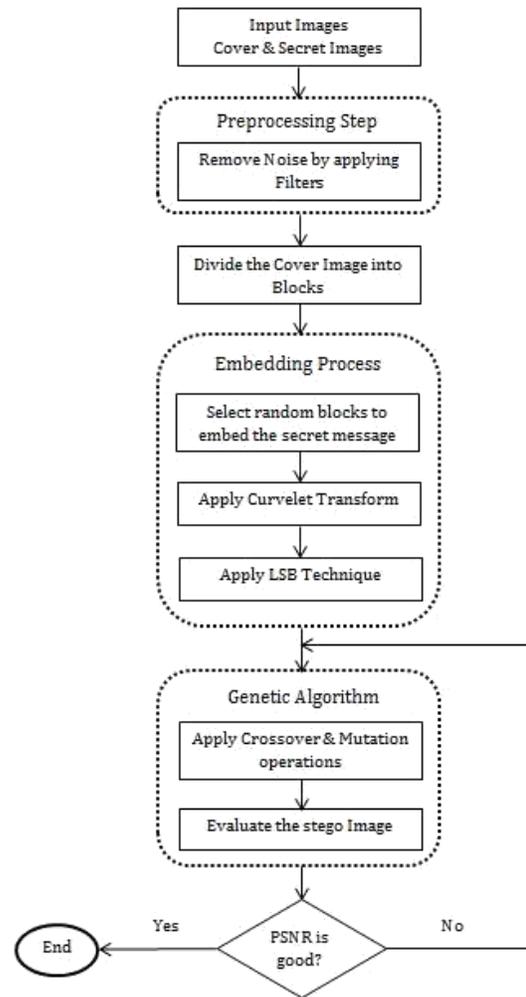


Fig. 6. Proposed HCTGA algorithm.

ALGORITHM 1: EMBEDDING ALGORITHM OF THE PROPOSED HCTGA

Algorithm 1 The Embedding Algorithm of the Proposed HCTGA

- a. Get the cover image with size 512 x 512. {Input cover image}
- b. Apply the denoising technique. {Preprocessing operation}
- c. Divide the cover & secret images into 16 Blocks.
- d. Rearrange the cover image blocks randomly.
- e. The curvelet transform is applied on the cover image.
- f. Embed the secret image by using LSB technique. {Hide secret image into cover image}
- g. Rearrange the cover image blocks to the original pattern.
- h. Apply the GA operations such as Crossover & Mutation, to generate a better stego images.
- i. Apply the fitness function (PSNR) that can measure the stego image quality.
- j. Choose the best stego image.

VI. EXPERIMENTAL RESULTS

A. Parameter Settings

There are three important parameters in the GA process, which are the population size, the crossover probability and the mutation probability. The values of these parameters are listed in Table 1.

B. General Performance of HCTGA

The performance of the proposed algorithm is shown in Fig. 7 and Table 2.

C. Comparison between HCTGA Algorithm and other Algorithms for Image Steganography

We compared the proposed HCTGA algorithm with three famous steganography techniques to test the efficiency of our algorithm. These methods are: the LSB, the wavelet transforms and the curvelet transform method without using genetic algorithm.

1) LSB

Least significant bit (LSB) is a spatial domain technique; it is usually used for steganography for hiding the secret messages by replacing the cover's least significant bits by the secret image bits. This technique makes steganography very easy mission to accomplish, it is fast, effective, enables high capacity embedding and the changes in the cover after steganography would be unnoticeable. The drawback of this technique is its weakness against attacks, but this is can be solved by using other image transformation with LSB.

2) Wavelet Transform

It is a frequency domain transform; steganography based wavelet transform was the most popular and frequently used technique over the last few years. It works by converting the cover from spatial to frequency domains, then hiding the secret message in the wavelet coefficients of this cover image [3].

Steganography based wavelets is very effective method, because it is resistible against the attacks unlike the LSB, and it provides a better capacity and robustness for the steganography process.

The drawback of wavelet transform is its representation of image edges and higher order singularities. To represent edges by wavelet transform, too many wavelet coefficients are needed to repeat edges at scale after scale in order to reconstruct the edges properly [4]. So it is not the best method for capturing higher order singularities for images. To overcome the shortage of wavelet transform in higher dimensions, and to represent the curves better, Candes and Donoho proposed curvelet transform [5].

3) Curvelet Transform

The steganography based curvelet was described in Section 3.

TABLE. I. GA PARAMETERS

Pop Size	16
Probability of Crossover Pc	Pc = 1
Probability of Mutation Pm	Pm = 0.001

TABLE. II. PERFORMANCE OF THE PROPOSED ALGORITHM

Test No.	MSE	PSNR
1	6.47	40.02
2	7.10	39.62
3	8.21	38.99
4	7.58	39.33
5	7.52	39.37
6	6.91	39.74
7	6.37	40.09
8	7.12	39.61
9	8.31	38.93
10	10.53	37.91
11	5.97	40.37
12	6.29	40.14
13	6.75	39.84
14	5.37	40.83
15	9.33	38.43
16	9.19	38.50
17	7.99	39.11
18	7.12	39.61
19	5.11	41.05
20	8.97	38.60
21	4.98	41.16
22	5.61	40.64
23	8.41	38.88
24	7.95	39.13
25	7.13	39.60
26	6.29	40.14
27	6.59	39.94
28	6.91	39.74
29	7.10	39.62
30	7.37	39.46





Fig. 7. The performance of HCTGA Algorithm (a) cover image, (b) secret image, (c) stego image.

We apply the four methods on 10 images and the PSNR values of the methods are calculated and recorded in Table 3 and Fig. 8. This comparison illustrates that the proposed HCTGA algorithm can obtain better image quality than the other compared methods.

TABLE III. COMPARISON BETWEEN LSB, WAVELET, CURVELET AND THE PROPOSED HCTGA ALGORITHM

Test No.	LSB	Wavelet	Curvelet	HCTGA
1	29.12	31.32	40.01	40.02
2	28.57	30.07	37.13	39.62
3	21.01	29.49	37.24	38.99
4	28.10	32.18	38.44	39.33
5	20.39	30.21	39.31	39.37
6	19.34	21.11	38.93	39.74
7	21.81	25.98	39.22	40.09
8	21.52	25.65	38.67	39.61
9	29.31	27.28	38.81	38.93
10	23.89	33.41	37.17	37.91

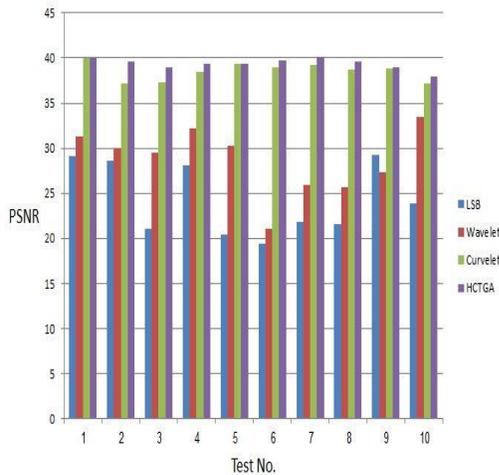


Fig. 8. Comparison between LSB, Wavelet, Curvelet and the proposed HCTGA algorithm.

VII. CONCLUSION AND FUTURE WORK

A new hiding algorithm has been illustrated in this paper by using the curvelet transform and genetic algorithm. Some filters are applied to the cover image as a preprocessing technique to remove the noise and enhance the quality of the images. Discrete curvelet transform is applied to cover images and the secret messages had been embedded in the curvelet coefficients by using LSB technique. Finally the genetic algorithm had been used to choose the best embedding criteria which lead to a high performance and unsuspected stego images.

Our proposed HCTGA algorithm has been tested and analyzed over fifty cover and secret images to test the performance of the proposed method and the results were promising.

Applying other evolutionary optimization algorithms besides genetic algorithm, such as, Differential Evolution (DE), Scatter Search (SS), Tabu Search (TS), Neighborhood search (NS) and Simulates Annealing (SA), is the main idea of the future work.

REFERENCES

- [1] M. R. Garg, M. T. Gulati, Comparison of Lsb and Msb Based Steganography in Gray-Scale Images, International Journal of Engineering Research and Technology (Vol. I, NO. 8 October 2012). ESRSA Publications, 2012.
- [2] H. Patel and P. Dave, Steganography Technique Based on DCT Coefficients, International Journal of Engineering Research and Applications, 2(1), 713-717, 2012.
- [3] P. Rajkumar, R. Kar and H. Dharmasa, A Comparative Analysis of Steganographic Data Hiding within Digital Images, International Journal of Computer Applications, 53(1), 1-6, 2012.
- [4] J. Zhang and W. Yinghui, A comparative study of wavelet and curvelet transform for face recognition, Image and Signal Processing (CISP), 2010 3rd International Congress on. Vol. 4. IEEE, 2010.
- [5] E. Candes, L. Demanet, D. Donoho and L. Ying, Fast discrete curve let transforms, Multiscale Modeling & Simulation, 5(3), 861-899, 2006.
- [6] H.M. Reddy and K.B. Raja, High capacity and security steganography using discrete wavelet transform, International Journal of Computer Science and Security (IJCSS), 3(6), 462, 2009.
- [7] Yuan-Yu Tsai, Chung-Ming Wang "A novel data hiding scheme for color images using a BSP tree". Journal of systems and software, vol.80, pp. 429-437, 2007.
- [8] L. Y. Por, W. K. Lai, Z. Alireza, T. F. Ang, M. T. Su, B. Delina, "StegCure: A Comprehensive Steganographic Tool using Enhanced

- LSB Scheme,” Journal of WSEAS Transactions on Computers, vol. 8, pp. 1309-1318, 2008.
- [9] M. A. Bani Younes and A. Jantan, “A New Steganography Approach for Images Encryption Exchange by Using the Least Significant Bit Insertion,” International Journal of Computer Science and Network Security, vol. 8, no. 6, pp.247-257, 2008.
- [10] Chang-Chu Chen, and Chin-Chen Chang, “LSB-Based Steganography Using Reflected Grey Code,” The Institute of Electronics, Information and communication Engineers Transaction on Information and System,” vol. E91-D (4), pp. 1110-1116, 2008.
- [11] B. Ahuja and, M. Kaur, “High Capacity Filter Based Steganography,” International Journal of Recent Trends in Engineering, vol. 1, no. 1, pp.672-674, May 2009.
- [12] E. Walia, P. Jain, Navdeep, “An Analysis of LSB & DCT based Steganography”, Global Journal of Computer Science and Technology, Vol. 10 Issue 1 (Ver 1.0), April 2010.
- [13] E. Ghasemi, J. Shanbehzadeh and N. Fassihi, “High Capacity Image Steganography using Wavelet Transform and Genetic Algorithm”, International multi conference of engineers and computer scientists, 2011.
- [14] D. Babya, J. Thomasa, G. Augustine, E. Georgea and N. R. Michaela, “A Novel DWT based Image Securing Method using Steganography”, International Conference on Information and Communication Technologies (ICICT 2014), 2015.
- [15] Z. Yin and B. Luo, “MDE-based image steganography with large embedding capacity”, SECURITY AND COMMUNICATION NETWORKS, 2015.
- [16] M. Saidi, H. Hermassi, R. Rhouma, S. Belghith, “A new adaptive image steganography scheme based on DCT and chaotic map”, Multimed Tools Appl (2016).
- [17] Al-Ataby, A. A., & Al-Naima, F. M. “High capacity image steganography based on curvelet transform”. *Developments in E-systems Engineering (DeSE)*, 2011.
- [18] J.H. Holland, “Adaptation in Natural and Artificial Systems”, Univ. of Michigan Press, Ann Arbor, Mich., 1975.
- [19] H. Mostafa, A.F. Ali, and G. Eltaweal, “Hybrid curvelet transform and least significant bit for image steganography”, 2015.
- [20] J.L. Starck, E. Candes, and D. Donoho “The Curvelet Transform for Image Denoising”, IEEE Trans. Image Processing , Vol. 11, PP. 670-684, June 2002.

Automatic Music Genres Classification using Machine Learning

Muhammad Asim Ali

Department of Computer Science
SZABIST
Karachi, Pakistan

Zain Ahmed Siddiqui

Department of Computer Science
SZABIST
Karachi, Pakistan

Abstract—Classification of music genre has been an inspiring job in the area of music information retrieval (MIR). Classification of genre can be valuable to explain some actual interesting problems such as creating song references, finding related songs, finding societies who will like that specific song. The purpose of our research is to find best machine learning algorithm that predict the genre of songs using k-nearest neighbor (k-NN) and Support Vector Machine (SVM). This paper also presents comparative analysis between k-nearest neighbor (k-NN) and Support Vector Machine (SVM) with dimensionality return and then without dimensionality reduction via principal component analysis (PCA). The Mel Frequency Cepstral Coefficients (MFCC) is used to extract information for the data set. In addition, the MFCC features are used for individual tracks. From results we found that without the dimensionality reduction both k-nearest neighbor and Support Vector Machine (SVM) gave more accurate results compare to the results with dimensionality reduction. Overall the Support Vector Machine (SVM) is much more effective classifier for classification of music genre. It gave an overall accuracy of 77%.

Keywords—K-nearest neighbor (k-NN); Support Vector Machine (SVM); music; genre; classification; features; Mel Frequency Cepstral Coefficients (MFCC); principal component analysis (PCA)

I. INTRODUCTION

Nowadays, a personal music collection may contain hundreds of songs, while the professional collection usually contains tens of thousands of music files. Most of the music files are indexed by the song title or the artist name [1], which may cause difficulty in searching for a song associated with a particular genre.

Advanced music databases are continuously achieving reputation in relations to specialized archives and private sound collections. Due to improvements in internet services and network bandwidth there is also an increase in number of people involving with the audio libraries. But with large music database the warehouses require an exhausting and time consuming work, particularly when categorizing audio genre manually. Music has also been divided into Genres and sub genres not only on the basis on music but also on the lyrics as well [2]. This makes classification harder. To make things more complicate the definition of music genre may have very well changed over time [3]. For instance, rock songs that were made fifty years ago are different from the rock songs we have today. Luckily, the progress in music data and music recovery has considerable growth in past years.

According to Aucouturier and Pachet, 2003 [4] genre of music is possibly the best general information for the music content clarification. Music engineering encourages the practice of categories and family based operators like to organize their sound accumulations by this clarification, so the requirement of involuntary organization of audio files into categories improved extensively. In addition, the latest improvements in category organization here are still an issue to accurately describe a type, or whether mostly rely on a consumer understands and flavor.

In order to establish and explore increasing composition groups we implemented an automatic technique that can be used for data mining for valuable data about audio composition direct from the audio file. Such data could incorporate rhythm, tempo, energy distribution, pitch, timbre, or other features. Most of the classifications depend on spectral statistical features timbre. Content collections relating to further musicological contents such as pitch and rhythm are too suggested, however their execution time is very less and furthermore they are closed by tiny info collections pointing at different audio arrangements. The inadequateness of audio descriptors will positively have a limitation on music categorization methods.

In this paper, we use machine learning algorithms, including k-nearest neighbor (k-NN) [5] and Support Vector Machine (SVM) [6] to classify the following 10 genres: blues, classical, rock, jazz, reggae, metal, country, pop, disco and hip-hop. In addition, we perform a comparative analysis between k-nearest neighbor (k-NN) [5] and Support Vector Machine (SVM) [6] with and without dimensionality reduction via principal component analysis (PCA) [7]. The k-nearest neighbor is automatically non-linear, and it can sense linear or non-linear spread information. It inclines to do very well with a lot of data points. Support Vector Machine can be used in linear or non-linear methods, once we have a partial set of points in many dimensions the Support Vector Machine inclines to be very good because it easily discovers the linear separation that should exist. Support Vector Machine is good with outliers as it will only use the most related points to find a linear separation (support vectors).

In our research we used Mel Frequency Cepstral Coefficients (MFCC) [8] to extract information from our data as prescribed by past work in this field [9].

II. LITERATURE REVIEW

The prominence of programmed music genre classification has been developing relentlessly for as far back as couple of years. Many papers have proposed frameworks that either model songs as a whole or utilize SVM to build models for classes of music. Below some of the related work is mentioned.

Kris West and Stephen Cox [10] in 2004 prepared a confounded classifier on many sorts of sound elements. They demonstrated capable outcomes on 6-way type characterization errands, with almost 83% grouping precision on behalf of their greatest framework. As indicated by them the detachment of Reggae and Rock music was a specific issue for the component extraction plan which was assessed by them. They also shared comparative spectral characteristics as well as comparable proportions of harmonic to non-harmonic substance.

Aucouturier & Pachet [11] worked on single songs through Gaussian Mixture Model (GMM) [12] and utilize Monte Carlo procedures to assessment the KL divergence [13] among them. Their setup was focused on an audio information recovery structure where the situation is calculated in articulations of recovery accuracy. Authors did not utilize a propelled classifier, as their outcomes are positioned by k-NN. They conveyed some important component sets for a few models that we use in our examination, in particular the MFCC.

Li, Chan and Chun [14] recommend an alternate technique to concentrate musical example included in sound music by methods for convolutional neural framework. Their tests demonstrated that convolution neural network (CNN) has vigorous ability to catch supportive components from the deviations of musical examples with unimportant earlier information conveyed by them. They introduced a system to consequently extricate musical examples high-lights from sound music. Utilizing the CNN relocated from the picture data retrieval field their element extractors require insignificant earlier learning to develop. Their analyses demonstrated that CNN is a practical option for programmed highlight mining. Such revelation supported their hypothesis that the inherent attributes in the assortment of melodic data resemble with those of picture data. Their CNN model is exceedingly versatile. They also presented their revelation of the perfect parameter set and best work on using CNN on sound music type arrangement.

Xu, Maddage and Fang [15] mutually used SVM on events of brief time highlights from whole classes. They then sorted the edges in test melodies and after that they let the edges vote for the class of the whole melody. They said in spite of the fact that the test informational indexes they utilized as a part of their examinations they are not adequate to sum up the superior of both the features and the SVM classifier. It can be seen that musical score is measurably distinguishable with great execution (more than 85 %) with particularly fundamental three classes (i.e. a, b and c). The characterization multifaceted nature can be diminished by various leveled arrangement steps. By presenting CAMS they built the general execution by 3-4%. One of the disadvantages

of this framework is high computational many-sided quality in figuring distinctive feature orders for various arrangement steps.

Perdo and Nuno [13] used SVM on different record level components for speaker ID and speaker affirmation assignments. They showed the Symmetric KL difference based piece and moreover considered showing a record as a single full-covariance Gaussian or a mix of Gaussians. They approved this approach in speaker ID, confirmation, and picture arrangement errands by contrasting its execution with Fisher part SVM's. Their outcomes demonstrated that new technique for consolidating generative models and SVM's dependably beat the SVM Fisher portion and the AHS strategies. It regularly outflanks other grouping strategies for example, GMM's and AHS. The equivalent blunder rates are reliably better with the new piece SVM techniques as well. On account of picture grouping their GMM/KL divergence-based piece has the best execution among the four classifiers while their single full covariance Gaussian separation based portion beats most different classifiers. All these empowering demonstrate that SVM's can be enhanced by giving careful consideration to the way of the information being displayed. In both sound and picture errands they simply exploit earlier years of research in generative techniques.

Andres, Peter and Larsen [16] used short-time features to hold the information of the first flag and compact to such a point that small dimensional classifiers or relationship estimations can be functional. Most extraordinary conclusions have been set in brief time highlights which enter the data from a little measured window (much of the time 10ms - 30ms). In any case, as often as possible the outcome time probability is extent of minutes. They consider differing approaches for component blend and late data fusion for music type categorization. A novel element blend system, the AR model, is suggested and clearly overwhelms normally utilized mean change features.

Li and Ogihara [17] in their paper prescribe Daubechies Wavelet Coefficient Histograms (DWCHs) as a list of capabilities appropriate for categorization of music type. The list of capabilities outlines vastness contrasts in the sound flag. In this paper they proposed DWCHs, another feature extraction strategy for music genre grouping. DWCHs analyze music motions by registering histograms on Daubechies wavelet coefficients at different recurrence groups which has enhanced the arrangement accuracy. They gave a relative investigation of different feature extraction and grouping techniques and research the order execution of different characterization strategies on various feature sets.

An extensive assessment with mutually personal and substance created likeness calculation done through different types of questions [18]. They tended to the topic of contrasting distinctive present song comparability methods and furthermore elevated the interest for a typical assessment record.

A few other models have been made to take care of music genre classification with the million song dataset [19], which utilizes sound features and expressive features. The Model forms a sack of words for the expressive features. For the

sound features, they utilized the MFCC (Mel-recurrence cepstral coefficients) [20]. Their work was one of a kind by utilizing expressive features.

Similarly, another paper automatic musical genre classification of audio signals [21] in which a vector of size 9 (Mean-Centroid, Mean-Rolloff, Mean-Flux, Mean-Zero-Crossings, std centroid, std Rolloff, std Flux, std Zero-Crossings, Low-Energy) was utilized as their Musical-Surface Features vector. Musicality features were resolved and their model was assembled utilizing both the vectors.

A wide range of information is hidden inside a music waveform which ranges from auditory to perceptual [19]. In an experiment by Logan and Salomon [22] they organized playlists with the closest neighbors of a seed song. As indicated by them they depicted a technique to analyze songs construct exclusively in light of their sound substance. They assessed their separation measure on a database of more than 8000 songs. Preparatory goal and subjective outcomes demonstrated that their separation measure jam numerous parts of perceptual comparability. For the twenty songs judged by two clients they saw that all things considered 2.5 out of the main 5 songs returned are perceptually comparable. They additionally observed that their measure is powerful to basic humiliation of the sound.

Tzanetakis & Cook [21] also computed music related features arranging songs into genre with k-NN in view of GMMs prepared on music information. Authors basically had 100 capabilities routes for every class. They displayed these modules with GMMs requiring few segments in light of their mean utilization of feature measurements. As per the authors in spite of the fluffy way of genre limits, musical genre arrangement can be performed consequently with results altogether superior to possibility, and execution similar to humanoid type characterization. Three feature sets for speaking to tumbrel surface, rhythmic substance and pitch substance of music signs were suggested and were assessed utilizing measurable acknowledgment classifiers.

Gjerdingen and Perrott [23] investigated people to evaluate an excerpt and assign it to any one of 10 genre labels. The authors thought that the participants will be good in this task but the speed at the task was performed by the participants was as short as $\frac{1}{4}$ second which was unexpected.

Another review [24] led the investigations on song type characterization by 27 social audience members. Every person listened to focal thirty seconds of every song and be solicited to pick one out from six song types. These audience members accomplished between member genres understanding rate of just 76%. An arrangement of investigations looking at human and programmed musical genre grouping was exhibited. The outcomes demonstrate that there is noteworthy subjectivity in genre comment by people, and puts the consequences of programmed genre grouping into appropriate setting. Also, the utilization of computationally concentrated sound-related model didn't bring about enhanced outcomes contrasted with features figured utilizing MFCCs. These outcomes showed that there is huge bias in music type comment by people. That is, distinctive individuals arrange melody type in an unexpected way, prompting numerous irregularities.

Liu and Huang [25] in 2002 proposed another approach for substance based sound ordering utilizing GMM and represent another formula for separation calculation amongst 2 representations. Sound association strategies that contain non discourse signals have been prescribed. A large portion of these groupings point the arrangement of communicates audiovisual in general gatherings as audio, discourse, and ecological noises. The issue of judgment among song and discourse has set up huge consideration on or after the underlying effort of [26] where straightforward method of the normal zero-intersection level and vitality structures is utilized compare to the effort of [27] where different structures and measurable example acknowledgment classifiers are admirably assessed. The multidimensional classifiers manufactured gave an amazing and powerful segregation amongst discourse and music motions in computerized sound.

In another experiment by Kimber and Wilcox [28] sound signs were portioned and ordered into "music", "discourse," "giggling," and non-discourse sounds. An exploratory run founded framework for the division and association of sound signs from motion pictures. This visual-based preparing frequently prompted a very fine division of the varying media succession concerning the semantic significance of information. For instance, in the video grouping of a song execution there might be shots showing up group of the artists, a band, gathering of people and some other outlined perspectives. As indicated by the visual data, these shots will be filed independently.

Boyce, Li and Nestler [29] managed a more tough issue of finding performing voice sections in musical signs. In their framework a programmed discourse acknowledgment association is utilized as the feature vector for arranging singing portions.

Experiments by Tzanetakis and Cook [21], and Foote and Uchihashi [30] components were figured particularly on the substantial time-scale. They attempt to get the perceptual hits in the melody which creates them primitive and easy to check alongside the melody. Amazingly, brief time highlights must be attempted roundabout through e.g. their execution in a course of action undertaking. The paper also explains that much of the time executed via the mean and fluctuation of the brief timeframe highlights over the decision time horizon (cases are [31], [32] and [33]). However, the question is the measure of the applicable element stream they can get as an attempt to get the components of the brief span highlights.

Mckinney and Breedbaart [34] uses an otherworldly decay of the MFCC into four assorted repeat gatherings. An alternative method by Lu and Zhang [35] precedes the extent of characteristics overhead and steady circumstances the mean as the long haul highlight. Their brief span elements are zero intersection rate and brief time energy.

Anders, Peter and Larsen [36] in their experiment proposed a new model called the AR Model for genre classification which outperformed the commonly known mean-variance features. They investigated the decision of genre classification by short time feature integration.

Jonathan and Shingo [37] in 2001 introduced a method of beat spectrum to analyze the tempo and rhythm of audio and music. They found that high structure will have strong spectrum peaks which would help to reveal the tempo and relative strength of different beats. With this they were able to distinguish between different kinds of rhythms.

Li and Khokhar [38] utilized the comparable dataset to relate numerous arrangement strategies and data groups then offered the utilization of the nearby feature line design grouping procedure.

Scheirer [39] characterized a continuous beat following order for sound signs. For this grouping, a filter bank is connected with a system of brush channels that track the flag periodicities to convey an assessment of the primary beat and its quality.

III. DATA GATHERING

Music Analysis, Retrieval, and Synthesis for Audio Signals (Marsyas) is an open source World Wide Web for sound handling with particular complement on audio data uses. For our experiments we used GTZAN dataset which has a collection of thousand sound files. Each of the file is thirty seconds in length. Ten genres are present in this dataset containing hundred tracks each. Each track has 16-bit audio file 22050Hz Mono in .au format [40]. We have chosen ten genres: blues, classical, rock, jazz, reggae, metal, country, pop, disco and hip-hop. Our total data set was 1000 songs.



Fig. 1. MFCC flow.

V. ALGORITHMS

A. *K-Nearest Neighbour (k-NN)*

The first machine learning technique we utilized was the k-closest neighbors (k-NN) [5] as it is very famous for its simplicity of execution. The k-NN is by design non-linear and it can detect direct or indirect spread information. It also slants with a huge amount of data. The essential computation in our k-NN is to measure the distance between two tunes. We handled this by methods of the Kullback-Leibler divergence [10].

B. *Support Vector Machine (SVM)*

The second technique we used is the support vector machine [6] which is a directed organization method that discovers the extreme boundary splitting two classes of information. During this the information is not directly distinct in the feature space; if this is the case then they can be put into an upper dimensional space through method of Mercer kernel. Actually, the internal results of the information focuses in this higher dimensional space are essential, so the projection can

IV. MEL FREQUENCY CEPSTRAL COEFFICIENTS (MFCC)

It is used for audio handling. The earlier music classification studies directed us to MFCCs [8] as a methodology to characterize time domain waveforms as little frequency domain coefficients. To process the MFCC, we at first analyze the middle portion of the waveform and took 20ms diagrams at a parameterized break. For independent layout we used hamming window to smooth the points of time. After this, we proceeded with the Fourier change to develop the repeat modules. We then put the frequencies to the Mel scale which models human perspective of changes in pitch, which is generally immediate below 1kHz and logarithmic more than 1kHz. These mapping packs the frequencies into 20 containers by figuring triangle window coefficients in perspective of the Mel scale. Copying these by the frequencies and taking the log we then took the discrete cosine transform, which fills in as a figure of the Karhunen-Loeve transform to de-correlate the repeat fragments. Finally, we kept the underlying 15 of these 20 frequencies since higher frequencies are the purposes of point of interest that have a lesser degree an impact to human acknowledgment and contain less information about the melody. Finally, we displayed each uneven tune waveform as a grid of cepstral components where every section is a vector of 15 cepstral frequencies and 20ms plot for a parameterized number of edges per tune (see Fig. 1).

be understood if such an inner item can be figured straightforwardly.

The space of potential classifier tasks comprises of biased direct arrangements of key preparation occurrences in this kernel space [41]. The SVM training algorithm selects these weights and support vectors to improve the boundary amongst classifier boundary and training orders. Since training instances are specifically utilized in characterization, utilizing complete tracks as these samples supports very well with the issue of track taxonomy. SVM can be used in direct or indirect strategies once we have an incomplete set of points in various dimensions SVM inclines to be real because it has the capacity to discover the straight separation that should exist. SVM is great with outliers as it will just utilize the most related points to find a true separation.

VI. METHODOLOGY

Before starting, we added necessary toolboxes to the search path of MATLA. These were as follows:

- Utility Toolbox.

- Machine Learning Toolbox.
- SAP (Speech and Audio Processing) Toolbox.
- ASR (Automatic Speech Recognition) Toolbox.

We wrote a script to read in the audio files of the hundred tracks per category and extracted the MFCC features used for individual track. We additionally reduced the dimension of each track because extracted features are based on MFCC's statistics [8] comprising mean, std, min, and max along respectively dimension. Since MFCC has 39 dimensions, the extracted file-based features have $39 \times 4 = 156$ dimensions. To conclude, we used k-NN and SVM machine learning techniques via compact features set as well as with all features set of each track.

Below is the list of platform and MATLAB version that we utilized as a part of our investigation Platform: PCWIN64.

MATLAB version: 9.0.0.341360 (R2016a)

A. Data Collection

We gather all the sound files from the directory. The sound files have extensions of "au". These files have been sorted out for simple parsing, with a sub folder for each class.

Result.

Collecting 1000 files with extension "au" from "D:/szabist/matlab/GTZAN/genres"...

B. Feature Extraction

For every song, we separated the comparing feature vector for classification. We utilized the function `mgcFeaExtract.m` (which MFCC and its measurements) for feature extraction. We additionally put all the dataset into a single variable "dataset" which is less demanding for further handling which includes classifier development and assessment. Since feature extraction is extensive, we just loaded the dataset.mat. As discussed above the extracted features are based on MFCC's, so the extracted file-based features had $39 \times 4 = 156$ dimensions.

Result.

Extracting features from each multimedia object...

```
100/1000: file=D:/szabist/matlab/test1/GTZAN/blues/..  
200/1000: file=D:/szabist/matlab/test1/GTZAN/classical/..  
300/1000: file=D:/szabist/matlab/test1/GTZAN/country/..  
400/1000: file=D:/szabist/matlab/test1/GTZAN/disco/..  
500/1000: file=D:/szabist/matlab/test1/GTZAN/hiphop/..  
600/1000: file=D:/szabist/matlab/test1/GTZAN/jazz/..  
700/1000: file=D:/szabist/matlab/test1/GTZAN/metal/..  
800/1000: file=D:/szabist/matlab/test1/GTZAN/pop/..  
900/1000: file=D:/szabist/matlab/test1/GTZAN/reggae/..  
1000/1000: file=D:/szabist/matlab/test1/GTZAN/rock/..  
Saving dataset.mat...
```

C. Data Visualization

Since we had all the necessary information stored in "dataset", we applied different functions of machine learning toolbox for data visualization and classification. For example, we displayed the size of each class (see Fig. 2):

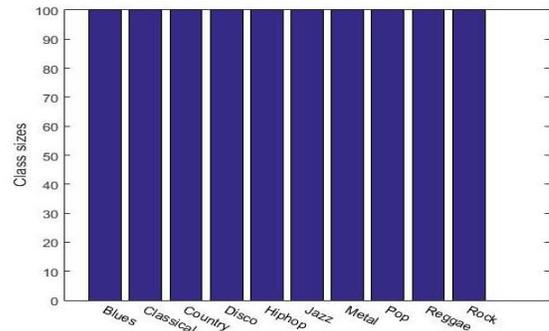


Fig. 2. Class sizes.

156 features
1000 instances
10 classes

We plotted the range of features of the dataset (see Fig. 3):

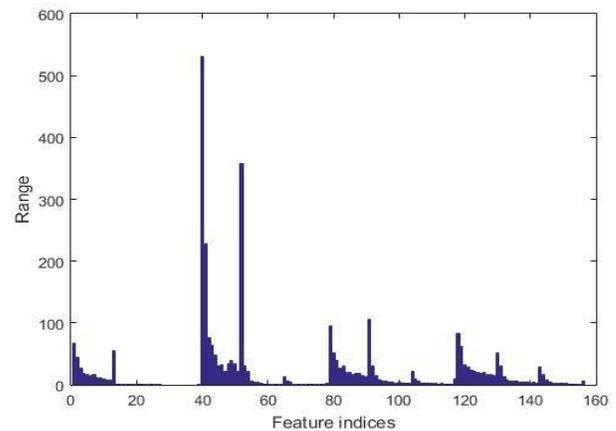


Fig. 3. Features range.

D. Dimensionality Reduction

The measurement of the feature vector is very large:

Feature measurement = 156.

We considered dimensionality reduction via PCA (principal component analysis) [7]. Initially the cumulative variance gave the descending eigenvalues of PCA (see Fig. 4):

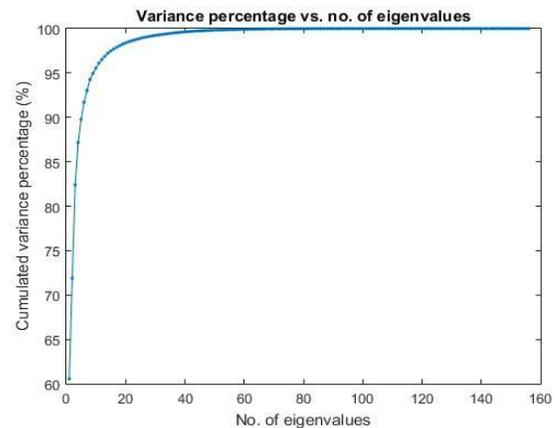


Fig. 4. Variance percentage vs. No. of eigen values.

A realistic choice is to maintain the dimensionality such that the cumulative variance percentage is greater than the threshold which is 95%.

We reduced the dimensionality to 10 to keep 95% cumulative variance via PCA (see Fig. 5).

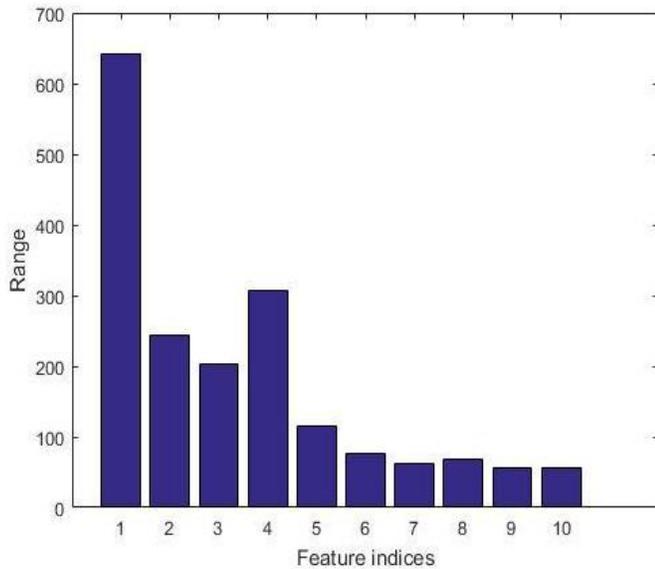


Fig. 5. Features range with reduced dimensions.

E. Classification and Results

At first we used the k-NN (k-nearest neighbor classifier) [5] for classification.

Result.

RR = 52.3 % for original dataset.

	blues	classical	country	disco	hiphop	jazz	metal	pop	reggae	rock
blues	52.00% (52)	0	11.00% (11)	6.00% (6)	1.00% (1)	1.00% (1)	8.00% (8)	0	8.00% (8)	13.00% (13)
classical	0	84.00% (84)	1.00% (1)	0	0	9.00% (9)	1.00% (1)	2.00% (2)	1.00% (1)	2.00% (2)
country	8.00% (8)	1.00% (1)	47.00% (47)	8.00% (8)	3.00% (3)	6.00% (6)	2.00% (2)	5.00% (5)	12.00% (12)	8.00% (8)
disco	5.00% (5)	0	16.00% (16)	37.00% (37)	8.00% (8)	1.00% (1)	5.00% (5)	4.00% (4)	6.00% (6)	18.00% (18)
hiphop	4.00% (4)	1.00% (1)	4.00% (4)	11.00% (11)	42.00% (42)	1.00% (1)	6.00% (6)	5.00% (5)	21.00% (21)	5.00% (5)
jazz	5.00% (5)	4.00% (4)	8.00% (8)	3.00% (3)	0	68.00% (68)	1.00% (1)	2.00% (2)	4.00% (4)	5.00% (5)
metal	3.00% (3)	0	3.00% (3)	4.00% (4)	6.00% (6)	2.00% (2)	71.00% (71)	0	2.00% (2)	9.00% (9)
pop	0	0	8.00% (8)	10.00% (10)	6.00% (6)	4.00% (4)	0	65.00% (65)	4.00% (4)	3.00% (3)
reggae	4.00% (4)	0	9.00% (9)	6.00% (6)	15.00% (15)	4.00% (4)	1.00% (1)	3.00% (3)	55.00% (55)	3.00% (3)
rock	8.00% (8)	1.00% (1)	12.00% (12)	23.00% (23)	2.00% (2)	6.00% (6)	11.00% (11)	1.00% (1)	6.00% (6)	30.00% (30)

Fig. 6. k-NN with reduced dimensions.

After k-NN we used other classifier in order to get a better result hence SVM [9] was used. Before using SVM for

classification we used a function `mgcOptSet.m` to put all the Music Genre Classification related options in a single file.

Using this classifier, we achieved following result.

Training Recognition Rate = 84.69%

Validating Recognition Rate = 64.20%

The recognition rate is 64%, indicating SVM is a much more effective classifier. We plotted the confusion matrix for better understanding of the results (see Fig. 7).

Our experiment showed that if PCA is used for dimensionality reduction, the accuracy will be lower. As a result, we kept all the features for further exploration.

Again the k-NN (k-nearest neighbor classifier) was used but with all features. The result was just over 55% which is less than what was achieved before.

Result.

R = 55.1 % for dataset after input normalization.

The confusion matrix was plotted for this in Fig. 6.

	blues	classical	country	disco	hiphop	jazz	metal	pop	reggae	rock
blues	61.00% (61)	2.00% (2)	12.00% (12)	0	2.00% (2)	3.00% (3)	8.00% (8)	0	2.00% (2)	10.00% (10)
classical	0	92.00% (92)	0	0	1.00% (1)	6.00% (6)	0	0	0	1.00% (1)
country	5.00% (5)	0	61.00% (61)	9.00% (9)	2.00% (2)	2.00% (2)	1.00% (1)	3.00% (3)	6.00% (6)	11.00% (11)
disco	1.00% (1)	0	5.00% (5)	56.00% (56)	3.00% (3)	2.00% (2)	4.00% (4)	8.00% (8)	8.00% (8)	13.00% (13)
hiphop	2.00% (2)	3.00% (3)	4.00% (4)	5.00% (5)	58.00% (58)	0	7.00% (7)	5.00% (5)	13.00% (13)	3.00% (3)
jazz	3.00% (3)	9.00% (9)	7.00% (7)	1.00% (1)	0	72.00% (72)	1.00% (1)	0	2.00% (2)	5.00% (5)
metal	3.00% (3)	0	0	2.00% (2)	4.00% (4)	2.00% (2)	80.00% (80)	0	1.00% (1)	8.00% (8)
pop	0	2.00% (2)	10.00% (10)	9.00% (9)	6.00% (6)	0	0	70.00% (70)	3.00% (3)	0
reggae	5.00% (5)	3.00% (3)	10.00% (10)	7.00% (7)	12.00% (12)	1.00% (1)	1.00% (1)	3.00% (3)	55.00% (55)	3.00% (3)
rock	9.00% (9)	1.00% (1)	10.00% (10)	19.00% (19)	1.00% (1)	6.00% (6)	11.00% (11)	1.00% (1)	5.00% (5)	37.00% (37)

Fig. 7. SVM with reduced dimensions.

RR = 57.6% for original dataset

RR = 64.9% for dataset after input normalization

Now the recognition rate is improved from 55% to 64% which is equivalent to the previous recognition rate of SVM, it shows that with all features k-NN is more effective classifier (see Fig. 8).

Not achieving the satisfied result, we again used SVM but with all features.

Result.

Training RR=99.01%

Validating RR=77.00%

So now the training rate is improved from 84.69% to 99.01% and recognition rate is improved from 64% to 77% (see Fig. 9 below).

	blues	classic	country	disco	hiphop	jazz	metal	pop	reggae	rock
blues	49.00% (49)	0	12.00% (12)	13.00% (13)	2.00% (2)	0	12.00% (12)	0	5.00% (5)	7.00% (7)
classic	0	90.00% (90)	3.00% (3)	0	0	6.00% (6)	0	0	0	1.00% (1)
country	0	0	62.00% (62)	5.00% (5)	2.00% (2)	6.00% (6)	1.00% (1)	4.00% (4)	1.00% (1)	19.00% (19)
disco	5.00% (5)	0	8.00% (8)	55.00% (55)	3.00% (3)	0	2.00% (2)	2.00% (2)	5.00% (5)	20.00% (20)
hiphop	1.00% (1)	0	3.00% (3)	9.00% (9)	57.00% (57)	0	2.00% (2)	7.00% (7)	14.00% (14)	7.00% (7)
jazz	0	7.00% (7)	7.00% (7)	1.00% (1)	0	73.00% (73)	2.00% (2)	1.00% (1)	0	9.00% (9)
metal	1.00% (1)	0	3.00% (3)	5.00% (5)	4.00% (4)	0	76.00% (76)	0	0	11.00% (11)
pop	0	1.00% (1)	8.00% (8)	7.00% (7)	3.00% (3)	0	0	71.00% (71)	5.00% (5)	5.00% (5)
reggae	0	0	4.00% (4)	17.00% (17)	11.00% (11)	0	1.00% (1)	9.00% (9)	54.00% (54)	4.00% (4)
rock	0	0	15.00% (15)	13.00% (13)	1.00% (1)	4.00% (4)	4.00% (4)	0	1.00% (1)	62.00% (62)

Fig. 8. k-NN with all dimensions.

	blues	classic	country	disco	hiphop	jazz	metal	pop	reggae	rock
blues	83.00% (83)	0	2.00% (2)	2.00% (2)	1.00% (1)	2.00% (2)	5.00% (5)	0	0	5.00% (5)
classic	0	94.00% (94)	0	0	2.00% (2)	2.00% (2)	0	0	0	2.00% (2)
country	4.00% (4)	0	70.00% (70)	5.00% (5)	0	1.00% (1)	0	6.00% (6)	2.00% (2)	12.00% (12)
disco	2.00% (2)	0	3.00% (3)	66.00% (66)	7.00% (7)	1.00% (1)	2.00% (2)	4.00% (4)	6.00% (6)	9.00% (9)
hiphop	3.00% (3)	0	0	4.00% (4)	74.00% (74)	0	2.00% (2)	1.00% (1)	14.00% (14)	2.00% (2)
jazz	0	5.00% (5)	1.00% (1)	1.00% (1)	0	90.00% (90)	0	0	0	3.00% (3)
metal	6.00% (6)	0	2.00% (2)	3.00% (3)	1.00% (1)	0	83.00% (83)	0	0	5.00% (5)
pop	0	0	9.00% (9)	4.00% (4)	2.00% (2)	0	0	80.00% (80)	2.00% (2)	3.00% (3)
reggae	5.00% (5)	0	4.00% (4)	6.00% (6)	8.00% (8)	0	0	4.00% (4)	71.00% (71)	2.00% (2)
rock	4.00% (4)	0	12.00% (12)	11.00% (11)	0	3.00% (3)	7.00% (7)	1.00% (1)	3.00% (3)	59.00% (59)

Fig. 9. SVM with all dimensions.

VII. CONCLUSION

Accuracy of classification by different genres and different machine learning algorithms is varied. The success rate of SVM was 83% but the blues genre was misjudged as rock or metal genre. The k-NN did badly while recognizing blues with a recognizing percentage of 49%. The SVM also misidentified classical genre as jazz or hip-hop, but the rock genre was accurately identified with success rate of 94%. The K-NN did also well when identifying classical with success rate of 90%. Similarly, the SVM did also well with recognizing entire categories but on the other hand it also inaccurately identified

disco with rock and reggae with hip-hop. The success rate of country was 70% but with rock genre it was just 12%. Hip hop genre had the success rate of 74% but had difficulty differentiating between reggae with highest inaccuracy of 14%. Jazz was identified with the accuracy rate of 90% but had difficulty in recognizing classical genre. Rock has the lowest success rate of 59% having difficulties with many other genres. K-NN had difficulty differentiating between other genres with blues with lowest success rate of 49%. The success rate of rock genre was 62% which was better than the SVM. Overall we found that SVM is more effective classifier which gave 77% accuracy.

VIII. FUTURE WORK

In the end our study creates a simple solution on the genre classification problem of music. However, it could be further extended out in a few ways. For instance, our research doesn't give an absolute reasonable correlation between learning strategies for classification of music genre. The exact similar methods utilized in this research could be effortlessly stretched to categorize songs created on some further category, like including extra metadata content elements for example music album, track name, or lyrics.

REFERENCES

- [1] Chaturanga, Y. M. ., & Jayaratne, K. L. (2013). Automatic Music Genre Classification of Audio Signals with Machine Learning Approaches. *GSTF International Journal of Computing*, 3(2).
- [2] Serwach, M., & Stasiak, B. (2016). GA-based parameterization and feature selection for automatic music genre recognition. In *Proceedings of 2016 17th International Conference Computational Problems of Electrical Engineering, CPEE 2016*.
- [3] Dijk, L. Van. (2014). Radboud Universiteit Nijmegen Bachelorthesis Science Finding musical genre similarity using machine learning techniques, 1–25.
- [4] Aucouturier, J., & Pachet, F. (2003). Representing Musical Genre: A State of the Art. *Journal of New Music Research*, 32(February 2015), 83–93.
- [5] Leif E. Peterson (2009) K-nearest neighbor. *Scholarpedia*, 4(2):1883.
- [6] Mandel, M. I., Poliner, G. E., & Ellis, D. P. W. (2006). Support vector machine active learning for music retrieval. *Multimedia Systems*, 12(1), 3–13.
- [7] Jolliffe, I. T. (2002). *Principal Component Analysis*, Second Edition. *Encyclopedia of Statistics in Behavioral Science*, 30(3), 487.
- [8] Logan, B. (2000). Mel Frequency Cepstral Coefficients for Music Modeling. *International Symposium on Music Information Retrieval*, 28, 11p.
- [9] Fu, Z., Lu, G., Ting, K. M., & Zhang, D. (2011). A survey of audio-based music classification and annotation. *IEEE Transactions on Multimedia*, 13(2), 303–319.
- [10] West, K., & Cox, S. (2004). Features and Classifiers for the Automatic Classification of Musical Audio Signals. *Proc. International Society for Music Information Retrieval Conference*, 1–6.
- [11] Aucouturier, J.-J., & Pachet, F. (2004). Improving timbre similarity: How high's the sky? *Journal of Negative Results in Speech and Audio Sciences*, 1(1), 1–13. Retrieved from
- [12] Bilmes, J. A. (1998). A gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *ReCALL*, 4(510), 126.
- [13] Moreno, Pedro, J., Ho, Purdy, P., & Vasconcelos, N. (2003). A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications. *Proceedings of Neural Information Processing Systems*, 16, 1385–1393.
- [14] Li, T. L. H., Chan, A. B., & Chun, A. H. W. (2010). Automatic Musical Pattern Feature Extraction Using Convolutional Neural Network.

- Proceedings of the International MultiConference of Engineers and Computer Scientists (IMECS 2010), I, 546–550.
- [15] Xu, C., Maddage, N. C., Shao, X., Cao, F., & Tian, Q. (2003). Classification using support machine, 429–432.
- [16] Meng, A., Ahrendt, P., & Larsen, J. (2005). Improving music genre classification by short-time feature integration. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, V, 497–500.
- [17] Li, T., Ogihara, M., & Li, Q. (2003). A comparative study on content-based music genre classification. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval SIGIR 03, 15(5), 282.
- [18] Berenzweig, A., Logan, B., Ellis, D. P. W., & Whitman, B. (2004). A Large-Scale Evaluation of Acoustic and Subjective Music-Similarity Measures. Computer Music Journal, 28(2), 63–76.
- [19] Liang, D., Gu, H., & Connor, B. O. (2011). Music Genre Classification with the Million Song Dataset 15-826 Final Report.
- [20] Rhodes, C. (2009). Music Information Retrieval, II(2008), 1–14. Retrieved from <http://www.doc.gold.ac.uk/~mas01cr/teaching/cc346/>
- [21] Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals: IEEE. IEEE Transactions on Speech and Audio Processing, 10(5), 292–302.
- [22] Logan, B., & Salomon, a. (2001). A Music Similarity Function based on Signal Analysis. IEEE International Conference on Multimedia and Expo 2001, 0(C), 952–955.
- [23] Gjerdingen, R. O., & Perrott, D. (2008). Scanning the Dial: The Rapid Recognition of Music Genres. Journal of New Music Research, 37(2), 93–100.
- [24] Lippens, S., Martens, J. P., & De Mulder, T. (2004). A comparison of human and automatic musical genre classification. 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, 4, iv-233-iv-236.
- [25] Liu, Z., & Huang, Q. (2002). Content-Based Indexing and Retrieval-by-Example in Audio. Proceedings of the 2000 IEEE International Conference on Multimedia and Expo (ICME '00), 2(c), 877–880.
- [26] Saunders, J. (1996). Real-time discrimination of broadcast speech/music. Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on, 2, 993–996 vol. 2.
- [27] E. Scheirer and M. Slaney, “Construction and evaluation of a robust multi feature speech/music discriminator,” in Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on, vol. 2. IEEE, 1997, pp. 1331–1334.
- [28] Kimber, D., & Wilcox, L. (n.d.). Acoustic Segmentation for Audio Browsers 1 Introduction 2 Acoustic Segmentation.
- [29] Boyce, R., Li, G., Nestler, H. P., Suenaga, T., & Still, W. C. (2002). Locating singing voice segments within music signals, (October), 7955–7956.
- [30] Foote, J., & Uchihashi, S. (2001). The beat spectrum: A new approach to rhythm analysis. Proceedings - IEEE International Conference on Multimedia and Expo, 881–884.
- [31] Srinivasan, H., & Kankanhalli, M. (2004). Harmonicity and dynamics-based features for audio. Proc. ICASSP, 4, 321–324.
- [32] Zhang, Y., & Zhou, J. (2004). Audio Segmentation Based on Multi-Scale Audio Classification. IEEE International Conference on Acoustics, Speech, and Signal Processing, 349–352.
- [33] George Tzanetakis. 2002. Manipulation, Analysis and Retrieval Systems for Audio Signals. Ph.D. Dissertation. Princeton University, Princeton, NJ, USA.
- [34] Mckinney, M. M. F., & Breebaart, J. (2003). Features for Audio and Music Classification. Proc ISMIR, 4(November 2003), 151–158.
- [35] Lu, L., Zhang, H. J., & Jiang, H. (2002). Content analysis for audio classification and segmentation. IEEE Transactions on Speech and Audio Processing, 10(7), 504–516.
- [36] P. Ahrendt, A. Meng and J. Larsen. (2004). Decision time horizon for music genre classification using short time features, 12th European Signal Processing Conference, Vienna, pp. 1293-1296.
- [37] J. T. Foote, (1997). Content-based retrieval of music and audio, Int. Soc. Opt. Photon. Voice Video Data Commun., pp. 138-147.
- [38] Li, G., & Khokhar, A. A. (2000). Content-based indexing and retrieval of audio data using wavelets. In IEEE International Conference on Multi-Media and Expo (II/TUESDAY ed., pp. 885-888)
- [39] Scheirer, E. D. (1998). Tempo and beat analysis of acoustic musical signals. The Journal of the Acoustical Society of America, 103(1), 588–601.
- [40] “Marsyas data sets.” [Online]. <http://marsyas.info/downloads/datasets.html> Available: Accessed on 20th July 2017
- [41] Cristianini, N., & Schölkopf, B. (2002). Support Vector Machines and Kernel Methods: The New Generation of Learning Machines. Artificial Intelligence Magazine, 23(3), 31–42.

The Identification of Randles Impedance Model Parameters of a PEM Fuel Cell by the Least Square Method

M^{ed} Selméne Ben Yahia

Laboratory of Energy Efficiency
Application and Renewable Energy
LAPER, Tunis El Manar University,
Faculty of Sciences, Tunisia

Hatem Allagui

Laboratory of Energy Efficiency
Application and Renewable Energy
LAPER, Tunis El Manar University,
Faculty of Sciences, Tunisia

Abdelkader Mami

Laboratory of Energy Efficiency
Application and Renewable Energy
LAPER, Tunis El Manar University,
Faculty of Sciences, Tunisia

Abstract—One of the problems of industrial development of fuel cells is the reliability of their performances with time. The solution of this problem is through by the development of improved diagnostic methods such as the identification of parameters. This work focuses on the modeling and the identification of the impedance model parameters of a Proton Exchange Membrane (PEM) fuel cell. It is based on the Randles model represented by specific complex impedance at each cell. We have used the “Least square” method to determine the parameters model using measured reference values. The proposed authentication method is valid for Randles model, but it can be generalized to be applied to others.

Keywords—Randles model; impedance; Proton Exchange Membrane (PEM) fuel cell; modeling; parameters identification; least square

I. INTRODUCTION

The fuel cell is not a source of energy, but a converter that directly transforms the chemical energy of a fuel into electrochemically powered and its working principle was discovered in 1839 by W. Grove [1], [2]. It is an efficient means of electrical production in terms of efficiency. However, behind the displayed simplicity of its operating principle, it may indeed appear as a relatively complex to be technically implemented. The modeling of a fuel cell is an important factor in describing the operation of the fuel cell system through a well-defined model, and more precisely the fuel cell impedance model that describes the frequency behavior. The objective is not to describe in detail and exhaustively all the models present in the fuel, but rather to highlight the dynamic model in which one can easily identify its parameters.

Our choice was the impedance model. There were two reasons for this choice; the first is the validity of the impedance model of the fuel cell at the electrical level and the multiphysics phenomena. This model permits to generate the polarization curve and the Nyquist diagram as a function of the frequency from a well-defined scan. The second reason is the importance of the impedance model to facilitate the diagnosis of the fuel cell and to prolong the life of the fuel cell. The identification of impedance model parameters is necessary to model diffusion phenomena, hence the need to use optimization and identification methods. The objective of this

article is to judge the methodology for identifying the impedance model parameters of the fuel cell used and to validate it by experimental results. The use of parametric identification requires to be coupled to another approach, most often using decision to accomplish the task of diagnosis [3]. Among the methods using an equivalent electrical circuit, we can cite the author’s work [4] which has established a model based on a so-called current interruption measure for the estimation of parameters of the PEM fuel cell in order to diagnose its state. Other authors [5] have developed another equivalent electrical circuit consisting of three RC cells placed in series to model the non-linear relationship between voltage and current. This method was used to detect the phenomenon of flooding. In order to diagnose the cell, a method has been found to determine the correlation between the resistance value of the diffusion layer and the water content in the cell [6]. The parameters were studied using an extended Kalman filter, in the case of dynamic stresses in current [7]. Another method of identification based on the generalized moments of input and output from current-jump measurements allowed knowing the evolution of the electrochemical phenomena during an aging process.

The identification methods have the advantage of being simple to implement; however, the diagnostic method may present operational difficulties due to the use of test signals. This is why its implementation must be accompanied by a complementary strategy, especially during real-time applications. In [8], the authors propose the extrapolation of the laws applied to electrical circuits (law of nodes and meshes) on other hydraulic or pneumatic systems. The proposed model for the fuel cell is capable of monitoring the evolution of anode and cathode gas dynamics (pressure and flow) and predicting the fuel voltage. The model allows the detection of the deterioration of the membrane, the drying and the flooding of the cells. Diagnosis is based on the identification of resistance to flow. For example, a flooding defect is located at the cathode if the drop in the cell voltage is gradual, the resistances equivalent to the cathode increase and those of the anode remain invariant.

We will start with the presentation of the PEM fuel cell. Then, we will explain the impedance circuit and the analysis of the effect of the frequency behavior on the defined scan. In the

third part, we will detail the procedure for identifying the parameters of the proposed model to be developed by a Matlab environment program.

II. DESCRIPTION OF THE PEM FUEL CELL

The principal role of the full cell is to convert the chemical energy of a fuel into electrochemically electricity. There are several cell technologies to fuel characterized by the nature of the electrolyte, the operating temperature, the gas consumed.

The five categories of fuel cell are:

- AFC (Alkaline Fuel Cell)
- PEMFC (Proton Exchange Membrane Fuel Cell)
- PAFC (Phosphoric Acid Fuel Cell)
- SOFC (Solid Oxide Fuel Cell)
- MCFC (Molten Carbonate Fuel Cell).

Each type contains a fuel cell, an oxidizer, an electrolyte, electrodes, an important and different temperatures for each type, the chemical reactions at the anode and cathode also are important for the functioning of the pile.

A. Fuel Cell Exchange Membrane PEM Proton

The PEMFC (Proton Exchange Membrane Fuel Cell) is a fuel cell operating at low temperature (between 30° C and 90° C). Its basic principle is the electrochemical combustion of hydrogen and oxygen [9]. The simplest system allows from hydrogen and oxygen, supply water, electricity and heat.

Fig. 1 expresses the principle of operation of a PEM cell.

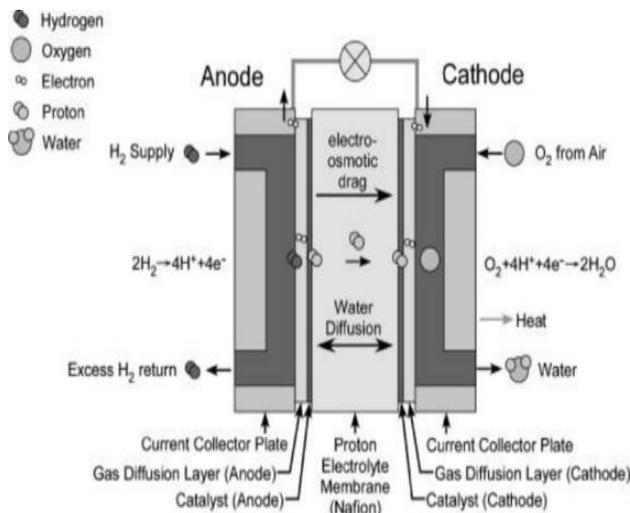


Fig. 1. The Operating principle of a PEM fuel cell.

A fuel cell consists of three main elements [10]:

- The anode which is fed by a fuel (hydrogen, methanol, etc.).
- The cathode which is fed with an oxidizer (oxygen).
- The electrolyte, solid or liquid, which separates the two electrodes and to provide the distribution of the

intermediate ion of the oxidation reaction of the fuel cell.

The electrolyte must prevent the passage of electrons through the electrical circuit. The reaction produced is the reverse reaction of the electrolysis of water. The overall reaction is:



B. Electrical Characteristics of PEMFC

We recall a few laws of thermodynamics necessary to understanding the external characteristics of fuel cells. In general, the balance of a reaction is written [11]:



The gas activity is defined by: $a = P/P_0$

Where, P and P^0 , respectively represent the partial pressure of gas and the standard pressure. The variation of the Nernst equation for the reaction is expressed by [12]:

$$E = E_0 + \frac{RT}{nF} \ln \left(\frac{a_{\text{A}}^{\alpha} a_{\text{B}}^{\beta}}{a_{\text{C}}^{\delta}} \right) \quad (3)$$

Where, R is the universal constant of gas equal to 8,314 J.mole⁻¹.k⁻¹.

This last equation is the general form of the Nernst equation. It shows the dependency of the load voltage with pressure at constant temperature. Thus, the theoretical voltage of a cell increases when the activity of the reactants increases and the activity of the products decreases.

Any electrode bringing together the oxidized and reduced forms of a redox couple has what is called an electrode potential. This is achieved through the Nernst law which connects to the activity of the reactants and the products of electrochemical reaction occurring at the electrode. The ideal performance of a fuel cell is determined through the evaluation of the potential on each electrode in each fuel cell technology. In the case of the PEM cell, the previous equation becomes [9]:

$$E = E_0 + \frac{RT}{2F} \ln \left(\frac{P_{\text{H}_2}}{P_{\text{H}_2\text{O}}} \right) + \frac{RT}{4F} \ln (P_{\text{O}_2}) \quad (4)$$

III. MODELING OF CIRCUIT IMPEDANCE OF THE PEM FUEL CELL

For a fuel cell, several models can be developed depending on the objective to be reached [4]-[7]. The integration of a fuel cell in an electrical environment requires knowledge of its electrical model.

The characterization tools were presented in the literature. We propose here to detail those which we used in our experimental tests:

- The automatic drawing of voltage-current curves.
- The impedance spectroscopy.
- The study parameters of this method depend on the richness of the information desired.
- The scanning frequency.

- The current sweep pattern (sine, linear).

The model must be simple, precise and must predict the electrical behavior in both static and dynamic conditions.

The simplest model can be an input-output model (equivalent circuit, for example) that would allow description of fuel cell behavior in its environment.

A. Characterisation and Simple Modeling of Fuel Cells

The polarization curve of the PEM fuel cell presented in Fig. 2 is generally described as the sum of four terms: the theoretical open circuit voltage E , V_{act} activation overvoltage, resistance overvoltage V_{ohm} and the concentration overvoltage V_{conc} [12]:

$$V_{fuel} = E - V_{act} - V_{ohm} - V_{conc} \quad (5)$$

Fig. 2 gives the polarization curve of the PEM fuel cell.

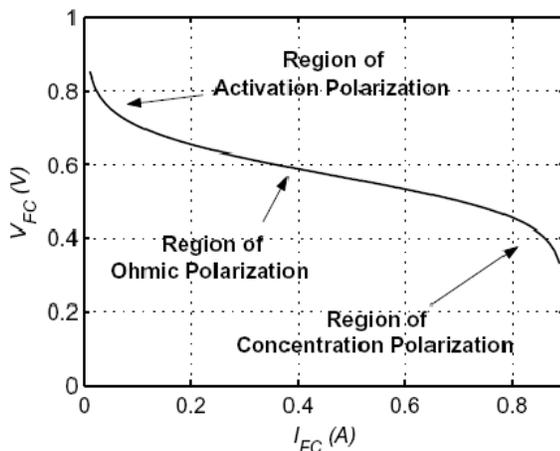


Fig. 2. Voltage-current characteristic of the fuel cells.

1) Activation polarization

The activation losses are due to starting of the chemical reactions at the anode and cathode. A portion of the available energy is used to re-form chemical bonds to the electrodes. If these losses occur at the two electrodes, the oxidation reaction of hydrogen at the anode is much faster than the reduction of oxygen at the cathode. It follows that activation losses are mainly due to cathodic reactions. The relationship between activation losses and the current density is given by equation Tafel [12]:

$$V_{act} = A \cdot \ln\left(\frac{I_{FC} + i_n}{i_0}\right) \quad (6)$$

Where I_{FC} is the current delivered by the fuel cell, the exchange current i_0 characterizing empty the electrode-electrolyte exchanges, in the internal power to take account of a possible passage of gas and / or electrons through the electrolyte and to the slope of the Tafel.

2) Ohmic polarization

The ohmic losses are due to the resistance being the electrodes and the bipolar plates to the flow of electrons and the electrolyte to the passage of protons. The corresponding voltage drop is written [12]:

$$V_{ohm} = R_m(I_{FC} + i_n) \quad (7)$$

Where R_m is the total resistance of the fuel cell.

3) Concentration polarization

The gas consumption depletes the gas mixtures and reduces its partial pressure. This pressure reduction depends on the current issued and characteristics of gas circuits. This voltage drop is expressed in terms of a current limit i_L , for which all the fuel being used, its pressure drops to zero, and constant B called constant transport transportation or mass transfer [12]:

$$V_{conc} = -B \cdot \ln\left(1 - \frac{I_{FC} + i_n}{i_L}\right) \quad (8)$$

4) Relationship Between the Polarization and the Impedance Model

In a fuel cell, several models can be developed according to the objective. The integration of a fuel cell in an electric environment requires knowledge of its electric model. The model must be simple, accurate and must allow predicting the electrical behavior both in static conditions and in dynamic regime. The simplest model can be an input-output model type (equivalent circuit, for example) which allow the description of the fuel cell behavior in its environment. The simplest representation of the fuel cell as an electric model is to put a DC voltage source in series with electrical impedance in Fig. 3.

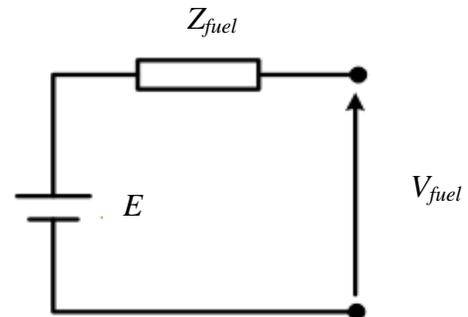


Fig. 3. Representation of a fuel cell using a voltage source associated with its electrical impedance [12].

B. Impedance Model of PEM Fuel Cell

1) Fuel cell impedance spectroscopy model

The Impedance spectroscopy is a measure conventionally used in electrochemistry. It is concerned with determining the impedance of a material or a system in response to electrical excitation.

The determination of the electrical impedance of the system is carried out in three steps [13]:

- Superposition of a sinusoidal component of small amplitude (small perturbation Signal δI) to the direct current (or voltage) imposed on the battery for A whole range of frequencies ($\omega_n = 2 \cdot \pi \cdot f_n$ the pulsation of the excitation associated with the second frequency). The DC component of the current corresponds to a point of operation on the polarization curve $V=f(I)$ of the stack.
- Measurement of the amplitude and phase shift of the sinusoidal component of the response In voltage (δV) of the fuel.

- Calculation of the complex impedance $\bar{Z}(\omega)$ (respectively the admittance) of the fuel cell over the range of frequencies studied. This impedance is defined Such as the ratio of the voltage to the current in the frequency domain; the VDC and IDC are eliminated [13].

In this section we propose an equivalent electrical circuit of the fuel cell. This circuit brings together electrical elements representing the electrical and electrochemical phenomena (loss of activation, ohmic loss, loss of concentration, double layer phenomenon) [14], [15]. These phenomena are coupled with those studied in the diffusion approach to establish the final model.

2) Impedance model of randles

In a fuel cell, several models can be developed according to the objective. The integration of a fuel cell in an electric environment requires knowledge of its electrical model. The model must be simple and should be used to predict the electrical behavior both in static mode only loaded dynamically.

The simplest model can be an input-output model type (equivalent circuit for example) that allows the description of the fuel cell behavior in its environment. The simplest representation of the fuel cell in the form of an electrical model is to a DC voltage source in series with electrical impedance (Fig. 3). The electrical impedance includes a capacitance of the double layer C_{DC} and a resistance R_T characterizing charges transport phenomena to the electrodes. The resistance R_M represents the membrane resistance and the different others resistances in contact with the membrane in this case, diffusion phenomena are neglected.

The simplest representation of a fuel cell stack in the form of an electrical schematic diagram is given in Fig. 4. In this case, diffusion phenomena are neglected [12], [16]-[18].

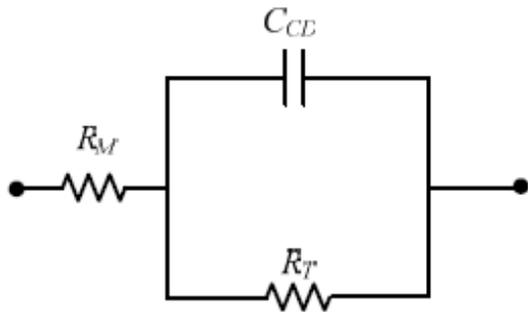


Fig. 4. Single impedance of an electrochemical cell.

So the cell impedance becomes:

$$Z(\omega) = R_M + \frac{R_T}{(1+j\omega R_T C_{DC})} \quad (9)$$

And the simple model impedance scheme has the following form in the Nyquist plane as illustrated in Fig. 5.

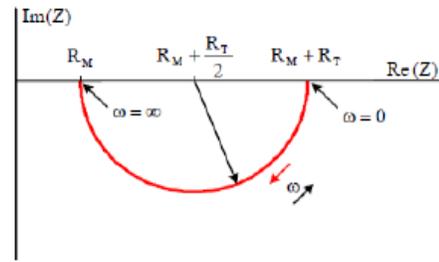


Fig. 5. Simple model impedance in the Nyquist plane.

The plot of this impedance in the Nyquist plane corresponds to a semicircle which the center is $(R_M+R_T/2, 0)$ and a radius equal to $R_T/2$ (Fig. 3). At low frequencies, the impedance tends to R_M+R_T . At high frequencies; it tends to R_M .

The maximum of the capacitive effect corresponds to point $(R_M+R_T/2, -R_T/2)$ and is obtained for:

$$\omega R_T C_{DC} = 1 \quad (10)$$

The given experimental points represent the real and the imaginary part of the measured impedance presented in a table.

The reference impedance $Z(\omega)_{ref}$ is:

$$Z(\omega)_{ref} = \text{Real}_{ref}Z(\omega) + j \cdot \text{Im}_{gref}Z(\omega) \quad (11)$$

The impedance of the selected model is a resistor in parallel with a capacitance both in series with a further resistor.

$$Z(\omega) = R_M + \frac{1}{\frac{1}{R_T} + j\omega \cdot C_{DC}} \quad (12)$$

After a mathematical calculation, the expression is divided into a real and an imaginary part

$$Z(\omega) = R_M + \frac{R_T}{1+\omega^2 R_T^2 C_{DC}^2} - j \frac{\omega R_T^2 C_{DC}}{1+\omega^2 R_T^2 C_{DC}^2} \quad (13)$$

We set $1 + \omega^2 R_T^2 C_{DC}^2 = a$

Expression of the impedance of the model becomes:

$$Z(\omega)_{mod} = R_M + \frac{R_T}{a} - \frac{j(\omega R_T^2 C_{DC})}{a} \quad (14)$$

The real impedance equal model:

$$\text{Re}_{el_{mod}}(Z(\omega)) = R_M + \frac{R_T}{a} \quad (15)$$

The imaginary impedance of the model:

$$\text{Im}_{g_{mod}}(Z(\omega)) = -\frac{R_T^2 \omega C_{DC}}{a} \quad (16)$$

IV. PARAMETERS IDENTIFICATION OF THE IMPEDANCE MODEL OF PEM FUEL CELL

A. The principle of identification method

The principle of a method of identification is the iterative search of a set of parameters θ allowing minimizing a criterion of deviation between a model and experimental measurements,

as illustrated in Fig. 6. The principle is to apply the same set of excitations u_k to experimental measurements and to a model. The outputs \hat{y}_k of the model and the outputs y_k of the measurement are then compared and the set of parameters θ is adjusted to reduce the measurement-model deviation on the set of N iterations resulting from the applied excitation [13].

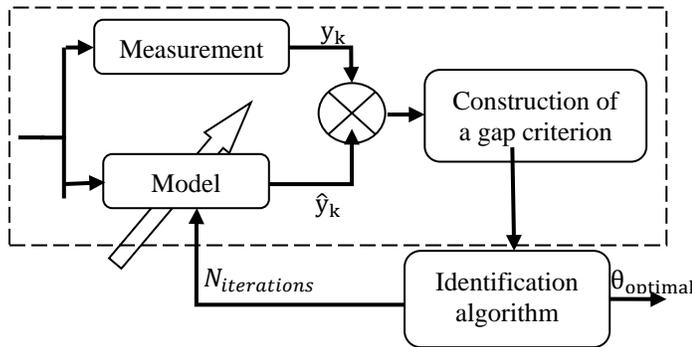


Fig. 6. Identification principle [13].

The formulation of an identification problem is reduced to the formulation of a monocriter optimization problem whose objective function is constituted by a quadratic error criterion F_θ defined by the sum of the model-measurement errors squared. The set θ of M parameters can be subject to domain constraints. In the end, the problem to solve is expressed by (17) [13].

$$\min[F_\theta] = \min[\sum_{k=1}^N (\hat{y}_k(\theta) - y_k)^2] \quad (17)$$

Or $\theta = [\theta_1 \dots \theta_M]$ With M the number of parameters of the model and $\theta_{i_{\min}} \leq \theta_i \leq \theta_{i_{\max}}$ with $i \in [1 \dots M]$

B. Representation of Impedance Measurements by Nyquist Diagram

As a result of the measurements, the information can be plotted in the complex plane, the abscissa axis giving the real part of the impedance and the ordinate axis the imaginary part. This mode of representation is called a Nyquist diagram. Each point of the Nyquist diagram corresponds to the total impedance measured for a given excitation frequency during impedance spectroscopy. The Nyquist diagram is rather used by electrochemists because it facilitates the reading of phenomena with different dynamics [13]. Fig. 7 expresses the Randles model representation of the Nyquist plane.

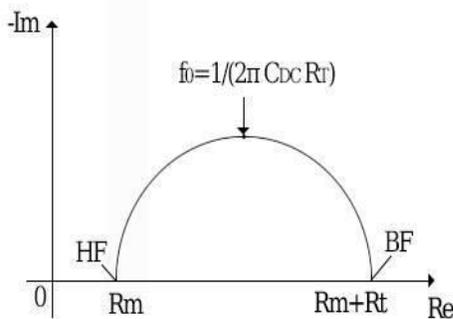


Fig. 7. Randles model representation of the Nyquist plane.

The Nyquist diagrams can provide useful information for fuel cell analysis. Thus, we usually extract the value of the real part of the intersection of the curve with the abscissa axis at the low frequencies (which corresponds to the resistance attributed to the membrane of the models presented later), the frequency at the top of the curve, and the width of the circle between the two intersections of the curve with the abscissa axis (which corresponds to the sum of the electrical resistances of the models).

C. Methodology of Identification: Parameters Extraction

The parametric identification of the electrochemical field differs from one researcher to another, after having obtained experimental data, they must be analyzed. For this, he chose to parameterize the models described by the researcher, so that their answers correspond to the experimental test. The set of parameters obtained will make it possible to characterize, at least in part, the state of the fuel cell at the time of the measurement. To analyze degradation, I want to compare the parameters extracted from the experimental tests carried out under degrading conditions. For impedance spectroscopy and current scans, the principle of parameter extraction is to compare the experimental reading with the result given by the model and to adjust the different parameters of the model in order to make “Paste” the Response of the latter with the Experimental. Procedure for identification by the least squares method. The objective of this section is to present the method of identifying the parameters of the impedance model using the measured values. The organizational chart of the working methodology used for the identification is given in the following Fig. 8.

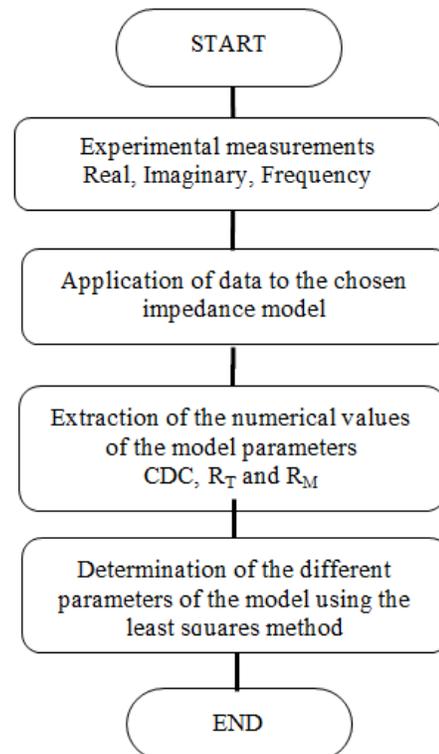


Fig. 8. Organizational chart of the methodology used for identification.

After obtaining the reference values used in the works of Arafet and Rouane [19] and choosing the model described in Fig. 4. We can calculate and estimate the parameters of the model. The experimental points given represent the real part and the imaginary part of the measured impedance presented in a table. The reference impedance $Z(\omega)_{ref}$ is:

$$Z(\omega)_{ref} = Real_{ref}Z(\omega) + j.Img_{ref}Z(\omega) \quad (24)$$

The fuel cell impedance model is obtained by the frequency behavior of the complex impedance of the fuel cell as follows :

$$Z(\omega)_{ref} = Real_{model}Z(\omega) + j.Img_{mod}Z(\omega) \quad (25)$$

Where $Real_{model}Z(\omega)$ is the real of the impedance model of the fuel cell and $Img_{mod}Z(\omega)$ is the imaginary part of the fuel cell impedance model.

The experimental data generated by the electrochemical impedance spectroscopy method are mainly analyzed using an electric circuit model. The most of the circuit elements used in the model are the electrical elements such as resistance, capacitor, inductance and the electrochemical elements such as the Warburg impedances.

The minimization of the error between the impedance of the calculated model and the measured impedance makes it possible to find the optimal parameters minimizing the difference between each reference point and corresponding point in the model. At the start of the calculation, the initial values are assigned to these parameters. The comparison between the reference impedance and the impedance of the model is done by the error criterion or the minimization criterion of the least squares method (J).

$$J = \sum_{i=1}^n \{ (Reel_{ref}Z_i - Reel_{mod}Z_i)^2 + (Img_{ref}Z_i - Img_{mod}Z_i)^2 \} \quad (26)$$

n represents the number of experimental points excited by the frequency.

When the error identification criterion is validated, the resulting data set is the optimal set. The calculation stops when a stopping condition is reached, namely:

- The maximum number of iterations reached.
- The variation of tolerance of each parameter reached.
- Tolerance on the implementation error.

Before passing to the development of the proposed method, I want to remind some important points for the determination of impedance electrochemical model parameters by impedance spectroscopy. The Nyquist diagram is a set of points identified on the real part and the imaginary part measured for a frequency ranging from 0.1 Hz to 12 kHz.

These points from the spectrum.

So we go to the next step which is determining R_M . Indeed, the determination of R_M , R_T and C_{DC} depends on minimizing the error criterion J.

1) The determination of R_M

After reading the measurements table, we obtained three vectors which the coordinates are the frequency, the real part and the imaginary part corresponding.

The high frequency HF = [FHF ReHF ImHF]

The low frequency BF = [FBF ReBF ImBF]

The average frequency MF = [FMF ReMF ImMF].

At the beginning of the program, R_M , R_T and C_{DC} parameters are initialized to zero. A high frequencies approximation is assumed that the imagination of the reference value is zero, we obtain the following equation:

$$J_{RM} = (Reel(HF) - R_M)^2 \quad (27)$$

We start the iterations and the stop condition is as follows:

$J_{RM} \leq \epsilon$ Or the maximum number of iterations is attained. So, is assigned ϵ a value equal to 10^{-8} and begins to increment R_M by a space no value h_1 , whose value depends on the estimated value of the parameter to calculate. Therefore varies R_M to have J_{RM} exceeding ϵ or we reached the maximum number of iterations. In this case one can say that R_{Mf} is found.

Determining the flowchart of the R_M is given in Fig. 9.

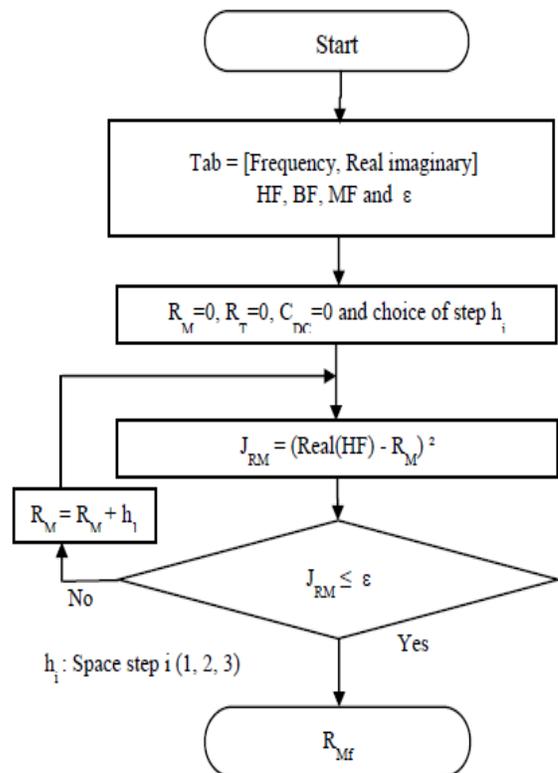


Fig. 9. Determining the flowchart of the R_M .

2) The determination of R_T

In this stage R_M is known (computed in the previous step). At low frequencies, it is assumed that the reference imaginary is equal to zero. Therefore we obtain the following equation:

$$J_{RT} = (\text{Re}(BF) - (R_{Mf} + R_T))^2 \quad (28)$$

R_T is incremented by a space not h_2 value is incremented and R_T to be J_{RT} exceeding ϵ it reached the maximum number of iterations.

Determining the flowchart of the R_T is given in Fig. 10.

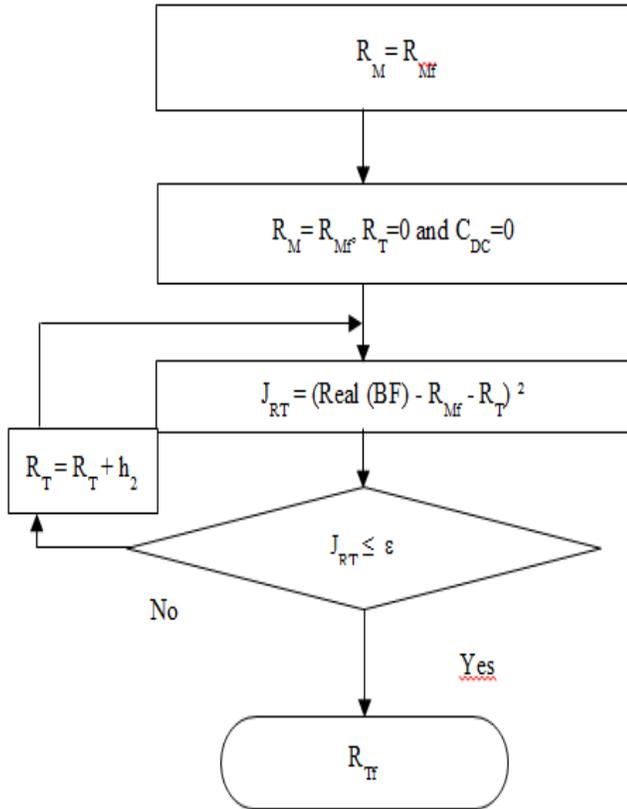


Fig. 10. Determining the flowchart of the R_T .

3) The determination of C_{DC}

The determining C_{DC} is performed by choosing any point in the measurement chart provided that its imaginary part is not zero. An average frequency and seeks to minimize the criterion J_{CDC} example we take while incrementing the C_{DC} value one step h_3 . The equation of the criterion to be minimized is:

$$J_{CDC} = \left(\text{Re}(MF) - \left(R_{Mf} + \frac{R_{Tf}}{1 + \omega^2 R_{Tf}^2 C_{DC}^2} \right) \right)^2 + \left(\text{Im}(MF) - \frac{R_{Tf}^2 \omega C_{DC}}{1 + \omega^2 R_{Tf}^2 C_{DC}^2} \right)^2 \quad (29)$$

This step continues until the number of iterations is reached or $J_{CDC} \leq \epsilon$.

Determining the flowchart of the C_{DC} is given in Fig. 11.

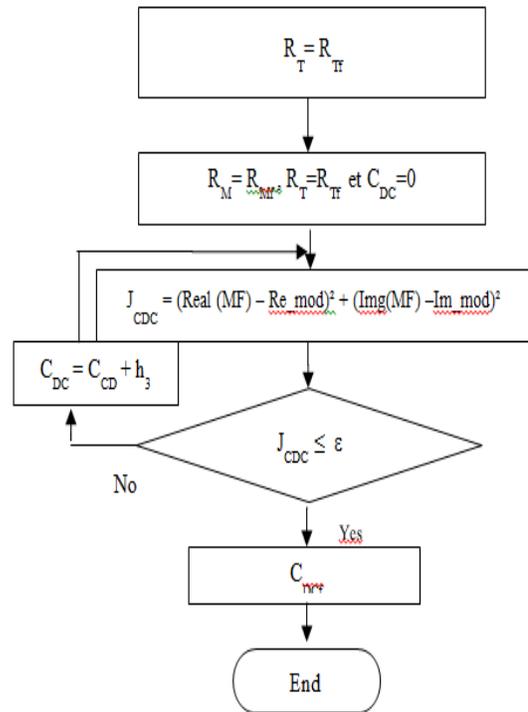


Fig. 11. Determining the flowchart of the C_{DC} .

The principle of extracting the parameters using this program is summarized by the steps indicated in Fig. 12.

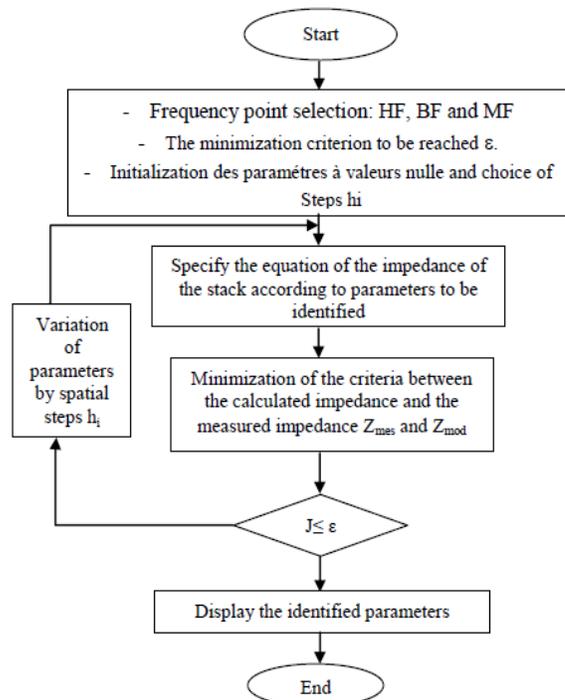


Fig. 12. Main chart calculation.

D. Description of the Algorithm in the Frequency Selection Program

It is assumed that the first value in the table is the maximum frequency (high frequency).

We run the table while comparing each value to the first. It is the same routine which is repeated for the lower frequency.

For the average frequency dividing the number of rows of the table by two, the resulting value provides an indication of the value of the average frequency which will be used. Fig. 13 presents gives details about selection of frequency.

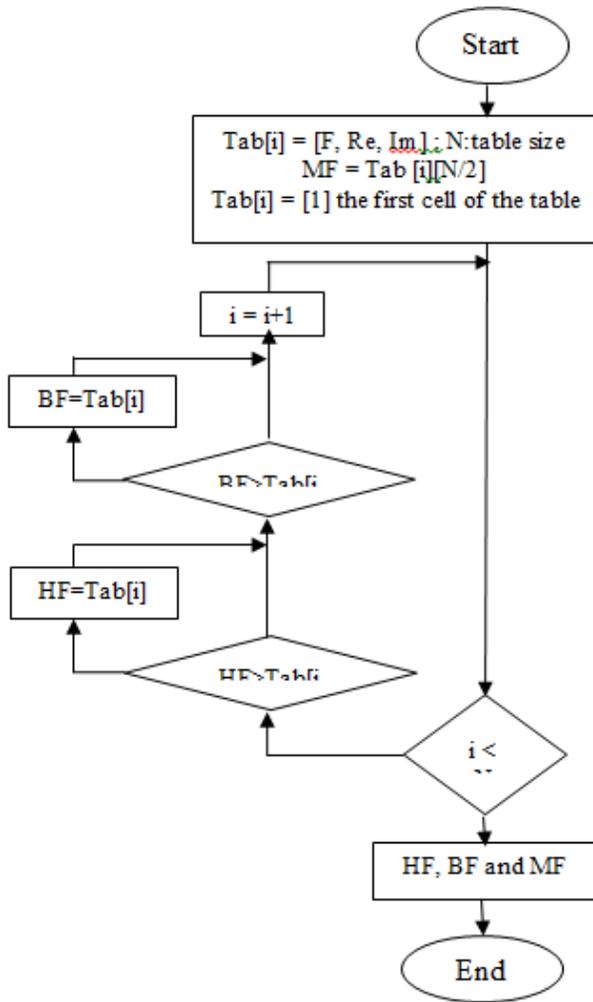


Fig. 13. Frequency selection sub-program flowchart.

V. THE EXPERIMENTAL WORK

The identification algorithm is developed in Matlab, software. It contains the main program, the sub-programs of the frequency selection and the parameters display. This program after simulated gives waited results as those found by Arafet and Rouane [19] for a PEMFC Nexa Fuel Cell type Ballard whose power is 1200W.

This validates the results found by our algorithm. Moreover, these values are almost identical to those obtained by M. Selmene et al. using the genetic algorithm [20]. These results are also close to the measured values found by Reddad [21] et al. and Rouane [22] who worked on a fuel cell having the same characteristics as the Nexa fuel cell.

The following tables show the calculating step and extract the results found during the minimization of the criterion J for each parameter (Tables 1, 2, and 3).

TABLE I. SOME POINT R_M

SOME POINT CHANGES R_M	
R_M	J_{RM}
0	1E-4
0.001	8.11E-5
0.003	4.9E-5
0.006	1.6E-5
0.00973	7.2E-8
0.01001	1.2E-10
0.0201	1.02E-4
0.031125	4.46E-4
0.041905	1.01E-3

TABLE II. SOME POINT R_T

SOME POINT CHANGES R_T		
R_M	R_T	J_{RT}
0.01001	0	0.0081
0.01001	0.001	0.0079
0.01001	0.003	0.0075
0.01001	0.006	0.007
0.01001	0.021	0.0047
0.01001	0.045	0.002
0.01001	0.078	1.44E-4
0.01001	0.091	1.0E-6
0.01001	0.12	9E-4

TABLE III. SOME POINT C_{DC}

SOME POINT CHANGES C _{DC}			
R _M	R _T	C _{DC}	J _{CDC}
0.01001	0.091	0	0.00602
0.01001	0.091	0.0001	0.00561
0.01001	0.091	0.001	0.00205
0.01001	0.091	0.0022	1.67E-4
0.01001	0.091	0.0028	7.66E-6
0.01001	0.091	0.003001	6.54E-8
0.01001	0.091	0.006	4.8E-4
0.01001	0.091	0.01	0.001
0.01001	0.091	0.015	0.0013

The convergence curves of R_M, R_T and C_{DC} are given in Fig. 14 to 16.

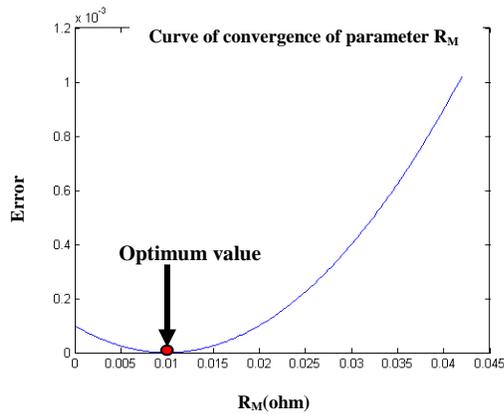


Fig. 14. Curve of convergence of parameter R_M.

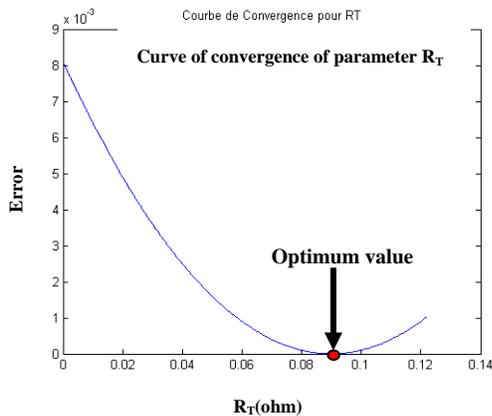


Fig. 15. Curve of convergence of parameter R_T.

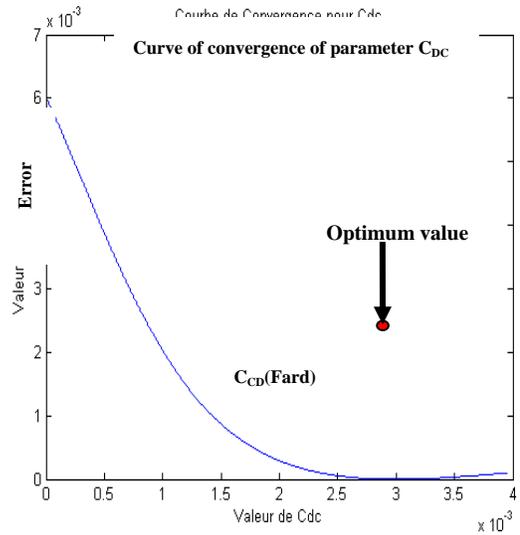


Fig. 16. Curve of convergence of parameter C_{DC}.

The convergence curves of different parameters admit an absolute minimum error which describes the value of each parameter so min J_{xx} gives the exact values.

The following tables show the calculating step and extract the results found during the minimization of the criterion J for each parameter (Table 4).

TABLE IV. VALUES OF THE PARAMETERS IDENTIFIED

Figure	Fig. 10	Fig. 11	Fig. 12
Error rate J _{RM} (%)	1.2E-10	-	-
R _M (Ω)	0.01001	-	-
Error rate J _{RT} (%)	-	1.0E-6	-
R _T (Ω)	-	0.091	-
Error rate J _{CDC} (%)	-	-	6.54E-8
C _{DC} (F)	-	-	0.003001

The identification algorithm used is the least-squares method which makes it possible to identify the resistive and capacitive elements such as the values R_M, R_T and C_{DC}. Model given by the table presented in Table 5.

TABLE V. THE OPTIMAL PARAMETERS OF THE COMPLEX IMPEDANCE OF THE NEXA PEM FUEL CELL.

Parameters	Numerical Value
C _{DCA} (F)	0.003
R _M (Ohm)	0.01
R _{Tc} (ohm)	0.091

Table 6 shows the comparison of two different method of identification and the parameters of the impedance model of a fuel cell identified in each methodology.

TABLE VI. COMPARISON OF TWO DIFFERENT METHOD OF THE IMPEDANCE MODEL PARAMETERS IDENTIFICATION

Identification methodology	Parameters to be identified
Classical method "least square"	Electrical parameters " $R_M R_T C_{DC}$ " (our case)
Advanced method "genetic algorithm"	Electrochemical parameters " $R_M R_T C_{DC} Z_w L \dots$ "

The identification of the electrochemical parameters differs from one researcher to another. Philippoteau [23] choose to parameterize the models described so that the parameters obtained correspond to the experimental test. On the other hand Morin [24] was identified static and dynamic parameters to be able to simulate the dynamic model.

The objective of Morin is to determine the parameters (if included Lelec) represented on the model associated with the activation phenomena, species diffusion and charge conduction losses. Among these parameters, four (the C_{diff} XXX and Cd) concern the dynamic aspects and all the others concern the static aspect. My identification work is based on the least squared method which is closer to Philippoteau's methodology [23] at the level of parameter set obtained "criterion to minimize" will permit to characterize, at least in part, the state of the stack at the time of measurement.

But this methodology does not allow to identifying the electrochemical parameters such as the Warburg impedance and the CPE phase constant element.

In this case, we must use advanced methods like the Methaeristic technique or the network of neuron which permit to evaluate an identification work based on one of the most intelligent techniques and compares their results by what find in this paper.

VI. CONCLUSION

In this paper, we have described a methodology for identifying the parameters of the complex impedance model of the fuel cell. This impedance model is based on electrical elements such as membrane resistance, load transfer resistance and double layer capacitor. It hangs on the mathematical equations of the elements from the experimental results by an algorithm of identification based on the method of least squares. The simulation results are presented by Nyquist diagrams to describe the different phenomena that occur inside the fuel cell.

REFERENCES

[1] H. Oman. Fuel Cells Personal Electricity. IEEE AES Systems Magazine, September 2000.

[2] 2013 Fuel Cell Technologies Market Report: Fuel Cell Technologies Office. U.S. DEPARTEMENT OF ENERGY. November 2014.

[3] Taher Hamaz, "Tools for characterization and diagnosis of a PEM fuel cell by external electromagnetic field measurement", Thesis, University Grenoble, 2015

[4] A. Forrai, H. Funato, Y. Yanagita, and Y. Kato. Fuel-Cell Parameter Estimation and Diagnostics, IEEE Transactions On Energy Conversion. Vol. 20 (3), pp. 668 - 675 , 2005.

[5] K. Sugiura, M. Yamamoto, Y. Yoshitani, K. Tanimoto, A. Daigo, T. Murakami, Performance diagnostics of PEFC by current-pulse method, J Power Sources, (2006), Vol.157, pp. 695- 702, 2006

[6] M.A Rubio, A. Urquia, S. Dormido. Diagnosis of PEM fuel cells through current interruption, J Power Sources, 171, 670-677, 2007

[7] T. Hamaz, C. Cadet, F. Druart. Détection de dysfonctionnements d'une pile à combustible PEMFC à partir de sauts de courant. CIFA 2012, Grenoble, France, 2012

[8] A. Hernandez, D. Hissel, R. Outbib. Modeling and Fault Diagnosis of a Polymer Electrolyte Fuel Cell Using Electrical Equivalent Analysis. IEEE TRANSACTIONS ON ENERGY CONVERSION, VOL. 25, NO. 1, pp. 148- 160, 2010

[9] Marielle MARCHAND. « Water Management in Fuel Cells ». Thesis of the National Polytechnic Institute of Grenoble. Novembre 1998

[10] Benjamin BLUNIER. " Modeling motor-compressors for the air management in the fuel cell systems - simulation and experimental validation". Thesis, University of Technology Belfort-Montbéliard. December 2007

[11] Amel LACHICHI. "Modeling and stability of a hybrid current regulatorApplication to converters for fuel cells". Thesis of the University of FrancheComté. November 2005.

[12] MajidZandi. "Contribution to the management of hybrid electric power sources." Thesis of the University of the National Polytechnic Institute Lorraine. University, Nancy Novevember 2010.

[13] Thomas GENEVE , "Methods of diagnosis of fuel cells", thesis of Institut National Polytechnique of Toulouse (INP Toulouse) February 2016

[14] Thomas Mennola, Mikko Mikkola, Matti NOPONEN, Tero Hottinen, Peter LUND, «Measurement of ohmic voltage Losses in individual cells of a PEMFC stack", Journal of Power Sources, 112: 261-272, 2002.

[15] K.R. COOPER et M. SMITH. «Electrical test methods for on-line fuel cell ohmic resistance measurement». Journal of Power Sources 160, pp. 1088-1095, 2006

[16] IdrisSadli. "Modeling impedance of a PEM fuel cell for power electronics in use".Thesis of the National Polytechnic Institute of Lorraine. December 2006.

[17] El-Hassan AGLZIM. "Characterization spectroscopy the complex impedance of impedance of a fuel cell in load Evaluation of the influence of moisture." Thesis, University Henri Poincaré-Nancy1. November 2009.

[18] Xiao-Yuan Zi · · Chaojie SONG Haijiang Wang.JiuJun ZHANG. "Electrochemical Impedance Spectroscopy in PEM Fuel Cells Fundamentals and Applications".

[19] Amar Rouane, El-Hassan AGLZIM, Bernhard KRAEMER, Reddad EL-MOZNINE, "Impedance Measurement of Fuel Cell is a Load". 9th International Conference electrical power quality and use, Barcelona october 2007

[20] Mohamed Selméne Ben Yahia, Hatem Allagui, Arafet Bouaicha, Abdelkader Mami "Fuel Cell Impedance Model Parameters Optimization using a Genetic Algorithm" The International Journal of Electrical and Computer Engineering (IJECE) Vol. 7, No. 1 , February 2017, pp. 196~205

[21] El-Hassan AGAIZIM Amar rouane, Reddad EL-MOZNINE "An Electronic Measurement Instrumentation of the impedance of a Loaded Fuel Cell Battery gold.Sensors", 2007 (ISSN 1424-8220 © 2007 by MDPI www.mdpi.org/sensors).

[22] El-Hassane AGLZIM, Amar ROUANE, Mustapha NADI and Djilali KOURTICHE, Signal Processing for the Impedance Measurement on an Electrochemical Generator, Sensors & Transducers Journal, Vol. 90, Special Issue, pp. 150-159, April 2008.

[23] Vincent PHILIPPOTEAU. "Tools and Methods for the Diagnosis of a Health State of a Fuel Cell". Thesis of the National Polytechnic Institute of Toulouse. July 2009

[24] Benoît Morin. "Hybridisation of a fuel cell by supercapacitors to a passive and direct solution". Thesis of the National Polytechnic Institute of Toulouse. February 2013.

Using the Facebook Iframe as an Effective Tool for Collaborative Learning in Higher Education

Mohamed A. Amasha

Department of Computer Teacher Preparation,
Damietta University.
Damietta, Egypt

Salem Alkhalaf

College of Science and Arts, Computer Science Department,
Qassim University
Alrass City, Saudi Arabia

Abstract—Facebook is increasingly becoming a popular environment for online learning. Despite the popularity of using Facebook as an e-learning tool, there is a limitation when it comes to presenting content: another platform is required to run the files. Presented in this paper is a case study of how the Facebook iframe code can be used as a hosting environment tool to support collaborative activities in higher education at Qassim University. The study was conducted on a sample of (N=45) university students who were enrolled in Selected Topics in Information Systems (INFO491) at the Faculty of Art & Science at Qassim University. We used Facebook markup language (FBML) to design and implement the course. An online questionnaire was used to investigate the students' perceptions about using Facebook iframe for the course. *Descriptive statistical analysis* and *chi-square test* were used to analyze the data. According to our results, the participants reported that using the Facebook iframe page increased their understanding and improved their learning performance. In addition, for the majority of students, it enabled them to learn more quickly. Our findings also revealed that a Facebook iframe page is a distinctive hosting environment for presenting content.

Keywords—Facebook iframe; collaborative learning; Facebook markup language (FBML); hosting environment

I. INTRODUCTION

Facebook has become an effective tool for e-learning and the most popular social networking site for college students [2], [18]. It is used as an e-learning tool in enhancing learner course outcomes. In addition, it offers great potential for teaching and learning [21]. It enables teachers to engage in three major types of interaction: learner-to-instructor interaction, learner-to-learner interaction, and learner-to-subject interaction [10], [31]. With Facebook, a structural equation model is constructed which examines the relationships between factors affecting this adoption process in relation to the user's existing purposes [25]. A previous study entailed the use of Facebook for three-way communication and as an interaction tool [16]. As a useful and meaningful learning environment, it can support, enhance, and/or strengthen student learning [14], [23]. As previous studies have shown, there are some constraints of using Facebook as an e-learning tool [1]: 1) Facebook does not support many file formats that need to be uploaded directly; 2) file size is limited for Facebook uploads; and 3) students cannot control the content via navigation buttons. One possible way of using Facebook is to use its iframe to host the course content and share resources, upload files, make

announcements, and conduct online discussions between teacher and students. The main purpose of this study was to describe how the Facebook iframe was used as a hosting environment of content and to report students' perceptions about it. We expected that the use of Facebook iframe would be a better way to host the content as well as deliver it to and share it with students. Hosting content in the Facebook canvas enables students to control the content and navigate through it. We hoped to improve each student's ability to use Facebook as a learning tool. The results were expected to provide insights into promoting the use of Facebook iframe in a collaborative learning environment.

II. LITERATURE REVIEW

A. Social Network

The 21st century has witnessed a great revolution, especially in the field of information and communications technology (ICT). This was accompanied by the appearance of some new techniques and applications, which make the user a positive participant in knowledge creation rather than being just a passive recipient. Social media is considered among these innovations of modern technology [17]. Social networks are no longer used only in communication and chatting with friends; they have become transparent and interactive learning atmospheres in which the learner is an active participant in the educational process. Using social media facilitates communication and creates an effective network worldwide [14]. Many teachers use online websites on a daily basis personally without being aware of how to activate those websites in education [13]. Facebook, for example, is one of the most popular social networks. It is ranked as the ninth best learning tool worldwide [2]. Nowadays, social media has become a part of our daily lives and changed the way people communicate. It is now used for communication, collaboration, and learning. It is one of the daily routines for learners, where professors can follow students' learning process as well. Facebook has become a tool for learning used by teachers in schools, and it is also used in universities as a learning management system (LMS) [5]. Facebook is considered the most widely used social network in the educational field; there are various schools and universities and educational institutions using it in the learning process. There are many ideas for using Facebook inside the classroom that both teachers and professors can apply easily. In addition, it can be used as a revision tool on which the teacher can post notes or a summary after each lesson [19]. Moreover, it can be

used as an advertisement board and as a way to share sites, files, and multimedia which support student learning and broaden their understanding [32]. Facebook can also be used as a study group. The teacher can divide the students into groups to discuss the lessons or the projects with each other, and the teacher's role is to follow the discussions and provide encouragement [24]. It is also a tool for communicating with parents to inform them about the level of their children or summer activities through special pages used for this purpose [27], [3].

B. Facebook as Learning Tool

Facebook is considered a learning platform, which hosts many new strategies for teaching as learning based on projects, solving problems, brainstorming, and teaching strategies. Facebook, with tools and potential, can contribute in raising the level of motivation of learners and improve the environment of the classroom, in addition to improving the relation between the students and teachers. The Facebook environment provides two types for interaction and feedback: syndication through the chat tool on Facebook and no syndication through Facebook pages or groups. Based on this research, Facebook was chosen to be the host site because it has distinguished features that are not available on other social networking sites. Presented in this study is a model of teaching and learning for using the Facebook iframe that is, it helps to host iframe code in a Facebook canvas [16]. Facebook is considered one of the most efficient sources of information for learners. It provides them with productive skills according to their needs, interests, and aspirations. The aim of this research is to make use of the available computer technology in the field of education. Facebook is effectively used as an LMS for e-learning and as a device to improve e-learning courses [24]. The most recent adaptation of e-Front is made with an arrangement of social apparatuses that encourage the utilization of Facebook as a LMS for e-learning. It comes with a simple Facebook integration plug-in for easy use. E-Front is an open source of learning management system with an attractive appearance and is SCORM certified. E-Front encourages group learning and maintains standards of aggregate information [8]. Utilizing familiar technologies that students are comfortable with, such as Facebook, helps in developing a successful learning environment. It makes good use of the creative, interactive, and collaborative nature of Facebook [4]. The goal of these technologies is to meet learners' needs, and they are useful in helping them accomplish their educational purposes [12].

On Facebook, an instructor can begin a closed group that is inaccessible in Facebook. This group will be effective for publishing information and getting feedback from the students [24]. Additionally, these groups will be useful for posting homework, links for further study, and declarations; sharing ideas; and socializing after school hours [30]. The reasons using Facebook in education was chosen for this study are as follows:

- The large number of network users.
- The availability of e-mail, forums, and chat, which work as a learning environment.

Therefore, the subject of Facebook was chosen to uncover the explanations behind its educational value [14]. The network enables us to socialize, chat, and communicate easily with our relatives and friends at no cost. We can share our opinions and thoughts about what is going on around the world. In addition, at the same time, we get immediate feedback on what we write or share. Thus, Facebook is considered the best means of interaction and communication. It also has strong privacy settings that maintain confidentiality according to the user's preferences [26].

C. Aim

The aim of the current study was to examine students' perceptions of using Facebook iframe as a hosting environment of the content of to support learning.

III. METHODOLOGY

A. Course Design and Implementation

1) Context:

In this study, Facebook markup language was used to design the elective course for undergraduate students at the Faculty of Science & Art at Qassim University. In this course 31 students were enrolled. Sessions were presented twice a week, and each session lasted for two hours. One session of the course per week was presented face-to-face, and the other session was presented online. The course addressed some topics in relation to the subject of e-learning.

2) Setting up Facebook iframe page and course materials:

A Facebook page was created, and then an iframe static tab was added to the page (<http://statictab.com/v4pa9jw>). The content of the course was created using PowerPoint and was converted into a video file using the SnagIt application. A Google drive (<https://drive.google.com/drive>) was used to upload the content and to obtain iframe tags and HTML (hypertext markup language) code. We developed HTML (hypertext markup language) code to add lectures to the Facebook iframe page. The following tags are a sample of HTML code inside a Facebook iframe page:

```
<p></p>
<div align="CENTER" style color = "ffff00" >
<h1>E-learning lecture</h1>
</div>
<iframe src="https://archive.org/embed/rrrrrrrrrr_453"
width="640" height="480" frameborder="0"
webkitallowfullscreen="true" mozallowfullscreen="true"
allowfullscreen></iframe><br><br>
<center><a href="http://statictab.com/zgo2sxm"
target="_blank"><input type="submit" style="width: 200px;"
value="main page " name="B1" /></a></center><br />
<center><a href="http://statictab.com/sfgmg47"
target="_blank"><input type="submit" style="width: 200px;"
value="lecture1" name="B1" /></a></center><br /><br />
```

3) Setting up a Facebook group

A Facebook closed group (INFO490) was created before starting the course. We asked students to set up their own profiles, and then they were invited to join the group. There were 45 students in this group. We provided the students with instructions and urged them to abide by the ethics of the group. After that, the researchers instructed the students to access the Facebook page and to participate in the activities of the page and view lectures on a regular basis.

4) Organizing lectures

Lectures were organized in the form of educational programs, allowing students to control and navigate to content. Weekly materials were added to the iframe page. Once the lecture material was created and uploaded on Google drive, the iframe code was generated and posted on Facebook iframe. The Facebook iframe page subscribed to the materials on the Facebook page, so Facebook became a hosting environment for the content and students had control of it. The researchers used the discussion board to ask questions, and this board was, for the most part, used on a weekly basis. Fig. 1 shows the educational program icon on the Facebook iframe page. Fig. 2 shows the content of the first lecture.



Fig. 1. Educational program icons on the Facebook iframe page.



Fig. 2. Shows the content of the first lecture.

B. Participants

The study was conducted at Qassim University during the first semester of the 2016 academic year. The participants were 45 graduate students who were enrolled in a course entitled Selected Topics in Information Systems (INFO491). All participants were invited to create a Facebook page if they had not already and invited to participate in the Facebook group (INFO 491). As was previously mentioned, the course sessions were held twice a week; one of these was face-to-face, and the other was online through the Facebook iframe page. After completing the learning of the course, all participants were asked to respond to an online survey at (<https://goo.gl/forms/EEJL2Xv7VZmZUX6I3>). In Table 1, the demographic profile is summarized. It includes participants' genders, ages, and opinions about using Facebook for learning. In total, 45 students participated in the study. Twenty-one of them were female (46.67%), and twenty-four were male (53.33%). The results in Table 1 show that 84.44% of Facebook users were 20–30 years old. Thirty-three participants (73.33%) use Facebook 5–10 times daily. About 42 participants (93.33%) do not mind using Facebook in e-learning courses, and a majority of them would not mind using Facebook iframe as a hosting environment tool to present content (93.33%). They think it would be useful for them, and it would give them the opportunity to exchange experiences and connect daily with their classmates.

C. Instrument

In this study, an online questionnaire was developed to investigate students' attitude toward the use of the Facebook iframe page as a learning tool. The questionnaire was designed with a 5-point Likert scale: (1= strongly disagree, 2= disagree, 3= slightly agree, 4= agree, 5= strongly agree). It consisted of 23 closed-ended questions and had five sections: *Facebook Self-efficacy (SE)*; *Attitude toward Facebook Iframe (AT)*; *Behavioral Intention (BI)*; *Outcomes and Usefulness (OU)*; and *Ease of Use (EU)*. Students were invited to follow the Google drive link to respond to the questionnaire, and we mailed the questionnaire to students that enrolled in the course (N=45). A Google spreadsheet was used to collect data. Ten experts validated the questionnaire, and its internal reliability was found to be good. The Cronbach's α coefficient was 0.745.

D. Data analysis

All analyses were conducted using the SPSS statistical software package version 20.0. Frequency distributions, mean score, and standard deviation were calculated for each item from the online questionnaires in the spreadsheet. Chi-square test was used to compare between actual and prospective students' responses regarding the use of the Facebook iframe page as an effective learning tool for the Selected Topics in Information System (INFO491) course. Statistical significance level was set at $p < 0.05$.

TABLE I. SAMPLE DEMOGRAPHIC INFORMATION

Items	Frequency (n=45)	(%)
Gender		
Male	24	53.33
Female	21	46.67
Age		
20-25	3	6.67
26-30	38	84.44
30+	4	8.89
Facebook usage		
One a day	6	13.33
5-10 times a day	33	73.33
10-15 times a day	5	11.12
More than 15 times a day	1	2.22
Opinion about Facebook in learning courses		
It isn't suitable as an educational tool.	1	2.22
Communicate to my friends.	2	4.44
I would not mind.	42	93.33

IV. RESULTS AND DISCUSSION

The aim of the current study was to examine students' perceptions of using Facebook iframe as a hosting environment of the content of to support learning. The results suggest that Facebook iframe may be useful in supporting students' learning outcomes. Also, students find the presentation of course content through Facebook iframe page is effective, efficient, and easy to use. Prior studies conducted revealed that, despite their agreement with the importance of Facebook as a learning environment, students had difficulty viewing the course content [33], [29], [14]. In our study, the use of the Facebook iframe page as a hosting learning tool changed students' attitudes toward using Facebook as a learning tool. This change was evident in our results, which show that students are influenced to adopt Facebook because they want display all the content on the Facebook page in the form of an educational program and not have to move to other web pages or obtain additional software. In addition, students want to establish or maintain contact with other people with whom they share interests and values. The results indicate that Facebook Self-efficacy (SE), Attitude toward Facebook Iframe (AT), Behavioral Intention (BI), Outcomes and Usefulness (OU), and Ease of Use (EU) have a significant positive influence on Adoption of Facebook [7]. The responses to the current study's survey items reveal that students prefer the use of the Facebook iframe rather than the Facebook wall as a learning tool. Unlike the findings of previous research, which indicated that "participants felt that a Facebook group's wall cluttered with posts would prevent users from noticing critical information and locating important specific content" (Lin, 2016), the results of this research show that using the Facebook iframe page as a hosting environment is more attractive and effective than the use of Facebook as a

content organizer. These results agree with the findings of previous studies that involved using Facebook as learning tool [22], [15], [20] and our study of using Facebook as a hosting environment. In this study, a correlational analysis between the constructs was conducted to explore associations among them. The results in Table 2 indicate that Outcomes and Usefulness is positively and significantly correlated with Attitude toward Facebook Iframe (0.707; $p < .01$); Behavioral Intention (0.105; $p < .01$); and Attitude toward Facebook Iframe (0.700; $p < .01$). Facebook Ease of Use is also found to be positively correlated with Facebook Self-efficacy (0.679; $p < .01$).

TABLE II. A CORRELATION MATRIX BETWEEN CONSTRUCTS

Constructs	SE	AT	BI	OU	EU
SE	-				
AT	0.598	-			
BI	0.130	0.039	-		
OU	0.700	0.707	0.105	-	
EU	0.685	0.662	0.130	0.679	-

2-tailed p values; * $p < 0.05$, ** $p < 0.01$.
Facebook Self-efficacy (SE), Attitude toward Facebook Iframe (AT), Behavioral Intention (BI), Outcomes and Usefulness (OU), Ease of Use (EU)

TABLE III. PERCEPTIVE AND CHI-SQUARE TESTING

	Mean(SD)	Cronbach's Alpha	Chi-square χ^2
Facebook Self-efficacy:			
I have the basic skills for using Facebook iframe page.	3.51(1.12)	0.644	14.88 ($p = 0.005$)
I have the skill to deal with Facebook group and interact with them.	4.21(.884)		16.76 ($p = 0.001$)
Facebook as learning tool is attractive, motivating.	3.69(1.08)		16.00 ($p = 0.003$)
Average	3.80(1.02)		
Attitude Toward Facebook iframe:			
Using Facebook iframe page makes the learning stay in long-term memory.	3.76 (1.04)	0.645	17.11 ($p = 0.002$)
Facebook has to include all subjects, not just the current course.	3.20(1.19)		15.33 ($p = 0.004$)
Using Facebook iframe page as a learning tool is attractive and amusing.	3.81(1.19)		20.42 ($p = 0.000$)
Using Facebook iframe page is helpful and more attractive than traditional learning.	2.82(1.31)		3.55 ($p = 0.469$)
Average	3.39(1.18)		
Behavioral Intention:			
I intend to use Facebook iframe to learn other courses.	3.80(1.14)	0.840	17.77 ($p = .001$)
I feel positive toward using Facebook iframe as a learning tool.	3.80(1.19)		20.44 ($p = .000$)

	Mean(SD)	Cronbach's Alpha	Chi-square χ^2
The Facebook iframe page was a good learning environment.	4.04(.92)		28.66 (p=.000)
Facebook iframe page was a good environment to exchange ideas and share resources.	3.87(1.08)		18.11 (p=.001)
Average	3.87(1.08)		
Outcomes and Usefulness:			
Using Facebook iframe page in learning course makes it easier to learn.	3.89(1.32)	0.598	10.55 (p=.029)
Using Facebook iframe page in learning course increased my understanding.	3.51(1.30)		10.88 (p=.028)
Using Facebook iframe page in learning course enabled me to accomplish learning more quickly.	4.11(.885)		16.77 (p=.001)
Using Facebook iframe page in learning helps focus on understanding the content.	3.22(1.16)		8.44 (p=.077)
Using Facebook iframe page improved my learning performance.	3.76(.857)		15.88 (p=.001)
I would find using Facebook iframe useful for learning the course and presenting the content.	2.89(1.22)		9.55 (p=.049)
Average	3.56(1.12)		
Ease of Use:			
I find learning the course through Facebook iframe easy to use.	3.78(1.085)	0.691	18.22 (p=.001)
I find it is easy to navigate between the parts of the application and understand its content.	3.71(1.014)		20.22 (p=.000)
Average	3.745(1.04)		
Overall average	3.673(1.08)		

A. Facebook self-efficacy

Most of the students have the ability to deal with Facebook; the results ensure that they have the basic skills for using Facebook iframe page (mean score = 3.51, $\chi^2=14.88$, $p = 0.005$). In addition, they reported that Facebook as a learning tool is attractive and motivating, and they have a desire for further learning through Facebook and learn by different ways (mean score = 3.69, $\chi^2=16.76$, $p = 0.003$).

B. Attitude Toward Facebook Iframe

The majority of students' have positive opinions that Facebook can be a hosting learning environment to improve their learning performance, and this result agrees with the findings of various other studies [14], [4], [3]. The results in (Table 3) show the mean score of all 19 items ranged from 3.39 to 3.74. The students reported that using the Facebook iframe page makes the learning stay in long-term memory (mean score = 3.76, $\chi^2=17.11$, $p = 0.002$), is attractive and amusing (mean score=3.81, $\chi^2= 20.42$, $p=0.000$), and is

helpful and more attractive than traditional learning (mean score=2.82, $\chi^2=3.55$, $p = 0.469$). The students also stated that Facebook has to include all subjects and not just the current course (mean score=3.20, $\chi^2= 15.33$, $p = 0.004$).

C. Behavioral Intention

The students reported that they intend to use the Facebook iframe to study other courses (mean score=3.80, $\chi^2=17.77$, $p=.001$), and they have positive views toward using the Facebook iframe as a learning tool (mean score=3.80, $\chi^2=20.44$, $p=.000$). Additionally, the students expressed that the Facebook iframe page is a good environment in which to exchange ideas and share resources (mean score=3.87, $\chi^2=18.11$, $p=0.001$). As for using Facebook as a hosting environment [28], [9], [6] they said that Facebook may be effective for this purpose because it helps them to study and view content without the need for other sites or additional software. During this study, they found they were able to learn the material easily.

D. Outcomes and Usefulness

In terms of enhancing outcomes, the students reported that Facebook enhanced and increased their understanding of course content (mean score=3.51, $\chi^2=10.88$, $p =.028$). This result is similar to other studies conducted elsewhere [11]. The respondents believe that using Facebook iframe page in learning course enabled them to accomplish learning more quickly (mean score=4.11, $\chi^2=16.77$, $p =.001$), and their confidence in using Facebook as a hosting environment improved. As for using the Facebook iframe page in learning to help focus on understanding the content, the students surveyed admitted their learning rate level increased (mean score=3.22, $\chi^2=8.44$, $p =.077$). The students explained that this increase occurred because they were able to create a podcast episode, online presentation, animation, and interact with the Google education app. Almost all of the students agreed that the Facebook iframe page could provide an environment for enhancing and improving their learning performance (mean score=3.76, $\chi^2=15.88$, $p =.001$). Indeed, students' responses indicate that they found the Facebook iframe useful for learning course material and presenting the content (mean score=2.89, $\chi^2=9.55$, $p =.049$).

E. Ease of Use

Students concurred that they found learning a course through the Facebook iframe easy in terms of using the application (mean score=3.78, $\chi^2=18.22$, $p =.001$). They reported that it was easy to navigate between the parts of the application and understand its content (mean score=3.71, $\chi^2=20.22$, $p =.000$).

V. CONCLUSION

In this paper is presented the small body of knowledge about using the Facebook iframe for learning purposes and the results of a study that suggest Facebook may be used as a hosting environment and for supporting students' learning. The majority of students surveyed reported that using the Facebook iframe page for studying a course was a positive experience. They perceived that the use of the Facebook iframe with preset content was an innovative method to support their learning outcomes and high stakes examination.

Using Facebook iframe allows students to explore different opinions, control their own learning, and view content presented in a more effective way. Additionally, presenting content as an educational program can help students understand that content and avoid indoctrination in learning. This may support students' learning by enhancing their self-efficacy and outcomes. Importantly, students reported that working with the Facebook iframe and Facebook group was useful in helping them to learn by engaging in peer learning and exchange of experiences. A further result was that students decided that the use of the Facebook iframe as a hosting environment made them feel better prepared and that they had a deeper understanding of the content of the course. Furthermore, in our study, the Facebook iframe page provided students with a safe environment in which to focus on understanding the content of the course and engage in a deeper level of learning prior to assessment. As such, this enhanced a feeling of self-efficacy around their ability to be successful in the examination. Based on the findings of our research, we recommend integration of the Facebook iframe page as a learning tool into the resources of more university courses.

REFERENCES

- [1] Baran, "Facebook as a formal instructional environment," *British Journal of Educational Technology*, vol.41, no. 6 (2010),pp. 146–149.
- [2] B.W.O'Bannon, J. L.Beard and V. G. Britt, "Using a Facebook group as an educational tool: Effects on student achievement," *Computers in the Schools*, vol.30, no. 3, (2013), pp. 229-247.
- [3] C.Pimmer, J.Chipps, P.Bryiewicz, F.Walters, S.Linxen and U.Gröbhiel, "Supervision on social media: Use and perception of facebook as a research education tool in disadvantaged areas," *The International Review of Research in Open and Distributed Learning*, vol.17, no. 5, (2016), pp. 201-214.
- [4] C.Pimmer, S.Linxen, and U.Gröbhiel, "Facebook as a learning tool? A case study on the appropriation of social network sites from mobile phones in developing countries," *British Journal of Educational Technology*, vol.43, no. 5 ,(2012), pp. 726-738.
- [5] E.Miron and G.Ravid, "Facebook Groups as an Academic Teaching Aid: Case Study and Recommendations for Educators," *Journal of Educational Technology & Society*, vol.18, no. 4, (2015), pp. 371-384.
- [6] G. Y. Lin, Effects that Facebook-based Online Peer Assessment with Micro-teaching Videos Can Have on Attitudes toward Peer Assessment and Perceived Learning from Peer Assessment," *Eurasia Journal of Mathematics, Science & Technology Education*, vol.12, no.9, (2016).
- [7] G.Grossecka, R.Branb, and L.Tiruc, "Dear teacher, what should I write on my wall? A case study on academic uses of Facebook," *Procedia-Social and Behavioral Sciences*, vol.15, (2011), pp. 1425-1430.
- [8] H. M.Tall, G.Kurtz and E.Pieterse , Facebook Groups as LMS: A Case Study. *International Review of Research in Open & Distance Learning*, 13(4), (2012), pp. 33-48.
- [9] J. C. Clements1, "Using Facebook to enhance independent student engagement: a case study of first-year undergraduates," *Higher Education Studies*, vol.5, no. 4 (2015), pp.131-146.
- [10] J. P.Mazer, R. E.Murphy and C. J. Simonds, "I'll see you on "Facebook": The effects of computer-mediated teacher self-disclosure on student motivation, affective learning, and classroom climate." *Communication Education*, vol.56, no. 1 (2007), pp. 1-17.
- [11] K. F.Hew and W. S. Cheung, "Use of Facebook: a case study of Singapore students' experience," *Asia Pacific Journal of Education*, vol.32, no. 2 (2012), pp.181-196.
- [12] K. M.Titi, and A. Muhammad, Improvement Quality of LMS Through Application of Social Networking Sites. *International Journal of Emerging Technologies in Learning*, 8(3), (2013). Pp.48-51.
- [13] K.huwe, "Twitter and Facebook Open the Door to Collaboration," *Computers in Libraries*, vol.32, no. 8 (2012), pp. 27-29.
- [14] M. K. Kabilan, N.Ahmad, and M. a. Abidin, "Facebook: An online environment for learning of English in institutions of higher education? " *The Internet and higher education*, vol.13, no. 4, (2010), pp. 179-187.
- [15] M. C.Pretorius and D.Villiers, "Evaluation of a Collaborative Learning Environment on a Facebook". *Electronic Journal of Information Systems Evaluation*, 16(1), (2013). Pp.58-72.
- [16] M.Amasha and S.Alkhalaf, "The effect of using facebook markup language (fbml) for designing an e-learning model in higher education," *international journal of research in computer science*, vol.5 (2015), pp. 1-9.
- [17] M.Amasha and S.Alkhalaf, "Using RSS 2.00 as a Model for u-Learning to Develop e-Training in Saudi Arabia". *International Journal of Information and Education Technology*, 6(7),(2016),p.516.
- [18] M.Tower, S.Latimer and J.Hewitt, "Social networking as a learning tool: Nursing students' perception". *Nurse Education Today*, vol.34, no. 6 (2014), pp. 1012-1017.
- [19] O.Sanghee and S. S.Yeon, "Motivations for sharing information and social support in social media: A comparative analysis of Facebook, Twitter, Delicious, You Tube, and Flickr", *Journal of the Association for Information Science & Technology*, 66(10), (Oct2015). Pp.2045-2060.
- [20] P.Ractham and L.Kaewkitipong, "The use of Facebook in an introductory MIS course: Social constructivist learning environment," *Decision Sciences Journal of Innovative Education*, vol.10, no. 2 (2012), pp. 165-188.
- [21] Q.wang, H. l.woo, C. l.quek, Y.yang, , and M.liu, "using the facebook group as a learning management system: an exploratory study . *brith journal of education technology* , 43(3), (2012), Pp. 428-438.
- [22] R. A. Sánchez, V.Cortijio and U. Javed, "Students' perceptions of Facebook for academic purposes," *Computers & Education*, vol.70 (2014), pp. 138-149.
- [23] R.Junco, The relationship between frequency of Facebook use, participation in Facebook activities, and student engagement." *Computers & Education*, vol.58, no. 1 (2012), pp. 162-171.
- [24] S. Aydin. "A review of research on Facebook as an educational environment". *Educational Technology Research & Development* , 60(6), (2012),PP. 1093-1106.
- [25] S. G.Mazman and Y. K. Usluel, "Modeling educational usage of Facebook." *Computers & Education*, vol.55, no. 2 ,(2010), pp.444-453.
- [26] S. H.Sarapin and P. L. Morris,"Faculty and Facebook friending: instructor–student online social communication from the professor's perspective," *The Internet and Higher Education*, vol.27 (2015), pp. 14-23.
- [27] S.Hamid, J.Waycott , S.Kurnia, and S.Chang, "Understanding students' perceptions of the benefits of online social networking use for teaching and learning," *The Internet and Higher Education*, vol.26, (2015), pp. 1-9.
- [28] S.Khana and , S. B. Tahir, "A Study on the Role of Facebook in E-Learning". *I.J. Education and Management Engineering*, 5, (2015),pp.1-11. doi:0.5815/ijeme.2015.05.01
- [29] S.Manca and M.Ranieri, "Facebook and the others. Potentials and obstacles of social media for teaching in higher education," *Computers & Education*, vol.95, (2016), pp. 216-230.
- [30] T. Arabacioglu and R. A.Vural, "Using Facebook AS A Lms?" *Turkish Online Journal of Educational Technology*, 13(2), (2014). PP. 202-215.
- [31] T. J.Sinclair and R.Grieve, "Facebook as a source of social connectedness in older adults." *Computers in Human Behavior*, vol. 66 (2017), pp. 363-369.
- [32] V.Kharitonov and V. T . Sergei, "Dynamics and Structure of Dispute in Open Group of Facebook Social Networking Service in Terms of Teenagers' Homosexual Relations Education," *European researcher, Series A 5-1*, (2014), pp. 882-910.
- [33] Y. J. Joo, K. Y.Lim and N. H. Kim, "The effects of secondary teachers' technostress on the intention to use technology in South Korea," *Computers & Education*, vol.95, (2016), pp. 114-122.

AES-Route Server Model for Location based Services in Road Networks

Mohamad Shady Alrahhah
Department of Computer Science
King Abdulaziz University (KAU)
Jeddah, Saudi Arabia

Muhammad Usman Ashraf
Department of Computer Science
King Abdulaziz University (KAU)
Jeddah, Saudi Arabia

Adnan Abesen
Department of Computer Science
King Abdulaziz University (KAU)
Jeddah, Saudi Arabia

Sabah Arif
Department of Computer Science
Superior university Lahore
Lahore, Pakistan

Abstract—The now ubiquitous use of location based services (LBS), within the mobile computing domain, has enabled users to receive accurate points of interest (POI) to their geo-tagged queries. While location-based services provide rich content, they are not without risks; specifically, the use of LBS poses many serious challenges with respect to privacy protection. Additionally, the efficiency of spatial query processing, and the accuracy of said results, can be problematic when applied to road networks. Existing approaches provide different online route APIs to deliver the precise POI, but mobile user demand not only Accurate, Efficient and Secure (AES) results, but results that do not threaten their privacy. In this paper, we have addressed these challenges by proposing an AES-Route Server (RS) approach for LBS, which supports common spatial queries, including Range Queries and k-Nearest Neighbor Queries. We can secure the user location through the proposed AES-RS model because it provides the query results accurate and efficiently. The proposed model satisfies the primary goals including accuracy, efficiency and privacy for a location base system.

Keywords—Mobile computing; location based services; location based services (LBS) privacy; LBS accuracy; LBS efficiency; ubiquitous computing

I. INTRODUCTION

Recent years have witnessed the emergence of mobile computing technology as both a ubiquitous and extremely popular paradigm [1], wherein mobile users are capable of accessing information about nearby points-of-interest (POI). The devices used (smart phones, tablets, etc.) are integrated with a global positioning system (GPS), thereby facilitating the usage of location-based services (LBS). In short, location-based services are value-added services that leverage a user's geographic location when making queries. By geo-tagging a query, users are able to receive more personal, and valuable, results. While helpful, this service depends on many factors, including Points-of-Interest, the precise information surrounding the user and their current location, and the inherent need for privacy protection [7].

A basic architecture for location-based services is depicted in Fig. 1, where a mobile user connects to the LBS Server

through a communication network. The user then posts a query to the LBS for some location by sending his current location. The LBS then responds to mobile user with the geographically appropriate set of results.

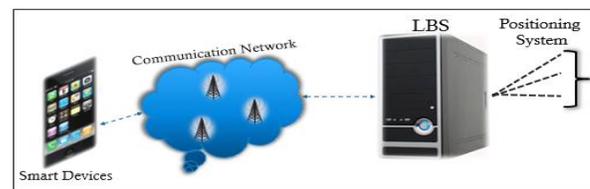


Fig. 1. A common LBS architecture.

In traditional mobile technologies, a mobile user posts a spatial query, q , such as a k Nearest Neighbor (k NN) or Range Query, to a server, requesting particular information; the server will then process the spatial query and return results to the mobile user with appropriate POI information [2], [3]. Without doubt, this “Point-to-Point access model” (POP) is quite ideal and easy to use. Unfortunately, several challenges arise for spatial query processing, such as when there are multiple users and issuing the same query, q , for their POI, or when all mobile users belong to the same location. In these scenarios, the server accrues additional overhead, and resources are wasted [3].

In a conventional mobile computing system, we find three primary goals with respect to a mobile user and the issuance of a spatial query:

- (G1) Accurate results,
- (G2) Efficient results and
- (G3) Privacy protection

G1 and G2 always present challenges due to the inherent realities of a mobile system. Accuracy and efficiency appear as luxuries in a system where both the user and the query are mobile. Additionally, LBS infrastructures and approaches have known limitations with respect to G1 and G2. In terms of G1 for LBS systems, a very famous framework “*SMashQ*” was proposed [5], which supports k NN query processing. The main

purpose of *SMashQ* was to leverage online route APIs, such as Google Maps, Yahoo Maps, Bing Maps, etc. to provide accurate query results for live travel in real road networks. While novel and an advancement in this research domain, *SMashQ* suffered with efficiency. Each time a user posts a query, q , to any LBS server, the LBS in turn would call the online route API for the most recent results and then return the results back mobile user. In short, the query response times were tragically slow. As expected, the proposed system was very accurate; the overhead of repeated queries on the server, followed by the server repeated calling route API, decreased the entire system efficiency. To overcome this problem, a more efficient approach was proposed, "*Route Server (RS)*" [6]. The primary goal of *Route Server* was to enhance the system efficiency with respect to query response time by reducing the number of route query requests. Furthermore, they used upper and lower limit calculation approach for this purpose. They also introduced a new mechanism such as "*Query Parallelism*" by parallelizing the query with different scenarios. *RS* was able to maintain accuracy while avoiding the repeated calls to the server and the online route API. The proposed approach seems to have addressed $G1$ and $G2$, leaving only $G3$.

The rapid growth and ever-increasing number of mobile users brings a variety of new challenges to LBS providers. Privacy protection, $G3$, is inherently challenging, as users, who want answers to their queries, must, in fact, reveal their locations and potentially sensitive personal data in order to receive answers to said queries.

- What if the mobile user's location is revealed?
- What kind of risks could be faced when mobile user's precise information becomes exposed?
- How can one protect mobile user's location privacy from bad actors?
- What factors should be involved under privacy protection?

These questions have formed the framework for a plethora of research within privacy and security of mobile data systems. A variety of approaches have been proposed to overcome privacy protection related challenges. Many depend on specific scenarios and basic privacy attributes such as the mobile user's identity, his current location, and time information [9]. For instance, a mobile user, who is at an unknown and unimportant location, may have no issue in sharing his personal data. But if the same mobile user is inside a residence or within its proximity, this location data may inadvertently reveal addition information that an adversary could misuse. Accordingly, many privacy attacks were identified, effectively creating a taxonomy of attacks, and respective solutions were proposed, each with its advantages and disadvantages.

A. Location Privacy Attacks

LBS location privacy attacks depend on protection attributes, described previously. Therefore, based on these protection attributes, we have classified Location Privacy Attacks into two major categories as follows (Fig. 2):

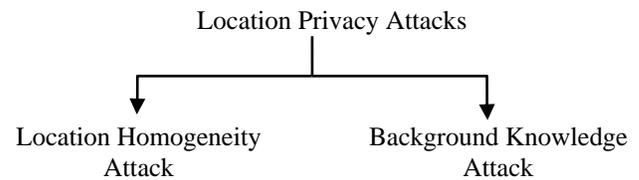


Fig. 2. Classification of location privacy attacks.

1) Location Homogeneity Attacks

Location Homogeneity attacks are one the most common attacks seen within LBS systems. They take advantage of the rare case in k -anonymity, where a sensitive value is indistinguishable and posted along a set of k -cluster values. Despite the dataset being k -anonymized, the sensitive value is revealed by any adversary [8], [9]. Additional homogeneity attacks include map utilization by reducing the area. In this case, the adversary reveals the diversity of the position information by analysing some location related information.

2) Background Knowledge Attack

In this attack, the attacker exploits the mobile user's contextual information and is able to accurately predict precise data. The contextual information of the user provides the background knowledge to the malicious attacker. In short, the attacker is able to leverage the background knowledge to prune the set of possible answers.

- **Maximum Movement Boundary Attack** is another background knowledge based attack approach used by adversaries to reveal mobile user's actual information. The adversary discovers the mobile user's region by identifying the maximum movement between two successful POI against posted queries in that specific region [10].
- **Multiple Query Attacks** The attacker follows the query log and identifies the query posted or updated frequently within a specific interval. The attacker effectively shrinks the specific region based on where he got consecutive query updates of a particular k -anonymity set and corresponding actual query [9], [11].
- **Context Linking Attacks** are categorized into three groups: personal context linking attacks, probability distribution attacks, and map matching attacks. Personal context linking attacks are related to the personal contextual information of a mobile user, which might be belong to his preferences or POI. Whereas probability distribution attacks are based on the high probability function of mobile user's location position. An adversary discovers the user's most frequent visited location position, along with a particular time span, and then applies a probability function to identify his precise information. Finally, Map matching is the third context linking attack, wherein a mobile user can be traced for a certain location by removing all irrelevant regions from the Map. Moreover, in order to leak the actual location information, an adversary could use the semantic information gained from the Map [12].

B. Location Privacy Approaches

A variety of approaches have been proposed to solve the aforementioned privacy attacks.

1) K-anonymity

One of the most commonly used approaches for location privacy preserving in LBS system is “K-Anonymity”, which insures that the precise information of targeted mobile user is indistinguishable from the value of set K-1 posts to LBS server. We can find out the probability [13] to trace the actual user’s data as follows:

Let’s have K a set of position of all anonymity users $K = \{k1, k2, k3, \dots, kn-1\}$. Therefore Probability of target user could be discovered as: $1/K$ (1)

The basic idea of k-anonymity to protect location privacy was demonstrated by Gruteser and Grunwald [31]. The theme of k-anonymity was that a mobile user can post a query, q, to the LBS server with an obfuscation area, along with k-1 anonymity positions of other users, rather than sending his precise location position. Certainty, k-anonymity approach is better in order to achieve the location privacy in LBS system; but in some cases, there are serious challenges when using this approach as follows:

- Homogenous Attack.
- Background Knowledge Attack.

2) Cryptography Based Approaches

Cryptography is another powerful approach to preserve a user’s location privacy from malicious attackers in a LBS system. The core idea behind cryptography based approach is utilization of encryption and decryption schemes for precise data that need to be sent over a network. A mobile user posts a query over the network; this query includes his secret data, which is encrypted by apply some particular algorithms at mobile user’s end. The same algorithm is available at server side to decrypt the data sent by user and utilized for further processing. The use of encryption and decryption schemes is dependent on the required level and kind of privacy. Cryptography approaches are classified into two main phases and then sub types as shown below in Fig. 3.

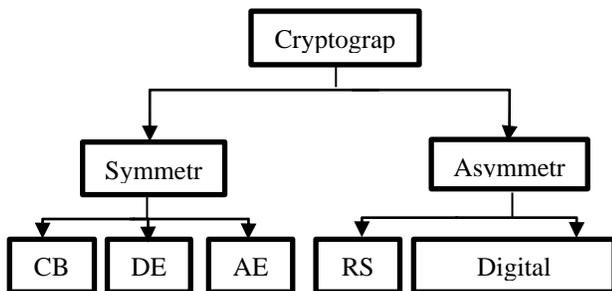


Fig. 3. Cryptography classification.

Certainly cryptography approach is very secured and implementable for LBS system but In contrast, a big challenge for cryptography based approach is the requirement of a massive level of computation during encryption and decryption takes more time than the system required. In LBS system, time is very significant attribute in order to provide the results

efficiently. However, implementation of cryptography might be costly regarding to this factor [4].

3) Mix Zones

Beresford introduced a new approach as “Mix Zone” for privacy location protection in mobile computing system [14]. Main theme of Mix zone was to conceal the precise location position of mobile user in his current locating region just like showing that “No existing in this area”. Once a mobile user enters in a mix zone area, his ID is shuffled by all other users belonging to that particular zone and the user’s precise location is protected. The major challenge for this approach is that an eavesdropper can easily find out the sensitive data of multiple mobile users through limited mix zone area [15].

4) Position Dummies

Leading to privacy location protection in LBS system, a new approach was introduced as “Position Dummies”. The fundamental principle of position dummies approach is that, user sends his actual position along with number of dummy location where mobile user’ precise information is indistinguishable [16]. Once user change his position from A to B with (x, y) coordinates, he posts a new query by sending his current position along with new dummies according to new place as shown in Fig. 4.

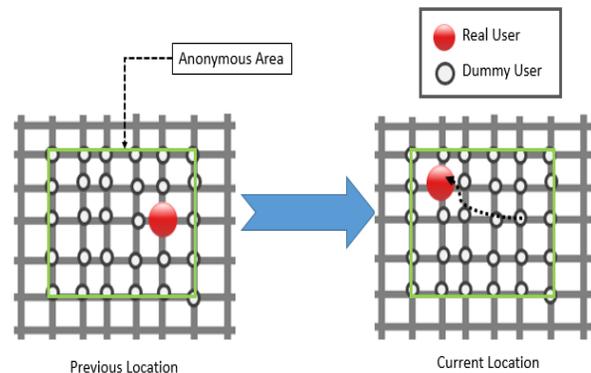


Fig. 4. Dummies on changing position.

In past, it has been remained a major challenge for dummy position that how to generate the number of dummies that have to post along user query to find any route path or POI [17], [26]. Later on, this challenge was overtaken by introducing different tools to generate these dummies [18]. In this paper, we also have proposed an efficient algorithm to generate the dummies at user end and then post to LBS along actual data. Position dummies have been considers the most approachable technique to secure the user’s precise location information. In our proposed AES-RS approach, we have implemented “position dummies technique” and made a secured route server approach for location based services in road network as discussed in other section.

The remainder of the paper is organized as follows. Section II describes related work, while the proposed AES-RS approach is discussed in Section III. In Section IV, we have discussed the implementation and results. Finally, Section V discusses the conclusion and future work directions.

II. RELATED WORK

In this section, we have illustrated the existing approaches utilized by others, advantages, limitations and future perspective directions to provide privacy for location protection of these approaches.

W. Sun, C. Chen and B. Zheng [19] emphasized road networks query processing approach. They proposed Network Partition Indexing (NPI) an Air Indexing that was supportive for spatial queries such as Range Query, CNN Query and kNN query. The basic idea of NPI was the processing of these spatial queries on road network by splitting the whole road network into small number of regions. They consider the road network concerning area as a grid G , and make its partition into number of cells where some information like upper and lower limit of each cell, border point and data segment was pre-computed to utilized in future query processing. Once mobile user posted any spatial query for a POI or route path, using these precomputed parameters, server broadcast the results in response through wireless network. They implemented NPI approach in real application and evaluated valued results. The one major challenge using this NPI approach was lost of information in case of link error over the network. They considered error-resilient and efficiency as future related challenges.

Z. Shao, D. Taniar and K. Adhinugraha presented Range-kNN queries supportive approach for privacy protection in [20]. The proposed algorithm was basically consisting of two major parts. In first part, they presented a new approach as Landmark Tree (LT) that was used to discover an appropriate landmark area by concealing the actual user's actual position. For LT, only a radius as parameter was required from mobile user for Range-kNN query implementation. After discovering the query range, another part as search algorithm was implemented to find out the most nearest neighbor from LT. In shortly, first part is responsible to find position inside the query range whereas second part is responsible to discover the location position from outside the range such as iNN ($i > 1$). The proposed algorithm was implementable limited to static objects but not for complex moving objects in real time applications.

B. Niu et al introduced Caching-Based approach for location privacy protection of user's position in Location Based Service system [21]. Caching based approach leads basically two algorithms such as CaDSA that was related to k-anonymization to improve privacy through utilizing caching dummy selections. Leading to CaDSA the author discovers some other performance effecting attributes such as how to normalize distance and how we can make sure the data freshness. Leading to privacy enhancement, the second algorithm called "enhanced CADSA" was proposed. Admittedly, the proposed algorithms provide privacy in location but the overhead of frequent queries to LBS makes the system performance down.

In [22], [23], the authors emphasized on location monitoring challenge for real time distributed system in mobile environment. According to author, the mobile objects should itself be responsive rather than increasing load on central server for objects related computation. In order to develop such

a responsive system, they make a set of assumptions such follows:

- The Moving Objects (MOs) have ability to locate its position.
- MOs have ability to determine their velocity vector.
- All MOs existing in mobile environment have ability of computation for assigning tasks.
- There is a synchronized clock among MOs.

They considered that in mobile computing system a distributed approach should be discovered that support *continues moving queries* along moving objects and proposed "MobiEyes". Furthermore, they brought in some optimization approaches constrict self-computation power at MOs end. Admittedly, the proposed approach is valuable but assumptions for such system are still challenges and future work for LBS system.

More on privacy protection, as discussed above Route-Server approach is one of the most efficient and accurate query results providing approach in LBS road networks. But the major challenge for RS was privacy protection of mobile user's precise information from adversary who can infer the faulty information in real data when a mobile user wants to post a spatial query for any route path or POI. We grouped privacy goal as G3 in above section.

Leading to G3, Privacy protection is another major challenge for LBS in road networks as it is very common practice to send some personal information when user issues any query for some POI information such as cinemas, bars, friend's location or any route path on a road network. For instance, Let's have a set Q of queries $\{q_1, q_2, q_3 \dots q_n\}$ where each $q \in Q$ belongs to Q set and posted as a route query, it will allow to an adversary to infer some false information by revealing mobile user's precise information [4] which is a big challenge for "Route Server" approach. In order to improve the privacy factor in Route Server algorithm we have proposed AES-RS a new secure approach presented in next section.

III. AES-RS SYSTEM MODEL

This section consists of proposed AES-RS system architecture which is essentially enhanced Route Server Architecture. One of the major components of AES-RS model is middleware Location Server that must be considered carefully. However before moving toward AES-RS model, we must introduce briefly the common models of location servers (LS) that are being used in LBS system [27], [28], [30]. These models are assorted into three basic categories including Untrusted Location Server (ULS), Trusted Location Server (TLS) and Peer to Peer based network (P2P) [29] where each model consist of three components as Mobile User Devices, Location Server and clients. In basic scenario each client interact with location server for desired POI or location finding, Location server further contact with clients to get the requested position. From Fig. 5(a) that elaborates the untrusted location server model, Fig. 5(b) shows the trusted location server using anonymizer that ensure trustworthy to deal with dummy position based request model or k-anonymity model

and Fig. 5(c) describes the third option as peer to peer network where each mobile user could interact with other mobile users or devices to find out the desired location or POI [8].

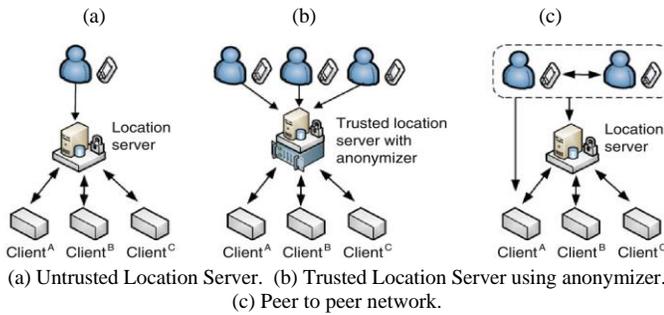


Fig. 5. Common LBS models.

Subsequently AES-RS is dummy position based model where a request is made along with number of dummy positions, however based on model features, we have selected second option as Trusted Location Server to ensure the provision of locations to mobile devices or users with privacy.

A. AES-RS System Architecture

AES-RS (Secured Route Server) architecture is enhancement by location privacy perspectives in Route Server Architecture proposed by [6]. AES-RS system architecture consists of three major entities such as mobile user, LBS and Route API. In AES-RS, mobile user part is now differ as in RS architecture as shown in Fig. 6. We implemented dummy position approach to protect user’s location privacy where a mobile user locating to grid area G post a query q along multiple dummies to AES-RS for any route path or POI. AES-RS executes that query, find out the required results from local Log “L” if find then return the required query results to user otherwise call Route API for the latest results.

In order to approach the goals G3 discussed in previous section, we have modified the definition 2 as “query results” for range and KNN query. Let’s a query q a set of dummy positions along actual location locating to Grid G and having time limited T, the results for Range query is:

$Q = \{k_1, k_2, k_3 \dots k_n\}$ then the query resulting definition should be modified q by Q, considering the multiple positions instead of single actual position. However,

$$R = \{p \in \mathcal{P} : \tau t_{now}(k \in Q, p) \leq T\}$$

And for KNN with K size

$$R = \{k \in Q \in \mathcal{P} : \tau t_{now}(k \in Q, p) \leq \tau t_{now}(k \in Q, p'), p' \in \mathcal{P} = -R\}$$

According to our AES-RS approach, before posting query to LBS, measure the minimum (L lower limit) and maximum (U upper limit) width and height of the specific area called grid “G”. The purpose to determine (L, U) coordinates is to make partition of “G” into equal number of cells “Ci”. Each cell (E, V) ∈ C representing that cells are connected through set of V Vertices and E Edges where (v ∈ V) and (e ∈ E) as shown in Fig. 7. Further to generate dummy positions, vertices are

calculated beyond each cell and one cell position is attached to mobile user’s actual position. Finally, an array is generated that contained all dummy K positions and index of actual user’s position by following the proposed algorithm DDA (Dummy Data Array).

Algorithm: DDA (Dummy Data Array)

Input: User location (X, Y), Anonymous_Area A, Anonymity_Number K;

Output: array[K(x,y) + (X,Y)]

Procedure:

- 1: $G(L, U)$ // Calculate Both Height and Width, U,L limit.
- 2: $C \leftarrow \sqrt{G}$ // Calculate Number of cells in G
- 3: $(V,E) \in C$ // Determine vertices and edges of each cell.
- 4: $P_x \leftarrow \text{Random}(0, v(C-1)), P_y \leftarrow \text{Random}(0, v(C-1))$
- 5: array[0 to C][0 to C] // Initialize 2-D array
- 6: $i = 0, j = 0, x, y = 0$ // Initialize values upto x-axis, y-axis
- 7: **While** ($i < (C-1)$) // Fill array with dummy positions
- 8: **While** ($j < (C-1)$)
- 9: **if** ($C_i.posX \neq X$ and $C_j.posY \neq Y$)
- 10: $x \leftarrow C_i.posX, y \leftarrow C_j.posY$
- 11: array[i][j] ← x, y
- 12: j ++; // Repeat step 8
- 13: **end if**
- 14: **end loop**
- 15: i ++; // Repeat step 7
- 16: **end loop**
- 17: add P_x, P_y in array
- 18: **Return** array

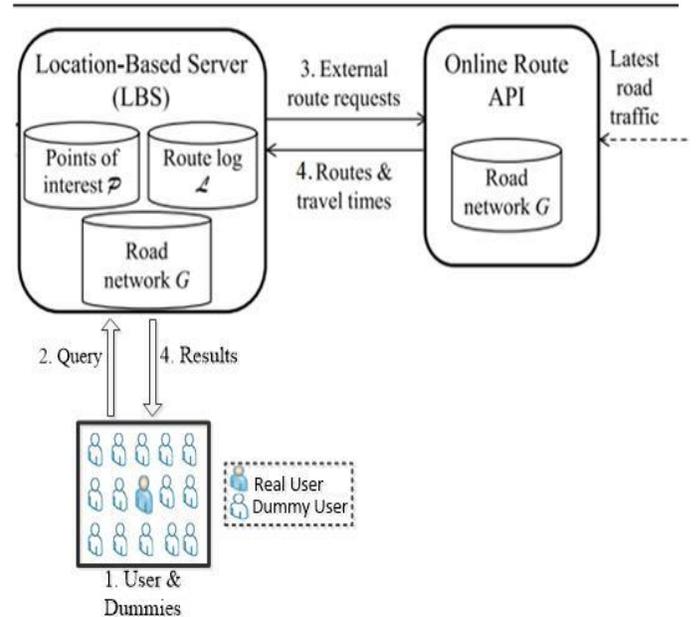


Fig. 6. AES-RS system architecture.

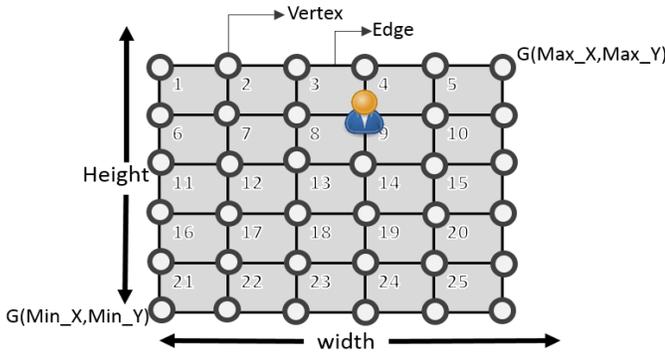


Fig. 7. Grid partition into cells.

According to DDA algorithm, it takes three input parameters as (X, Y) coordinates of user locating at current position, anonymous area A which is required to generate anonymity data and K number of dummies which are required to generate. It consider the anonymous area A as grid G and first calculate the upper and lower limits of whole anonymous area with respect to height and width denoted as $\langle \text{Min}_X / \text{Min}_Y, \text{Max}_X / \text{Max}_Y \rangle$. By using computed LU limits, anonymous area A partitioned into equal number of cells ($C_i \in G$) according to given input number of K as in equation 2 that was discovered by equation 3.

$$|C1|C2|C3|C4| \dots |Cn| = 1 \quad (2)$$

$$\text{Number of Cells} = \sqrt{G} \quad (3)$$

Once, number of cells are defined, it calculates the vertices and edges beyond each cell mentioned in step 3. Now, assign the mobile user's current location (P_x, P_y) to one random cell from G . Next, declare an array that will contain all the dummy positions and fill it according to number of cells because each cell is located with one dummy position. Once the array with dummy positions is filled, it adds the index of user's actual position in array and return.

B. AES-RS for Spatial Queries

As AES-RS is supportive for spatial queries such as Range query and KNN query as well. In this section, we present the consequences of Secured Route Server approach for spatial queries for a given query point q , along with a value d and data set P that reduce the number of requests. As described above, for AES-RS approach, we have used a Trusted Location Server (TLS) which ensure that only actual query request posted from mobile device to TLS along with set of dummy locations array will be computed to determine the POI or desired location. However, there will not be any change in spatial queries.

For range query in AES-RS, it first comport the distance range search for data set P on G road graph from q query point, denoted as $\text{range}(q, d, P) = \{o \mid o \in P \wedge \|o, q\| \leq d\}$ and then store the retrieved results from range in a set R . Similarly for KNN query with given query point q with data set P on G road network, a K Nearest Neighbor (KNN) query determine the k objects in P whole network distance which is represented as follows:

$$kNN(q, k, p) = \{O = U_{i \in [1, k] o_i} \mid O \subseteq P \wedge \forall o \in P - O, \forall o_i \in O, \|q, o\| \geq \|q, o_i\|\}$$

Unlike range query, KNN query doesn't have the fixed area for searching and contingent upon the current location of query point q and k value it find out the candidate point by defining upper and lower bounds.

C. AES-RS Effects on Accuracy

The one objective of RS algorithm was to provide accurate query results. As accuracy assurance in RS algorithm was achieved by calling route API frequently to get most updated query results and generate $\log L$ for Ψt routes that is validate till δ expiry time otherwise expire routes Ψt . In case of dummies along actual position, certainly it requires larger space to manage $\log L$ but no effect on accuracy in query results. However we can manage L by adding more memory space in the system.

D. AES-RS Effects on Efficiency

Efficiency was another essence factor in AES-RS and achieved by maintaining Ψt routes $\log L$. Definitely, it will affect on query response time because of requiring number of locations, doesn't matter it is dummy or actual location, LBS processing is required. But powerful approach as $\log L$, POI and Road Network G at LBS maintain route path and minimize the overhead of frequent route API calling.

IV. EXPERIMENTAL AND RESULTS

In this section we demonstrated our AES-RS approach and simulated to evaluate performance after enhancing RS approach by privacy factor. We used Riverbed Modeler academic edition 17.5 simulator tools that can be used to drive accuracy and performance in real network applications. Its old name was OPNet Modeler [24]. In our experiments, we used france_highway road network map provided in riverbed modeler. Further we selected multiple nodes as actual user location where he wants a route path to find out the nearest ATM from his current location using over the road network. In order to protect his precise information as current location, we draw multiple dummy positions $(k-1)$ then posted a query containing actual location along generated dummy positions to LBS server through a wireless network. The tenure in which multiple queries were posted to LBS and it respond back with query results was evaluated by setting 1 week duration. By following a basic wireless network routing approach, we used two Ethernet routers and sixteen dummy nodes from different locations were connected to each, which is further linked to an Ethernet switch and it post user's query to LBS for query results. Fig. 9 illustrates the rate at which data packets are being received by LBS server sending from Ethernet switch. The delay in transferring data packets to LBS server were calculated by using "Little's theorem" [25].

$$N(t) = A(t) + B(t) \text{ and } t \geq 0 \quad (4)$$

Where $A(t)$ is the number of data packets which are arrived at in time $(0, t)$ and $B(t)$ is the number of data packets that are depart from source location in time $(0, t)$.

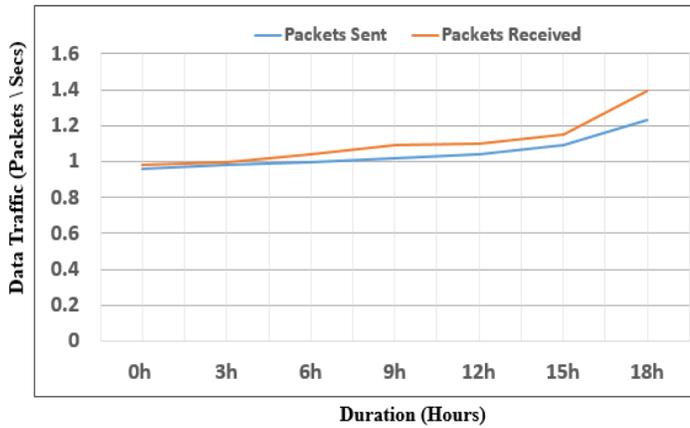


Fig. 8. Data transferring rate to LBS.

We observed that there were some other constituents like data transferring rate as shown in Fig. 8, the delay at Ethernet or wireless communication which could be cause of decreasing AES-RS system performance. In our case as shown in Fig. 9, during query transmitting over the network, delay size is very small in Ethernet and wireless which couldn't be reason to decrease system performance. In Ethernet, it becomes constant at a certain level by assuming that loss ratio in data packet is consistently zero. In contrast, delay variation increase and decrease after a certain time period which was overhead of using LBS as single server. It could be maintained by utilizing multiple LBS servers applying distributed approach.

The most significant part of AES-RS was to maintain LBS performance in order to provide user's query response accurately and efficiently by protecting mobile user's precise location. We evaluated LBS server performance when multiple query requests posted to it for any route path or POI and query processing at server side to return query results. Graph in Fig. 10 shows the number of requests posted to LBS server and its response quick by using log L, POI and Road Network G inside LBS.

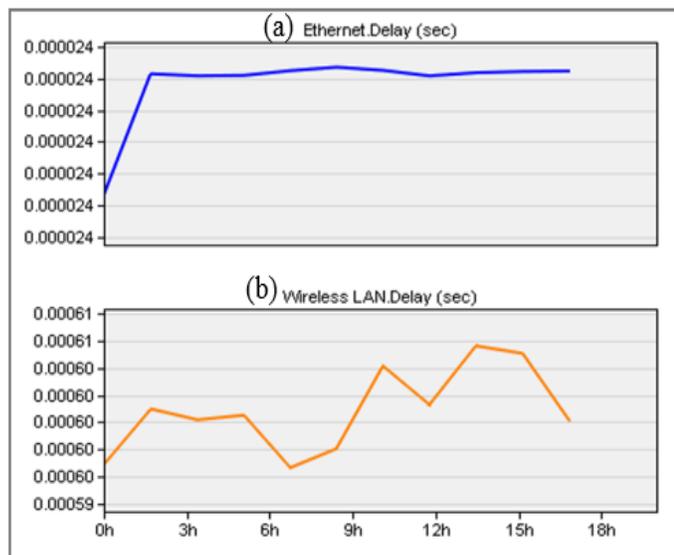


Fig. 9. Delay in ethernet and wireless LAN.

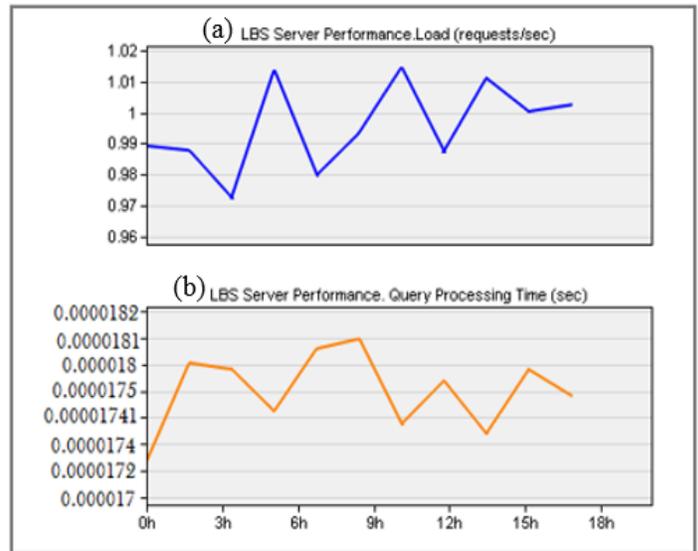


Fig. 10. LBS server performance.

We also evaluated the route API data access rate depicted in Fig. 10(a). The gradually decrease in graph 9 (a),(b) the clearly shows the advantage of log L, POI and Road Network G usage at server side that minimize route API hit rate due to availability of data at LBS server side. At initial stage, due to empty data in log it required to call route API for updated query results that increased route API retransmission attempts rate Fig. 10(b). But after a certain time t, when log L contained number of query results it decrease route API attempt rate. Furthermore, we assessed parallel route path approach proposed in RS algorithm and implemented in our experiments. Fig. 11(a), shows the results of data access delay through route API where we implemented parallel route path approach at LBS server side, it recognize firstly the required path against any mobile query, then it evaluate the relevant queries which are required route Path or POI from the same route. In this way, it minimizes the data access delay along query hits to LBS server.

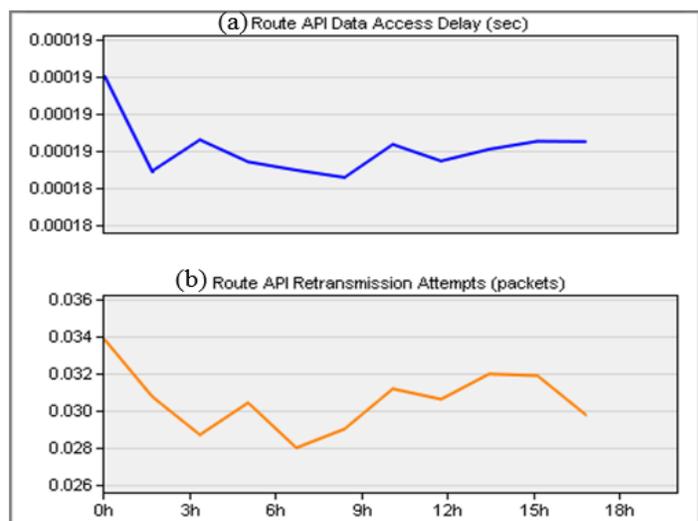


Fig. 11. Route API retransmission attempts and data access rate.

V. CONCLUSION

In mobile computing environment, every LBS system requires three primary goals such as accuracy, efficiency and privacy. A significant research has been attempted and delivered different LBS approaches to attain these goals. Route Server (RS) is one of the approaches that provide LBS system with accurate and efficient results for spatial queries. But RS algorithm didn't consider G3 as privacy goal to protect mobile user's precise information. However, by location privacy perspectives, we proposed AES-RS architecture which is an enhancement of RS algorithm and protect mobile user's precise location information from any adversary. On behalf of adversary attacks for LBS system, we discussed different kind attacks and various approaches to overcome these attacks. We also highlighted the advantages and limitations in existing approaches. After a critical analysis, we selected dummy position approach that ensure mobile user's privacy protection in RS algorithm and proposed a new approach AES-RS as Secure Route Server Architecture. As generating number of dummies for Dummy Position approach was a major challenge, we proposed an algorithm where dummy positions are generated at user end. Further in term of evaluation (G1, G2, G3) goals we simulated our approach using Riverbed modeler and generated different results. We discussed Ethernet and wireless WLAN as the factors that could be effective in efficiency in LBS wireless network system. From experiment results evaluation we can say AES-RS is an appropriate approach for LBS system which secure the user privacy for location protection by providing accurate and efficiently query results. By future perspectives, it required to examine the proposed solutions at large scale.

REFERENCES

- [1] G.H. Forman, and J. Zahorjan. "The challenges of mobile computing." *Computer* 27.4 (1994): 38-47.
- [2] W. Sun, et al. "An Air Index for Spatial Query Processing in Road Networks." *Knowledge and Data Engineering, IEEE Transactions on* 27.2 (2015): 382-395
- [3] M. Wernke, et al. "A classification of location privacy attacks and approaches." *Personal and Ubiquitous Computing* 18.1 (2014): 163-175
- [4] D. Zhang, C.-Y. Chow, Q. Li, X. Zhang, and Y. Xu. "SMashQ: Spatial mashup framework for k-NN queries in time-dependent road networks." *Distrib. Parallel Databases*, vol. 31, pp. 259-287, 2012.
- [5] L. Yu, and M.Y. Lung. "Route-Saver: Leveraging Route APIs for Accurate and Efficient Query Processing at Location-Based Services." *Knowledge and Data Engineering, IEEE Transactions on* 27.1 (2015): 235-249.
- [6] A. Civilis, C.S. Jensen, and S. Pakalnis. "Techniques for efficient road-network-based tracking of moving objects." *Knowledge and Data Engineering, IEEE Transactions on* 17.5 (2005): 698-712.
- [7] G.K. Shin, et al. "Privacy protection for users of location-based services." *Wireless Communications, IEEE* 19.1 (2012): 30-39.
- [8] W. Marius, et al. "A classification of location privacy attacks and approaches." *Personal and Ubiquitous Computing* 18.1 (2014): 163-175.
- [9] G. Ghinita, M.L. Damiani, C. Silvestri and E. Bertino. "Preventing velocity-based linkage attacks in location-aware applications" In: *Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems (GIS '09)*, Seattle, Washington, pp 246-255, 2009.
- [10] M. Gruteser and D. Grunwald "Anonymous usage of location based services through spatial and temporal cloaking". In: *Proceedings of the 1st international conference on mobile systems, applications and services (MobiSys '03)*, San Francisco, California, pp 31-42, 2009.
- [11] J. Krumm. "Inference attacks on location tracks", In: *Proceedings of the 5th international conference on pervasive computing (Pervasive '07)*. Springer, Toronto, pp 127-143, 2007.
- [12] G.Z. Ignatov, K.K. Vladimir, and R.S. Krachunov. "An improved finite-time ruin probability formula and its Mathematica implementation." *Insurance: Mathematics and Economics* 29.3 (2001): 375-386.
- [13] G. Singh, and A. Supriya. "A Study of Encryption Algorithms (RSA, DES, 3DES and AES) for Information Security." *International Journal of Computer Applications* 67.19 (2013): 33-38.
- [14] A.R. Beresford, F. Stajano. "Mix zones: user privacy in location-aware services". In: *Proceedings of the second IEEE annual conference on pervasive computing and communications workshops (PerCom '04 Workshops)*, pp 127-131, (2004).
- [15] M.L. Yiu, C.S. Jensen, J. Møller and H. Lu "Design and analysis of a ranking approach to private location-based services". *ACM Trans Database Syst* 36(2):1-42, (2011).
- [16] K.G. Shin, et al. "Privacy protection for users of location-based services." *Wireless Communications, IEEE* 19.1 (2012): 30-39.
- [17] J. Krumm. "A survey of computational location privacy." *Personal and Ubiquitous Computing* 13.6 (2009): 391-399.
- [18] W. Sun, et al. "An Air Index for Spatial Query Processing in Road Networks." *Knowledge and Data Engineering, IEEE Transactions on* 27.2 (2015): 382-395.
- [19] Z. Shao, D. Taniar, and K.A. Maulana. "Range-kNN queries with privacy protection in a mobile environment." *Pervasive and Mobile Computing* (2015).
- [20] B. Niu, et al. "Enhancing privacy through caching in location-based services." *Proc. of IEEE INFOCOM*. 2015.
- [21] B. Gedik and L. Liu. "Mobieyes: Distributed processing of continuously moving queries on moving objects in mobile system." *Advances in Database Technology-EDBT 2004*. Springer Berlin Heidelberg, 2004. 67-87.
- [22] K. Jürgen. *Continuous queries over data streams-semantics and implementation*. Diss. Universitätsbibliothek Marburg, 2007.
- [23] V. Sercan, and E. ERDEM. "Design and Simulation of Wireless Sensor Network Topologies Using the ZigBee Standard." *International Journal of Computer Networks and Applications (IJCNA)* 2.3 (2015).
- [24] Little, D.C. John, and C.G. Stephen. "Little's law." *Building Intuition*. Springer US, 2008. 81-100.
- [25] L. Hua, C.S. Jensen, and M.L. Yiu. "Pad: privacy-area aware, dummy-based location privacy in mobile services." *Proceedings of the Seventh ACM International Workshop on Data Engineering for Wireless and Mobile Access*. ACM, 2008.
- [26] A. Monika, and P. Mishra. "A comparative survey on symmetric key encryption techniques." *International Journal on Computer Science and Engineering* 4.5 (2012): 877.
- [27] T. Jawahar, and N. Kumar. "DES, AES and Blowfish: Symmetric key cryptography algorithms simulation based performance analysis." *International journal of emerging technology and advanced engineering* 1.2 (2011): 6-12.
- [28] P. Gilbert, L.P. Cox, J.W Jung. "Toward trustworthy mobile sensing". In: *Proceedings of the 11th workshop on mobile computing systems and applications (HotMobile '10)*, Annapolis, Maryland, (2010) pp 31-36.
- [29] K. Barker, M. Askari, M. Banerjee, K. Ghazinour, B. Mackas, M. Majedi, S. Pun and A. Williams "A data privacy taxonomy", *Proceedings of the 26th British national conference on databases: the final frontier (BNCOD 26)*, Birmingham, UK, (2009) pp 42-54.
- [30] M. Gruteser and D. Grunwald, Anonymous usage of locationbased services through spatial and temporal cloaking, in *ACM MobiSys'03: Proceedings of the 1st international conference on Mobile systems, applications and services*, pp. 31-42, 2003.
- [31] A.R Beresford, and F. Stajano. "Mix zones: User privacy in location-aware services." *Pervasive Computing and Communications Workshops*, 2004. *Proceedings of the Second IEEE Annual Conference on*. IEEE, 2004.

Visualizing Computer Programming in a Computer-based Simulated Environment

Dr. Belsam Attallah

Assistant Professor, Division Chair, Department of Computer Information Science, Higher Colleges of Technology (HCT), UAE
Senior Fellow, Higher Education Academy (HEA), UK
Member, British Computer Society (BCS), UK

PAPER OUTLINE

The research paper covers the following:

- 1) An introduction to the research, its aim and objectives.
- 2) Literature review on the problem formulation: This covers the complexity acknowledged by researchers and educators of the programming process at different levels.
 - a) Traditional programming.
 - b) Multithreading programming (concurrency and parallelism).
- 3) How visualization techniques are employed in a collaborative and simulated virtual environments to facilitate the learning of programming.
 - a) How collaboration environments are exploited in the learning of programming.
 - b) The employment of various visualization tools in the learning of programming.
 - c) The application of virtual world technologies in the learning of programming.
 - d) Former applications of virtual world technologies in the learning of programming.
- 4) The conclusion and future scope of the research.

I. INTRODUCTION

There is significant research acknowledging the level of complexity in the computer programming subject generally and at the Higher Education (HE) level. Programming skills require in-depth understanding of the complex theoretical concepts within this subject, which are recognized to be difficult to grasp by learners due to lack of real-life representation. Students who struggle in understanding and learning the abstract concepts of computer programming are likely to either withdraw from their course or choose another career path that does not involve programming [1].

This paper focuses on identifying the challenges faced by novice programmers and HE students in learning the different levels of computer programming, and provides recommendations on the techniques and platforms needed to overcome these challenges. In addition to visualization, the paper also explores the advantages of simulation, collaboration, interactivity and experimentation to support the process of learning computer programming.

Abstract—This paper investigated the challenges presented by computer programming (sequential/traditional, concurrent and parallel) for novice programmers and developers. The researcher involved Higher Education in Computer Science students learning programming at multiple levels, as they could well represent beginning programmers, who would struggle in successfully achieving a running program due to the complexity of this theoretical process, which has no similar real-life representation. The paper explored the difficulties faced by students in understanding this challenging, yet fundamental, subject of all Computer Science/Computing degree programmes, and focused on the advantages of visualization techniques to facilitate the learning of computer programming, with recommendations on effective computer-based simulated platforms to achieve this visualization. The paper recommended the application of virtual world technologies, such as ‘Second Life’, to achieve the visualization required to facilitate the understanding and learning of computer programming. The paper demonstrated extensive evidence on the advantages of these technologies to achieve program visualization, and how they facilitated enhanced learning of the programming process. The paper also addressed the benefits of collaboration and experimentation, which are ideal for learning computer programming, and how these aspects are strongly supported in virtual worlds.

Keywords—Computer programming; programming; object-oriented programming; programming language; parallelism; multithreading; multithreading; concurrency;; visual; visualization; visual environment; virtual worlds; second life; virtualization.

GOALS AND METHODS

The goals of this research are to assemble literature related to the difficulties faced by novice programmers and students learning computer programming at the Higher Education (HE) level, investigating the advantages of program visualization techniques to this process and recommending an effective computer-based simulated environment to achieve this visualization.

Both quantitative and qualitative research methods have been applied to achieve the outcomes of this research (questionnaires, observations and students’ feedback). An intensive literature review has been carried out to document the problem formulation, and to support the research outcomes and recommendations.

II. RESEARCHERS AND EDUCATIONISTS ACKNOWLEDGING THE DIFFICULTIES FACED BY THE STUDENTS TRYING TO LEARN COMPUTER PROGRAMMING

A. Traditional Programming

Programming is “a central element of the discipline of computing, an important practical skill for computing, and an essential component of the undergraduate curriculum” [2]. A large number of researchers and educators investigated and confirmed the complexity of the programming theory process. Programming curriculum is an essential and fundamental subject in Computer Science degree programmes that all students in this field are required to learn [3]. Programming languages have extensive and complex syntaxes, which results in great learning difficulties for novice learners and a high dropout rate from qualifications including this subject [3]. In spite of the advances within other Computer Science fields; learners still believe that their computing courses are dominated by programming subjects [2].

There are various factors which may contribute to the loss of students’ interest in Computer Science degree programmes, the most significant of which is the difficulties faced by students in the programming module of these courses; these difficulties result in high failure and dropout rates in preliminary programming modules at the HE level, which could reach as high as 30%-50% [4]. The difficulty in learning and teaching programming concepts is, therefore, confirmed by the high rate of failure and withdrawal in the introductory programming courses at universities [5].

Computer programming forms a common issue of concern amongst many universities due to the problems faced by their HE students in this subject in their first year of studies. In [6], authors confirmed that programming is a compulsory subject and an essential component in Computer Science curriculum, and that many novice learners often drop out from their degree courses due to either performing poorly or failing in programming subjects, which are considered the most hated and feared areas in a Computing qualification. In emphasizing the difficulty of the programming process, the reference clarified that programming techniques and skills are also hard to teach, not only because the traditional teaching methods are not very effective in the areas of scripting and problem solving, but also because such skills are best learned through experience. The difficulty in teaching this subject becomes even more challenging when trying to teach object-oriented programming to beginning learners [6].

In addition to Computer Science studies, programming is a very common subject in many fields of technology that are taught by a large number of universities in the world, although some courses only deliver the basics of it [7]. Unfortunately, learners usually face difficulties in understanding this subject even in the introductory courses, as these difficulties are not only because of the complex theory concepts in the subject, but also in various issues related to program construction, which often resulted in decreasing students’ retention rate [7]. Novice learners in introductory programming courses are required to comprehend the concepts, syntax, and semantics of a programming language and then be able to apply their understanding in coding a program and solving programming

problems; therefore, students of such courses consider learning to program as a difficult subject [8].

In [9], authors explained that “*Programming is one of the essential and most difficult skills to learn in the computer field and other disciplines. Programming can seem more troublesome for novices who have not learned programming concepts, usage and other basic programming skills*”. Beginning learners of programming find it non-inspiring to learn this subject, and this is one of the reasons why the majority of students in this field cannot do coding by themselves [9]. A non-user-friendly graphical environment makes the learning of programming difficult and programming problems more complex; while an interactive learning environment, where support and guidance are provided for students would help in overcoming a large amount of these difficulties [9].

In [10], authors confirmed that due to the various difficulties faced by beginning learners when trying to understand and learn computer programming, a large number of them fail this subject, and consequently withdraw from their Computer Science courses. Despite the fact that researchers and educators identified the challenges faced by novice learners in this subject, they are still struggling to recommend effective measures to support practitioners in this challenging area [10]. The reference explained the outcomes of the research carried out on beginning HE students, who considered computer programming as a traditional theoretical subject (similar to history), and that it is based on reading rather than practicing. These outcomes also showed that the students felt demotivated to get involved in the learning process as they failed to understand the programming instructions or achieve encouraging results. The reference indicated that the highest complexity faced by their students in learning programming is not only the understanding of the basic concepts, but also in how to successfully apply these concepts in a more advanced construct. Although certain students understand the syntax and semantics of a programming language, they fail to employ them correctly to achieve a functional program [10].

The major cause of non-completion in Computer Science degree programmes, is the difficulties faced by students in the transition period from Further Education (FE) to HE, where many of them having either little or no confidence in their programming skills; therefore, one of the significant challenges in HE Computer Science education is to have an effective learning platform in order to achieve major enhancements in students’ understanding, learning and achievement in the programming subjects [11].

Despite the fact that programming nowadays is considered a highly valuable skill, novice learners often express strong reactions to learning this subject due to the difficulties they face in understanding it [12]. Not only students face difficulties in this field, but also its lecturers, who sometimes find programming issues more challenging than students do, e.g. ‘understanding programming structures’ and ‘designing a program to solve a certain task’ [12]. Research confirms that teachers of programming continuously investigated new methods to support their learners to overcome the difficulties

they face at the start of Computer Science studies [13]. Physical lectures and traditional teaching methods failed most of the time in encouraging programming learners to get involved in relevant programming activities [13]. The skills required by learners to become good programmers are far beyond the syntax and semantics of a programming language, and the complexity of this subject results in high levels of failure at the start of Computer Science studies, as learners consider that they do not even understand the most basic concepts of programming due to their abstract nature, which has no similar real life representation [13].

The 'Grand Challenges in Computing Education' Conference, hosted by the British Computer Society (BCS), 2004, indicated the teaching and learning of computer programming as a major concern within the academic community worldwide. It clarified that learners view this subject as 'dry' and 'boring' rather than 'enjoyable' and 'creative', and this has demotivated people to apply for Computer Science qualifications. The Conference added that this was also accompanied by poor achievement and retention rates in Computer Science courses, which resulted in the opinion that, even after graduation, students of Computer Science studies clearly expressing their dislike of programming and their unwillingness to study it [2].

B. Multithreading Programming (Concurrency and Parallelism)

This area is considered one of the most complex subjects in Computer Science studies. This is due to the high degree of complexity in its theory concepts related to the threading mechanism that is applied by the computer operating system in the processor and memory units, which accordingly, makes the programming of it even more difficult.

The different executions of a multithreaded program may present different sets of results based on the structure of the threads and the way they communicate with each other within the program. This non-deterministic situation makes a multithreaded program difficult to write, test and debug [14].

A number of researchers and educators confirmed the complexity of coding a multithreaded (concurrent and/or parallel) program. Multithreaded programs are not only extremely difficult to write, but they are also very difficult to analyze, debug, and verify, as these processes are much harder than those in a sequential program [15]. Research in this area emphasized the negative impacts of the non-deterministic situation in the multithreading process. Conventional wisdom has assigned the difficulties of understanding this process to non-determinism, as repeated executions of the same program given the same input value(s) could well show different behaviors [15]. The complexity of multithreaded programs lies in the large number of states that the program could possibly be in at any given time [16]. The process of debugging a multithreaded program is a challenging task that requires certain specialized knowledge and tools; this is due to the difficulty in determining the state in which the program was at the time of failure, which is a frustrating situation for developers [16].

These complex multithreading concepts are difficult to grasp by novice learners; this is because of the large number of false assumptions made by students on the scheduling process of multiple threads in a program, and that they are unable to imagine what actually happens during the program execution due to the non-deterministic nature of threads scheduling, which makes it extremely possible that successive executions of the same multithreaded program produce different outcomes [17]. It is also difficult to teach multithreading programming, as lecturers need to find a way to visualize these complex concepts to students to facilitate their understanding and increase their confidence regarding program testing and debugging [17].

In [18], authors explained the complexity of the multithreading concepts by clarifying the process of having multiple threads within one program. It indicated that each thread is performing a task that works separately from the rest of the program, which makes the concept difficult to understand by many programmers. In sequential programs, the lines of code written by programmers are executed sequentially, which is the reason behind not understanding the situation of having a number of little programs (i.e. multiple threads), each of which has its own execution sequence, running inside one large program [18].

Due to the increased requirements on maximizing computer performance and productivity, multithreading nowadays is unavoidable for programmers; however, multithreaded programs are particularly difficult to write and debug correctly, and they are much more demanding and challenging than writing and verifying a sequential program [19]. The complexity of multithreading programming is widely acknowledged; however, the necessity of it has become more urgent [20]. People are quickly overwhelmed by the concept of concurrency, as they find it much more difficult to understand and learn compared to sequential code, as partially ordered operations could well make even careful people miss possible thread overlaps [20]; while parallelism caused the computer applications to become more complex resulting in increased difficulties in their design, implementation, verification, and maintenance, which has become widely acknowledged by developers [21].

III. EMPLOYMENT OF VISUALIZATION TECHNIQUES USING COLLABORATIVE AND SIMULATED VIRTUAL ENVIRONMENTS TO FACILITATE THE LEARNING OF PROGRAMMING

A. Utilizing Collaboration Environments

Many researchers highlighted the advantages of 'Collaborative Learning' to facilitate the understanding and learning of computer programming, and that collaborative learning is strongly achievable in virtual worlds.

In [22], authors defined 'Collaborative Learning' as an effective teaching and learning approach that is focused on adding value to students' understanding via interacting with others, where they are encouraged to share ideas and talks. Virtual worlds, e.g. Second Life, provide the students with a new opportunity to have the experience of interactive education in a computer-based simulated environment that facilitates achieving the objectives of collaboration,

engagement and experimentation [23]. Collaborative simulation activities form around half of the reviewed education literature in virtual worlds (over 100 academic papers) [24], while educators have long employed role-playing and simulation as a pedagogic tool in the education sector [25].

The proceedings of the 2013 International Higher Education Teaching and Learning Association Conference: Exploring Spaces for Learning, handled the issue of engaging and retaining HE students using ‘cutting-edge’ technologies and innovative pedagogies, one of the major areas of which was: ‘Collaboration and immersion discover best practices in a virtual world of Second Life’ [26].

In the field of computer programming, collaborative environments offer significant support to students in programming activities, which is an effective approach for learning this subject [1]. In [27], author discussed the use of scientific visualization in the field of ‘Big Data’, indicated that the visualization process of scientific data, which is key to its analysis and understanding, is not a simple task to achieve. As human beings are ‘optimized’ to interact within a 3D world, a virtual world environment such as Second Life or OpenSim enables people to walk into a representation of their data, while collaborating and interacting with each other within the same virtual space [27]. This is largely applicable to visualizing the data of program variables and the program execution during runtime. Constructivist activities or problem-based learning, e.g. in Computer Science simulations, form the strongest examples of the use of virtual tools, as virtual worlds provide a strong support for collaborative work and learner interaction in a simulated environment [24] (see Fig. 1 below).

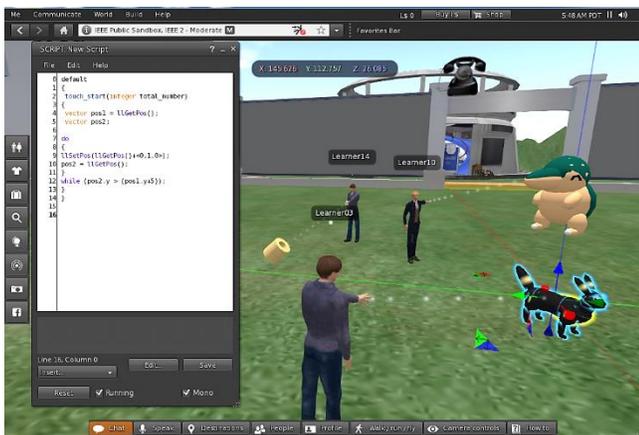


Fig. 1. Learning computer programming collaboratively in virtual worlds.

Research demonstrated that collaborative learning is considered an effective pedagogical feature for preliminary programming courses, as programming with peers is particularly appropriate for learning how to code a program [10]. The environment that promotes collaboration is able to offer important support for the activities to learn computer programming, as students need to communicate within their group, argue and give opinions, which encourages the type of reflection needed for effective learning of programming [10]. The virtual simulated environment “enables synchronous collaboration among students because the system permits two

or more avatars to edit the same object and share the same code while programming it” [10]. Constructivists and educators involved in constructionist learning might be able to recognize the potential in this environment, as it provides them with an accessible means for the creation of rich, immersive and appealing 3D framework for situated and experiential learning, and also communication tools to support dialogue and collaborative learning [10].

B. Application of Different Visualization Tools in the Learning of Computer Programming

In [1], authors explained the features and applications of different visualization tools/environments and a number of other similar program visualization software, which were created by developers to facilitate the learning and understanding of computer programming. These tools were ALICE, JELIOT, BlueJ and RAPTOR, which have been used to teach introductory computer programming courses (sources from 2000-2005). Students used these environments to drag and drop chunks of code into a canvas in order to achieve a visual representation of the computer program. This resulted in isolating these blocks of code from the rest of the program, which consequently, meant that these environments lacked both a comprehensive view of program visualization and also students’ engagement in a platform that does not support collaboration [1].

In [28], authors explained some other visualization tools such as jGRASP, which presented a static visualization of program execution, and ViRPlay3D/ViRPlay3D2, which presented some aspects of virtual world environments (avatars to represent learners exercising programming in a sandbox); however, this platform only facilitated the scripting process, but lacked support for collaborative learning, which is a strong feature offered by virtual world technologies.

C. Application of Virtual World Technologies in the Learning of Computer Programming

This paper focuses on a different visualization technique, which involves the application of virtual world technologies to visualize complex theory concepts of computer programming in order to enhance students’ understanding and learning of this subject at the HE level. The research involved HE in Computer Science students in a university center in England, UK. Visualization scenarios were designed in the virtual world of ‘Second Life’ to support the learning of challenging programming concepts as part of the HE Computer Science Year-1 and Year-2 programming courses. These visualization scenarios were scripted by the researcher using the programming language embedded within Second Life, called ‘Linden Labs Scripting Language (LSL)’. Many researchers confirmed the similarity of syntax and semantics between LSL and C++ language, which the selected HE students were studying as part of their Computer Science qualification. In [29], authors highlighted that the LSL’s main syntax and operators are expressive of those in Java and C++ programming languages. It explained that Second Life implements a compiler for the LSL language that contains C++ source code. In [10], authors confirmed the above by saying that the programming of objects in Second Life is performed by the use of LSL scripting language, the keywords

and structure of which are similar to those in C Language. The way the variables are declared in LSL language is the same as that in C++, and the multiple methods of creating a loop in LSL are almost identical to those in C++ [30].

In the visualization scenarios designed for this research, a number of eye-catching 3D objects were chosen to be programmed by learners within Second Life, e.g. Pokémon. This was meant to enable them to visualize the execution of challenging program instructions in order to improve their understanding of the relationship between the scripts and the actual implementation process and results. The type of the 3D objects was selected to add interest for learners and make their learning process enjoyable. These visualization scenarios enabled learners to view the immediate effects of script changes on each 3D object, i.e. visualizing the program execution. This assisted learners to understand how each program instruction works. Particular emphasis was placed on instructions related to loops and functions – for the Introduction to Programming course, and on classes and objects – for the Object-Oriented Programming course.

To demonstrate the benefit obtained from these visualization scenarios, below is an example of a program instruction handled by this research, which was visualized within Second Life. Learners found this instruction extremely difficult to understand and to imagine how it works and what the potential execution outcomes are. They considered visualizing this instructions' execution within Second Life very beneficial to their understanding of its function, structure and results. The advantages of visualizing programming instructions within the virtual platform were confirmed by students' answers to the following question asked by the researcher to the learners at the end of a whole session explaining the 'For Loop' in the physical classroom: "Which of these two For-Loop scripts result in moving the object six steps towards the X-axis?"

- (1)
For (i=0; i<6; i++)
 llSetPos(llGetPos()+<i,0,0>);
- (2)
For (i=0; i<6; i++)
 llSetPos(llGetPos()+<1,0,0>);

Some students were confident of their answer, and some were not. Those who were not 100% confident were permitted by the researcher to provide a prediction based on their current/background understanding of programming. It was a surprise to both the researcher and learners that all the answers of confident learners were wrong, while around half of not fully confident learners gave the correct answer; however, they were unable to correctly justify it. This was then followed by using the virtual environment to visualize the execution of the above code. When the students worked on moving their 'Pokémons' in Second Life, they were able to view the difference in the number of steps moved by the object as a result of the execution of each script sample. Following this visualization, they were able to provide confident explanations

on how each 'For Loop' of the above works. All learners confirmed that the process of explaining this instruction using the resources of a physical classroom, i.e. whiteboard, flipchart, projector, etc. did not facilitate the understanding of how this instruction works to the same degree that the visualization of it in virtual worlds did. This basic script was just an example of how confusing and complex programming instructions could be for novice learners [31].

The object-oriented programming visualization scenario, on the other hand, focused on using certain metaphors and sculptures within virtual worlds to visualize the challenging abstract concepts of 'classes' and 'objects'. This was meant to improve the understanding of what they mean, how they work, and why we need them in an object-oriented program. In this visualization scenario, learners were allowed the opportunity to compare, for example, between a portrait of a flower on the wall (a class) and an actual sculpture of the same flower planted in the ground (an object of the class) and their properties and functions, which mimicked classes and objects in an object-oriented program [31].

When learning programming collaboratively in virtual worlds, each student can have their own 3D object(s) to code; however, working collaboratively with other learners within the same virtual space enabled them the opportunity to communicate their ideas and script changes to each other. This allowed them to view the influences of these changes on the object behavior of their peers compared to that of their own objects, and consequently, modify their scripts successfully to achieve the required outcomes. Therefore, collaborative learning can strongly facilitate the learning of programming and develop the other set of skills necessary for this subject (see Fig. 2 below). The virtual world of Second Life "enables synchronous collaboration among students because the system permits two or more avatars to edit the same object and share the same code while programming it" [10].

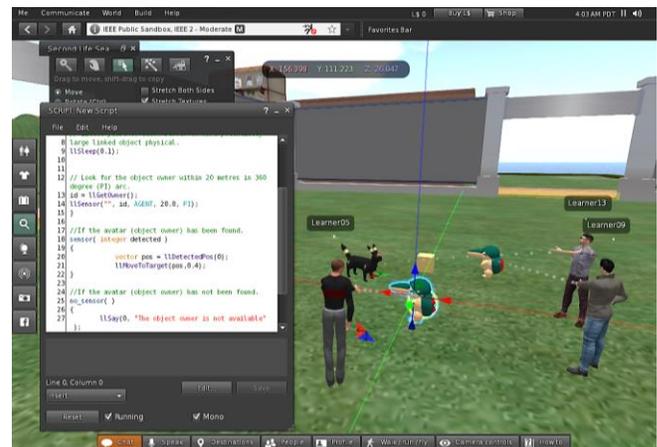


Fig. 2. Sharing the programming of 3D objects in virtual worlds between multiple learners.

A detailed questionnaire was distributed to students to capture their feedback on the application of virtual worlds to visualize the programming process [31]. The outcomes were as follows (see Fig. 3 below):

- The thoughts of slightly over 50% of learners, who initially considered this subject as difficult to understand and learn, were reversed following exercising programming in virtual worlds.
- Twenty-one percent more learners confirmed that effective understanding and learning of the complex theory concepts of this programming subject were achieved following the visualization activities in virtual worlds.
- Ninety-four percent of learners confirmed that affective quality was improved in the virtual platform. This figure was almost double the percentage obtained for affective quality in the physical world.
- Thirty-seven percent more learners confirmed that visualizing and learning this level of programming within virtual worlds is more engaging.

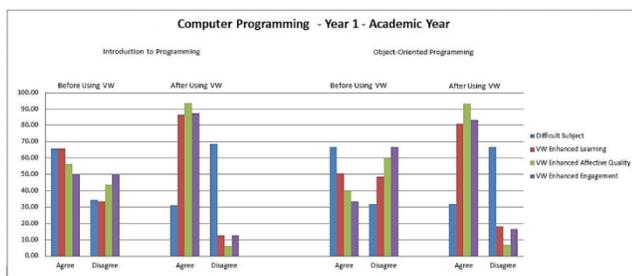


Fig. 3. Computer programming questionnaire, Year-1.

With regards to the introduction to object-oriented programming in Year-1, and as can be seen in Fig. 3 above, the questionnaire on the visualization scenario showed that:

- In agreement with the outcomes of the Introduction to Programming visualization scenario, the thoughts of slightly over 50% of learners, who initially considered this subject as difficult to understand and learn, were reversed following exercising programming in virtual worlds.
- Thirty-one percent more learners confirmed that effective understanding and learning of the complex theory concepts of this programming subject were achieved following the visualization activities in virtual worlds.
- Ninety-three percent of learners confirmed that affective quality was improved in the virtual platform. This figure was a lot higher than double the percentage obtained for affective quality in the physical world.
- Eighty-three percent of learners confirmed that visualizing and learning this level of programming within virtual worlds is more engaging. This figure was, again, a lot higher than double the percentage obtained for the physical world.

Moving to the Object-Oriented Programming, which is a more complex subject compared to normal programming (as highlighted earlier), this subject is delivered in Year-2 and the students were introduced to the subject towards the end of

their Year-1 introductory programming studies. It was reasonable to expect that students at this higher level (Year-2) would be more confident in learning programming compared to Year-1 students, as these Year-2 students have already studied the introduction to programming in their Year-1. However, the results of the below questionnaire revealed otherwise.

The questionnaire on the Object-Oriented Programming visualization scenario (which was designed by a different lecturer), revealed that as high as 77% of Year-2 students still consider the object oriented programming as a difficult subject. This confirmed that computer programming is an area of concern to HE students in Computer Science courses at all levels. The complexity faced by students in learning programming is not only the understanding of the basic concepts, but also in the process of applying these concepts correctly to achieve more advanced constructs. Although some students understand the syntax and semantics of a programming language, they fail to use it correctly to create a program [10].

The questionnaire showed that more than three quarters of students confirmed that Object-Oriented Programming is a difficult subject, and a slightly higher percentage of students agreed that the virtual world exercise improved their understanding and learning of the complex theory concepts of the subject, while 90% of them agreed to enhanced affective quality and 64% found this learning process more engaging [31] (see Fig. 4 below).

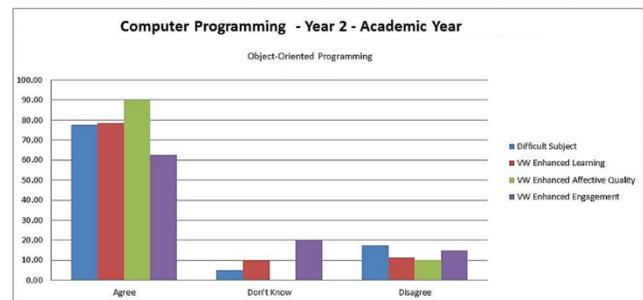


Fig. 4. Computer programming questionnaire.

On the other hand, the visualization scenario designed in virtual worlds to visualize the complex theory concepts of multithreading techniques (Concurrency and Parallelism) was applied on the BSc in Computer Science students, who found this module to be the most challenging amongst all the other modules. Before using virtual worlds to visualize the multithreading techniques, the researcher was used to drawing a number of sketches on the whiteboard for students to represent the computer Random Access Memory (RAM) and processor, with a number of arrows representing the data flow between these two components (for individual examples). More drawings and arrows were then added to show how the operating system controls the swapping and priority of tasks (threads) inside a computer and the time slots allocated to them within the processor. However, these sketches on the whiteboard could get very crowded and confusing for students, especially when more components are added (drawn on the board), e.g. the input/output devices and virtual

memory. In addition, there was no clear representation on the board of the sequence of actions and their individual consequences.

In order to visualize the multithreading techniques, the researcher carried out a thorough investigation for a comparable real-life example that requires a similar queuing process to receive the service (i.e. the queuing of threads in RAM by the operating system), and how the structure of the queue is affected by a higher priority arrival. The intension was to build the scenario in virtual worlds on the selected real-life example, in order to achieve a clearer and intuitive illustration for the students to enable them to compare the situation to that inside a computer system.

The outcome of the investigation was to choose a buffet restaurant example with a single restaurant keeper/waiter, where customers need to queue to get food, ice cream and drinks. The comparison between this real-life example and the multithreading techniques was as follows: The customers' queue represents the queue of tasks/threads within the computer RAM waiting to be served by the processor, while the food buffet, ice cream counter and the drink machine represent different resources/cores within the computer processor. The single restaurant keeper, who coordinates the providing of services, represents the operating system, while the restaurant tables and chairs represents the computer virtual memory having stand-by tasks (seated customers in the restaurant example) waiting for a space to join the queue in order to get served. Finally, the counter on the side having plates and cups, where the customer needs to go out of the queue to get a plate, represents the input/output devices in a computer system, where a task in RAM needing an input value cannot be served by the processor until it gets it. Fig. 5 and 6 show the virtual restaurant designed in Second Life to visualize the multithreading techniques.



Fig. 5. Multithreading Techniques visualization scenario (queuing technique in RAM with tasks in the virtual memory).



Fig. 6. Multithreading Techniques visualization scenario (swapping of tasks between the RAM and virtual memory).

In the virtual restaurant scenario, some students were required to act the role of customers queuing to get food, desserts or a drink (representing computer tasks queuing in RAM), while other students were required to sit down around the tables when the queue was full (representing tasks stored in the computer virtual memory) waiting for any of the queuing customers to finish, then the restaurant keeper (another student representing the role of the computer operating system) would ask them, one by one, to join the queue. Throughout this process another student is asked to act the role of a VIP customer who arrived to a busy restaurant having a full queue and a number of other seated customers waiting to be served (representing a high priority task joining a full RAM) [31].

Within this visualization scenario, a number of different multithreading situations were explained to the students, using the above restaurant metaphors, with their impacts and outcomes, e.g. when a higher priority task is placed by the operating system at the start of the queue in RAM changing the order of execution for all the remaining tasks, having a full queue with or without tasks in the virtual memory, and having a single core (Concurrency) or multiple cores/processors (Parallelism). Being part of this number of different situations and their visual impacts facilitated students' understanding of the complex abstract concepts of the multithreading process and the various factors affecting the execution of tasks in a computer system. In addition, the situation of role-playing the different computer components contributed greatly to this enhanced learning [31].

The observations by the researcher confirmed that the students found this visualization scenario extremely useful in understanding the different aspects of the multithreading process and the need for it. The researcher directed well-selected questions to the students (during and after the virtual exercise) to test their level of understanding and learning, and also to record their evaluation of their experience in virtual worlds.

The outcomes of the questionnaire distributed to students to capture and record their feedback on this visualization scenario showed the following outcomes (see Fig. 7 below):

- The thoughts of slightly over 50% of learners, who initially considered this subject as difficult to understand and learn, were reversed following exercising programming in virtual worlds.
- As high as 96% of learners confirmed that effective understanding and learning of the complex theory concepts of this programming subject were achieved following the visualization activities in virtual worlds. This figure was almost double the percentage obtained for the physical world.
- Hundred percent of learners confirmed that affective quality was improved in the virtual platform compared to 62% for the physical world.
- Hundred percent of learners confirmed that visualizing and learning this level of programming within virtual worlds is more engaging compared to 57% for the physical world.

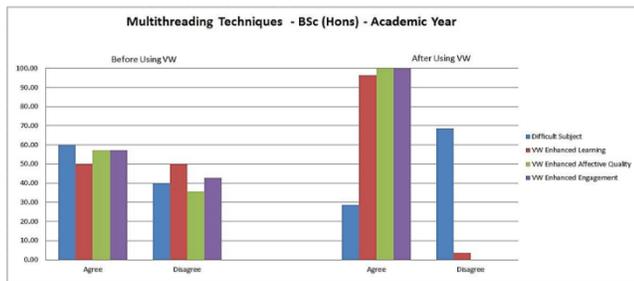


Fig. 7. Statistics of Multithreading Techniques, B.Sc. Students.

The process of visualizing the complex theory concepts of programming was more effective in virtual worlds as it engaged students in this immersive environment much more than the situation in the physical world where the viewer watches the program code passively. Interactivity and experiential learning were strongly achieved here. The virtual world environment inspired expressive and dynamic discussions on programming concepts, as students built their own visualization of the program, and followed the presentation of it more engagingly.

The researcher's observations throughout the different stages of this research demonstrated that students felt more relaxed in repeating their programming activities in virtual worlds when making mistakes or when not fully achieving their targets. This was due to the flexibility offered by the virtual world environment, and the fact that there were no physical consequences involved, e.g. being embarrassed in front of other students and/or the lecturer. This resulted in more engagement and involvement in the learning process, and consequently enhanced students' acquired skills and achievement in this field. As students were represented in avatars within the virtual world platform, they had less hesitation in asking basic questions or requesting more information. The facility of having a private channel in Second Life was very beneficial to students in carrying out

private chatting (via text) with the lecturer. This inspired more interaction especially for the shy students, and increased their self-confidence in discussing their concerns without feeling embarrassed for lagging behind others.

D. Previous Applications of Virtual World Technologies in the Learning of Computer Programming

As highlighted in the previous sections of this paper, research activities in the learning of programming area were explained by [10]. However, their paper indicated that although the main target of the research work was to investigate the possibility of using the Second Life virtual world as a platform for the teaching and learning of an imperative computer programming language, the research focused primarily on investigating the potential problems that could be faced by both teachers and students in this environment, and whether such problems could be solved and how.

In [1], authors carried out a study in Deakin University and Monash University, Australia, regarding the learning of computer programming in virtual worlds. It investigated the affordances of Second Life for 'experiential problem-based learning pedagogies', and the potentials and limitations of this platform for learning the programming subject. The study generated very positive answers in terms of the advantages of Second Life virtual world for learning computer programming.

In [32], authors explained an application of Second Life in the computing courses of the School of Computing, University of Portsmouth, UK. It described that Second Life was used in two areas: 1) Human Computer Interaction (HCI) Unit, and 2) Computer Engineering Projects Unit, both of which involve a great deal of programming requirements.

In [33], authors also studied the application of Second Life to engage and motivate the HE Computing students of the Computer Information Systems Department at Borough of Manhattan Community College, New York, USA. It explained that a teaching and learning platform was designed in Second Life to assist the students in overcoming the difficulties in their study. It clarified that the designed platform included a lecture area, group study rooms and interactive teaching and learning activities, which aimed at better engagement of students and the improvement of the retention data within the Computer Science programme.

In [13], authors introduced Second Life in the learning of computer programming in two higher education academic institutions in Portugal, where they used the 3D virtual world environment to visualize and contextualize some programming aspects. The use of visualization helped the programming students to better understand these aspects, because visual representations are easier to retain and handle, and that having an instant visualization of instruction results enabled students to directly judge whether their idea was right or wrong [13]. Second Life users were able to create avatars and 3D objects, and to program their behavior using the Linden Scripting Language (LSL); the benefit of this is the students' ability to execute the programming code concurrently and that several students are able to

simultaneously work over the same code and/or object, which provided the advantage of immediate presentation of program execution [13].

IV. CONCLUSION AND FUTURE SCOPE

Learning computer programming forms a cause of concern to a large number of novice programmers and students studying this field at the HE level. Research revealed that these concerns are the main reason behind HE students' withdrawal from their computing courses, achieving poorly or failing the modules that include programming concepts.

Research also showed that there are a number of software tools to visualize program structure for learners; however, the majority of them promoted static visualization, which did not generate the degree of support needed for the programming complex theoretical process.

This research demonstrated that there are strong indications of benefits of visualizing the program structure in virtual worlds, as this platform offers great advantages such as collaboration, simulation, interactivity and experiential learning, which are ideal for learning computer programming. This did not only cover enhancements to students' understanding of the programming complicated process, but also increased their engagement in the sessions, enhanced affective quality and improved their achievement.

The future scope could be utilizing virtual reality technology to facilitate the learning of programming with a comprehensive comparison between the advantages and limitations of both computer-based simulated environments. Aspects such as lecturer/students' satisfaction, ease of use and the technical issues involved could form the main points of the proposed comparison.

REFERENCES

- [1] Sajjanhar and J. Faulkner, "Exploring Second Life as a Learning Environment for Computer Programming," Deakin University and Monash University, Australia, Scientific Research, Creative Education, vol. 5, no. 1, pp. 53-62, January 2014.
- [2] McGettrick, R. Boyle, R. Ibbett, J. Lloyd, G. Lovegrove and K. Mander, "Grand Challenges in Computing Education," British Computer Society (BCS). The Computer Journal, vol. 48, no. 1, pp. 42-48, January 2005.
- [3] L. Morgado, B. Fonseca, M. Esteves, P. Martins, "Improving teaching and learning of computer programming through the use of the Second Life virtual world," The Polytechnic Institute of Leiria and the University of Trás-os-Montes e Alto Douro, Portugal, British Journal of Educational Technology, vol. 42, no. 4, pp. 624-637, July 2011.
- [4] S. Dasuki and A. Quaye, "Undergraduate Students' Failure in Programming Courses In Institutions Of Higher Education In Developing Countries: A Nigerian Perspective," American University of Nigeria. EJISDC, vol. 76, no. 8, pp. 1-18, 2016.
- [5] J. Kaasbøll, O. Berge, R.E. Borge, A. Fjuk, C. Holmboe and T. Samuelsen, "Learning Object-Oriented Programming," Norway: University of Oslo and InterMedia. 16th Workshop of the Psychology of Programming Interest Group, Carlow, Ireland, pp. 86-96, April 2004.
- [6] Miliszewska and G. Tan, "Befriending Computer Programming: A Proposed Approach to Teaching Introductory Programming," Victoria University, Melbourne, Australia. Issues in Informing Science and Information Technology, vol. 4, pp. 277-289, 2007.
- [7] Lahtinen, K. AlaMutka and H.M. Järvinen, "A Study of the Difficulties of Novice Programmers," Tampere, Finland: Institute of Software Systems, Tampere University of Technology. ACM Digital Library, vol. 37, no. 3, pp. 14-18, June 2005.

- [8] R. Matthews, H.S. Hinb and K.A. Chooc, "Practical use of review question and content object as advanced organizer for computer programming lessons," The University of Nottingham, Malaysia Campus & Multimedia University, Malaysia. Elsevier, Science Direct, vol. 172, pp. 215-222, January 2015.
- [9] M. Sasaki, S.M. Taheri and H.T Ngetha, "Evaluating the Effectiveness of Problem Solving Techniques and Tools in Programming," Gifu University, Japan. IEEE Xplore, Science and Information Conference, London, UK, July 2015.
- [10] Fonseca, M. Esteves, L. Morgado and P. Martins, "Using Second Life for Problem Based Learning in Computer Science Programming," Pedagogy, Education and Innovation in 3-D Virtual Worlds. The Polytechnic Institute of Leiria and the University of Trás-os-Montes e Alto Douro, Portugal. Journal of Virtual World Research, vol. 2, no. 1, ivvresearch.org, April 2009.
- [11] M. Huggard, "Programming Trauma: Can It Be Avoided?" Proceedings of the British Computer Society (BCS), Grand Challenges in Computing: Education. Newcastle, England, pp. 50-51, March 2004.
- [12] Costa and M. Piteira, "Learning Computer Programming: Study of difficulties in learning programming," IPS-ESTSetúbal & IPS-ESTSetúbal, Lisboa, Portugal. ACM Digital Library. ISDOC '13 Proceedings of the 2013 International Conference on Information Systems and Design of Communication, pp. 75-80, July 2013.
- [13] M. Esteves, B. Fonseca, L. Morgado and P. Martins, "Contextualization of Programming Learning: A Virtual Environment Study," Portugal: Polytechnic Institute of Leiria & University of Trás-os-Montes e Alto Douro. 38th ASEE/IEEE Frontiers in Education Conference, Saratoga Springs, NY, October 2008.
- [14] H. Cui, J. Wu, J. Gallagher, H. Guo and J. Yang, "Efficient Deterministic Multithreading Through Schedule Relaxation," Cascais, Portugal: Department of Computer Science, Columbia University. SOSP'11 - Proceedings of the 23rd ACM Symposium on Operating Systems Principles, pp. 337-351, October 2011.
- [15] J. Yang, H. Cui, J. Wu, Y. Tang and G. Hu, "Determinism Is Not Enough: Making Parallel Programs Reliable with Stable Multithreading," Columbia University, Communications of the ACM, vol. 57, no. 3, pp. 58-69, February 2014.
- [16] J. Roberts and S. Akhter, "An Introduction to Multi-threaded Debugging Techniques," Go Parallel, Intel Corporation, USA, 2011.
- [17] Malnati, C.M. Cuva and C. Barberis, "JThreadSpy: Teaching Multithreading Programming by Analyzing Execution Traces," ACM Digital Library, PADTAD '07 Proceedings of the 2007 ACM workshop on Parallel and distributed systems: testing and debugging, pp. 3-13, July 2007.
- [18] Rick, T. Mohiuddin and M. Nawrocki, "LabVIEW Advanced Programming Techniques," Chapter 9: Multithreading in LabVIEW. Boca Raton: CRC Press. Taylor & Francis Group, LLC, 2007.
- [19] M. Huisman and C. Hurlin, "Permission Specifications for Common Multithreaded Programming Patterns," France: INRIA Sophia Antipolis, December 2007.
- [20] E.A. Lee, "The Problem with Threads," Electrical Engineering and Computer Sciences, University of California, Berkeley, USA. Technical Report No. UCB/EECS-2006-1, IEEE Computer, vol. 39, no. 5, pp. 33-42, May 2006.
- [21] M. Duranton, D. Black-Schaffer, S. Yehia and K. De Bosschere, "Computer Systems: Research Challenges Ahead. The HiPEAC Vision 2011/2012," High Performance and Embedded Architecture and Compilation, Seventh Framework Programme. HiPEAC Compilation Architecture, October 2011.
- [22] A.E. Woolfolk, M. Hughes and V. Walkup, "Psychology in Education," Harlow: Pearson Longman, 2008.
- [23] Y. Huang, S. Backman and K. Backman, "Student attitude toward virtual learning in Second Life: A flow theory approach," Taylor & Francis [Online], Journal of Teaching in Travel & Tourism. Clemson University, Clemson, South Carolina, USA, vol. 10, no. 4, pp. 312-334, November 2010.
- [24] Duncan, A. Miller and S. Jiang, "A taxonomy of virtual worlds usage in education," British Journal of Educational Technology (BJET), vol. 43, no. 6, pp. 949-964, January 2012.

- [25] S. Fitzsimons, "An Exploration of Teaching and Learning in A Virtual World in The Context of Higher Education," PhD thesis, School of Education Studies, Dublin City University, Dublin, July 2012.
- [26] P. Blessinger and C. Wankel, "Proceedings of the 2013 International Higher Education Teaching and Learning Association Conference: Exploring Spaces for Learning," St. John's University, New York, USA. Conference organized by: UCF, HETL (Higher education Teaching & Learning), Faculty Center for Teaching & Learning, 2013.
- [27] Cioc, "Immersing Yourself in Your Data: Using Virtual World Engines to Solve "Big" Data," Astrobetter [Online], March 2013. [Accessed 12 July 2016].
- [28] J. Sorva, V. Karavirta and L. Malmi, "A Review of Generic Program Visualization Systems for Introductory Programming Education," Aalto University, Finland. ACM Digital Library. Vol. 13, no. 4, Article No. 15, November 2013.
- [29] W. Moldenhauer, J.C. Browne and C. Lin, "Bringing Verification to a Virtual World," CiteSeerX. Honors Thesis, Department of Computer Sciences, University of Texas, Austin, USA, May 2007.
- [30] J. Gomez, "Chapter 4: Logic," LSL Wiki [Online], 2012. [Accessed 02 May 2016].
- [31] Attallah, "The affordances of virtual world technologies to empower the visualisation of complex theory concepts in computer science: Enhancing success and experience in higher education," PhD, University of the West of England, June 2015.
- [32] J. Crellin, E. Duke-Williams, J. Chandler and T. Collinson, "Virtual Worlds in Computing Education," School of Computing, University of Portsmouth, UK. Taylor & Francis [Online], Computer Science Education Journal, Web-based technologies for social learning in computer science education, vol. 19, no. 4, pp. 315-334, December 2009. [Accessed 07 June 2016].
- [33] Y. Chen, J. Doong, and W. Ching-Song, "A 3D Virtual World Teaching and Learning Platform for Computer Science Courses in Second Life," Computer Information Systems Department, Borough of Manhattan Community College, CUNY, New York City, New York, U.S.A & Department of Information Management China University of Technology, Taipei, Taiwan. IEEE Xplore [Online], December 2009. [Accessed 09 May 2016].

Hybrid Technique for Java Code Complexity Analysis

¹Nouh Alhindawi

¹Faculty of Science and Information Technology
Jadara University
Irbid, Jordan

²Mohammad Subhi Al-Batah

²Faculty of Science and Information Technology
Jadara University
Irbid, Jordan

³Rami Malkawi

³Faculty of Information Technology and Computer Science
Yarmouk University
Irbid, Jordan

⁴Ahmad Al-Zuraiqi

⁴Faculty of Science and Information Technology
Jadara University
Irbid, Jordan

Abstract—Software complexity can be defined as the degree of difficulty in analysis, testing, design and implementation of software. Typically, reducing model complexity has a significant impact on maintenance activities. A lot of metrics have been used to measure the complexity of source code such as Halstead, McCabe Cyclomatic, Lines of Code, and Maintainability Index, etc. This paper proposed a hybrid module which consists of two theories which are Halstead and McCabe, both theories will be used to analyze a code written in Java. The module provides a mechanism to better evaluate the proficiency level of programmers, and also provides a tool which enables the managers to evaluate the programming levels and their enhancements over time. This will be known by discovering the various differences between levels of complexity in the code. If the program complexity level is low, then of the programmer professionalism level is high, on the other hand, if the program complexity level is high, then the programmer professionalism level is almost low. The results of the conducted experiments show that the proposed approach give very high and accurate evaluation for the undertaken systems.

Keywords—Complexity; java code; McCabe; Halstead; hybrid technique

I. INTRODUCTION

Java language is considered as one of the languages that has various advantages, these advantages includes its simplicity, safety, strength, impact, high level object-oriented ability, and many other advantages [1]. Complexity can here be defined as, the relationship between the internal parts of the program and how these parts can be interacted with each other, some of these parts will be connected to other parts of the program to make the program more complex and difficult to be analyzed or maintained. However, if these parts are less cohesive then the program will be less complex, in this case, the analysis would be easier to be analyzed and maintained [2]. The benefits of complexity measurement can be summarized as follows:

a) Complexity analysis of code can even be estimated from a design (whenever the design is easy and simple then the

code will be less complex, in contrast, if the design is more complex and unclear, then the program will be more complex).

b) The ability to distinguish between the simple and more complex program (allow the programmers to write a program in a way that has the following features: high quality, easy to understand, has few mistakes, easy to use and re-use, easy to maintain, easy to test, saves time and lower cost).

c) Good Complexity Measure provides continuous feedback (allowing us to follow the program continuously and to avoid most of the expected mistakes or problems).

TABLE I. LEVEL OF COMPLEXITY BY MCCABE MEASURE

Complexity	Risk Evaluation
1-10	A simple module without much risk
11-20	A more complex module with moderate risk
21-50	A complex module of high risk
51 or more	An untestable program of very high risk

Categorizing any source code complexity into good or bad will be helpful for code maintenance and evolution. Typically, the source code with good complexity is more maintainable, testable, understandable, and have less errors. On the other hand, any source code with bad complexity will be complex to be maintained, tested, understood by developers, and it will have a lot of errors.

Shrivastava [3] presented a measurement to provide a single ordinal number to be used to compare the program's complexity with other programs. This measurement used McCabe Complexity measures to analyze the system and find the complexity of the program, as follows:

$$CC = E - N + P$$

where

CC = Cyclomatic Complexity

E = Number of edges of the graph
N = Number of nodes of the graph
P = Number of connected components

The following is an example

```
public void ProcessPages()  
{  
    while(nextPage != true)  
    {  
        if((lineCount <= linesPerPage) && (status !=  
        Status.Cancelled) && (morePages == true))  
        }  
    }  
}
```

As shown in the above example, the routine is starting by adding 1 to the *while loop*, adding 1 to the *if statement*, and adding 1 to each && for a total calculated complexity of 5.

Davis and LeBlanc [4] studied a predictive value of various syntax-based problem complexity measures; they discussed McCabe and Halsted Complexity measures and analyzed the system to find the complexity of the program. Sheppard et al [5] compared three types of existing standard measures to find the complexity of the program, they used Halstead, McCabe's, and the length that measured by number of statements to analyze the system and find the complexity of the program.

Prabhu [6] applies McCabe's cyclomatic complexity and the Halstead metrics to evaluate the complexity of Simulink models. Prabhu notes that, the challenge of switching from programming languages to models is that, metrics have to be tailored and values obtained at the code or model level so that computed values are different. Olszewska [7] introduced new metrics specific to high-level design. They focus primarily on model counting, such as the average number of blocks per layer or the stability of the number of inputs/outputs across the model. Toularkis [8] distinguished between two classes of complexity measures which are: dynamic complexity measure and static complexity measure. Dynamic complexity to measure the amount of resources consumed during computation and static complexity to measure the size (e.g. program length) or structural complexity. Olabiyisi et al. [9] applied different software complexity metrics to searching algorithms, and the result showed that for both linear and binary search techniques, the languages do not differ significantly, therefore it is concluded that any of the programming languages is good to code linear and binary search algorithms.

Software complexity is different at the architecture level, where it is defined by how components communicate and are integrated, than at the code or behavior level, where it is defined by how components are implemented [10]. Delange et al. [11] demonstrate that maintaining low-complexity components and delivering high-quality models reduce maintenance activities and associated costs. Banker [12] estimates that software complexity itself can increase maintenance costs of commercial applications by 25% and increase the total lifecycle costs by 17%. Considering not only that safety-critical applications have stringent quality requirements but also that both the software and models of such applications must be maintained for decades, the real

costs could be higher than these estimates for critical applications.

There is substantial evidence that cyclomatic complexity is linearly correlated with product size [13]. Evidence shows that software complexity has increased significantly over time not only because of the increase in number of functions but also because of a paradigm shift in which more functions are realized using software rather than hardware [14]. The SEI's experience with high-reliability systems has been that a high-quality process leads to a low number of defects and reduces rather than increases cost [15], [16]. Nonetheless, actual industry practice and estimates of cost for high-reliability software vary widely [17]. Shull reports increases to development costs ranging from 50% to 1,000% due to more coding constraints and certification requirements (e.g., testing, validation) [18].

To the best knowledge, most of the previous modules used only one technique or one theory to measure the ratio of complexity of the programs. So the contribution of this paper is integrating two theories that are called Halstead and McCabe. In this way, the ratio of complexity will be more accurate, which helps programmers to make sure that their programs will be better and their work is more efficient. Whenever the ratio of complexity is more complex, then it will increase the mistakes and errors in the program, thus, the difficulty in maintenance and testing will be increased and the cost of the program will also be increased [19], [20]. For this reason, the ratio of complexity must accurately be measured to be more efficient, contains less errors, easy to test, easy to understand, easy to maintain and test, then this will decrease the cost of the program.

Microsoft visual studio 2010 with language (C#) are used for building the program which has been written to allow users to open any program written in Java, analyze the code, extract all Operators and Operands, all number of edges, number of nodes, and number of connected components, then finding the complexity measurement that allows to identify both the program and programmer levels is done.

This paper consists of five other sections organized as follows: Section 2 discusses McCabe and Halstead complexity measures, Section 3 includes the proposed approach, Section 4 contains the evaluation and discussion, Section 5 about the related work, and finally, Section 6 presents the conclusion and recommendations.

II. MCCABE AND HALSTEAD COMPLEXITY MEASURES

This paper focuses on McCabe [21] and Halsted Measures [22], here each mechanism will be discussed in more details.

A. McCabe Complexity

This theory is being used widely since it was issued; it depends on computing and controlling flow graph of the program, and measuring the number of linearly-independent paths [23]. Tables 1, 2 and 3 show an example about McCabe along with the complexity, $C = E - N + 2P$, where E is the number of edges, N is the number of nodes, and P is the number of connected components.

The following is an example of McCabe complexity Measure as shown in Fig. 1.

```
public static void sort(int x []) {
    for (inti=0; i < x.length-1; i++) {
        for (int j=i+1; j < x.length; j++) {
            if (x[i] > x[j]) {
                int save=x[i];
                x[i]=x[j]; x[j]=save
            }
        }
    }
}
```

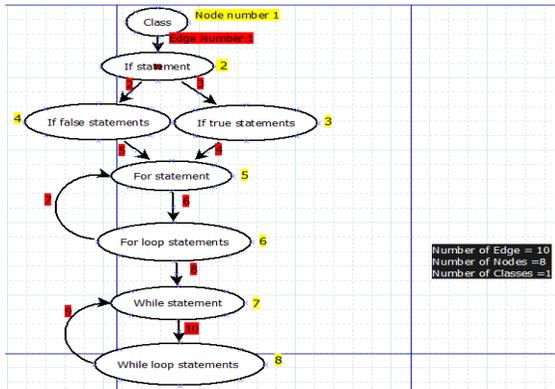


Fig. 1. Main steps for McCabe complexity.

TABLE II. MCCABE EXAMPLE

Measure	Symbol	Result
number of edges	\underline{E}	10
number of nodes	\underline{N}	8
number of connected components	\underline{P}	1
$C = E - N + 2P$ $C = 10 - 8 + (2 * 1) = 4$ 4 mean a simple module without much risk		

TABLE III. MCCABE EXAMPLE

Measure	Symbol	Result
number of edges	\underline{E}	13
number of nodes	\underline{N}	11
number of connected components	\underline{P}	1
$C = E - N + 2P$ $C = 13 - 11 + (2 * 1) = 4$ 4 mean a simple module without much risk		

The following is an example about the ratio of nested condition statements as shown in Table 4.

TABLE IV. NESTED CONDITION EXAMPLE

Over all condition statements	4
Nested condition statements	2
Ratio = Nested condition statements / Overall condition statements Ratio = 2/4	

B. Halstead Complexity

This theory is used to analyze and measure the complexity of the code; it relies on code division into two parts: Operators and Operands. In this way, the theory of Halstead that he believes can be interpreted as the followings: the program is a collection of operations performed on data, so in this case, each code in the program is either operation or operand. The following notations are used:

By using these parameters, Halsted theory can be defined as a set of complexity measures, including the program volume, program difficulty, program development time, and program bug fixing effort. Table 5 shows the symbol equation

n_1 = number of unique or distinct operators appearing in a program.

n_2 = number of unique or distinct operands.

$n = n_1 + n_2$, this is the vocabulary.

N_1 = total number of operators (implementation).

N_2 = total number of operands (implementation).

$N = N_1 + N_2$

for Halsted measure. Tables 6 and 7 show the operators and operand example, respectively.

The following is an example for Halsted complexity.

```
public static void sort(int x []) {
    for (inti=0; i < x.length-1; i++) {
        for (int j=i+1; j < x.length; j++) {
            if (x[i] > x[j]) {
                int save=x[i];
                x[i]=x[j]; x[j]=save
            }
        }
    }
}
```

TABLE V. SYMBOL EQUATION FOR HALSTED MEASURE

Measure	Symbol	Formula
Program length	N	$N = N_1 + N_2$
Program vocabulary	N	$n = n_1 + n_2$
Volume	V	$V = N * (\text{LOG } n)$
Difficulty	D	$D = (n/2) * (N_2/n_2)$
Effort	E	$E = D * V$

TABLE VI. OPERATOR EXAMPLE

Operator	Number of Occurrences
Public	1
Sort()	1
Int	4
[]	7
{}	4
for {;}	2
if ()	1
=	5
<	2
$n_1 = 17$	$N_1 = 39$

TABLE VII. OPERAND EXAMPLE

Operand	Number of Occurrences
X	9
Length	2
I	7
J	6
Save	2
0	1
1	2
$n_2 = 7$	$N_2 = 29$

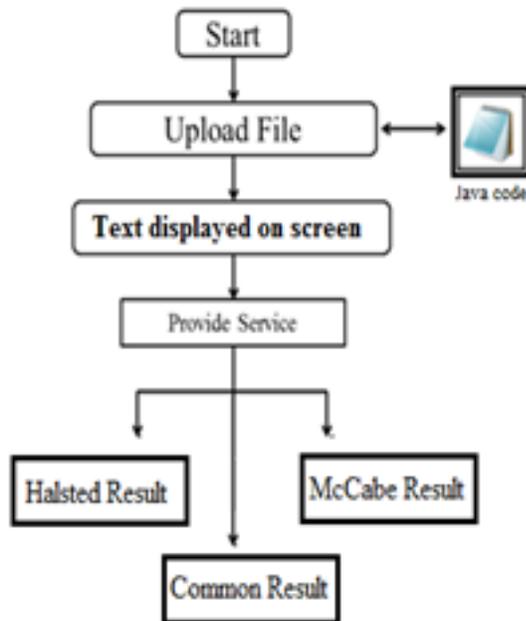


Fig. 2. Flow chart for the complexity analysis of JAVA code system.

III. PROPOSED APPROACH

The goal of this paper is to build a tool that measures the complexity of code to distinguish between the programs which have a little or more complexity, this can be made for the following reasons: to have a high-quality program, easy to be understood by the other programmers, has few mistakes, easy to use, easy to re-use, easy maintenance, easy to test, less of execution time, and lower cost.

Fig. 2 displays a flow chart for the complexity analysis of JAVA code system and working process. This system contains the main process of the first screen which uploads the file that contains Java code and then Enter, when the user start the code is displayed in the report, then the user selects what he/she needs. In this project there are 3 cases: Halsted Result, McCabe Result, and Common Result.

In order to create database for this program, all constants in the program must be selected, these constants such as all Java reserved words, and all Operators used in Java. Table 8 lists all words that are reserved, and Table 9 lists all Operators that are used.

TABLE VIII. JAVA RESERVED WORDS

abstract	Continue	For	new	switch
assert***	Default	goto*	package	synchronized
boolean	Do	If	private	this
Break	Double	implements	protected	throw
Byte	Else	import	public	throws
Case	enum****	Instance of	return	transient
Catch	Extends	int	short	try
Char	Final	interface	static	void
Class	Finally	long	strictfp**	volatile
const*	Float	native	super	while

TABLE IX. JAVA OPERATORS

Category	Operator	Name/Description
Arithmetic	+	Addition
	-	Subtraction
	*	Multiplication
	/	Division
	%	Modulus
	++	Increment
Logical	--	Decrement
	&&	Logical "and"
		Logical "or"
Comparison	!	Logical "not"
	==	Equal
	!=	Not equal
	<	Less than
	<=	Less than or equal
	>	Greater than
String	>=	Greater than or equal
	+	Concatenation(join two string)

IV. EVALUATION AND DISCUSSION

The main objective of this paper is to build a tool that measures the ratio of the complexity of JAVA programs. Typically, the best way to test the program is to have an example for it, in other words, a copy of the program must be available to have full evaluation for the program. This analysis is a dynamic based technique, where the program has been traced and inspected at running time. An example along with detailed steps about how the proposed approach works are presented and explained in this section.

Fig. 3 shows the Main window of the system which appears after clicking Enter in the Welcome window. It contains many buttons and empty space, these buttons such as: Browse of the project, Browse by Class, Clear Code area, McCabe Result, and Halsted Result. The main objectives of the buttons are as follows:

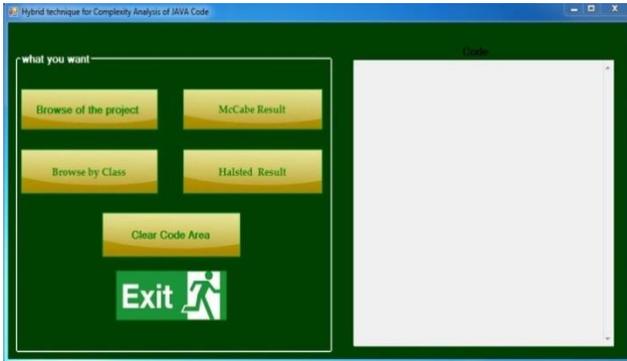


Fig. 3. Main window.

- **Browse of the project:** to open new screen in order to look for a folder containing some of classes written in Java,
- **Browse by Class:** to open new screen to look for any file containing some of codes written in Java.
- **Select code:** If you select a folder from (Browse of the project), this folder contains some of classes (message of number of file founded)
- **Clear Code area:** When the button is pressed, then any code in the code area is deleted.
- **McCabe Result:** the results screen is as shown in Fig. 4. It is designed for the following reasons: 1) Extract all number of edges, number of nodes, and number of connected components, 2) Make the necessary calculations, and 3) Find a level of complexity.



Fig. 4. Results screen.

In this window (Fig. 4), there are three main parts: Code Statement Analysis, Overall Code Analysis, and Final Result.

1) Code Statement Analysis: to extract number of comments, number of conditional statements, and number of loop statements in the project.

2) Overall Code Analysis: to extract number of edges, number of nodes, and number of connected components in the project.

3) Final Result: to calculate the complexity of the project using $C = E - N + 2P$ equation, then find the level of complexity using Table 1.

- **Halsted Result:** is designed to extract all Operators and Operands, make the necessary calculations, and find a level of complexity.

In this window (Fig. 5) there are three main parts: Operators, Operands, and result of Halted equation

1) Operators: to extract Operators with total number of each one.

2) Operands: to extract Operands with total number of each one.

3) Result of Halted equation: to calculate the Complexity of the project.

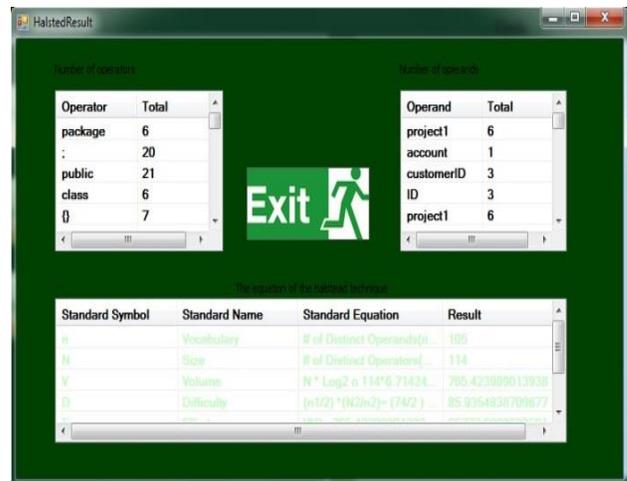


Fig. 5. Halsted result window.

- **Exit:** In all previous windows, click on the exit button to close the window or close the program.

Typically, code complexity correlates with the defect rate and robustness of the application program. In practice, the process of calculating the time complexity of a large program would be unproductive. Therefore, the developers must just focus on understanding the time complexity of the main functions of the program. Since that the time complexity of any program is considered as strong evidence and analysis for the complexity.

As shown in the above result, the output of the tool gives a detailed data about the undertaken code. Thus, by comprehending this data, the developers can know the exact complexity. This complexity can be used by the developers for any update or maintenance over the code specially when performing refactoring [24], [25]. The refactoring process over any source code is considered as a challenge for the developers, where the developers need to know previously the exact complexity information for the code. By presenting this information, the refactoring process will be easier and safer.

Moreover, the presented tools give a very accurate categorization for complexity risk. Furthermore, the presented approach helps the developers to find coding errors and programmers mistake if it exists. The presented approach was also evaluated by 10 master students, the student tried and evaluated the tool over two four open source software which are OpenCms which is website content management, Gwen view which is for 3D Modeling, K-3D which is for image viewer, and OLAT which is for Online Learning and Training. For each system, the students tried ten different test cases that mainly contains nested if statement and loops with all operators. The results show that having an analysis for Java programs using McCabe and Halstead theories together is very helpful for the developer. Moreover, the results can be used efficiently as a guide for software refactoring process, predicting effort, rate of error and time, and in scheduling projects.

V. RELATED WORK

A survey about software testing was presented by [26] which describes and presents the current approaches for software testing; the paper also presents an overview about the used models in software analysis and testing.

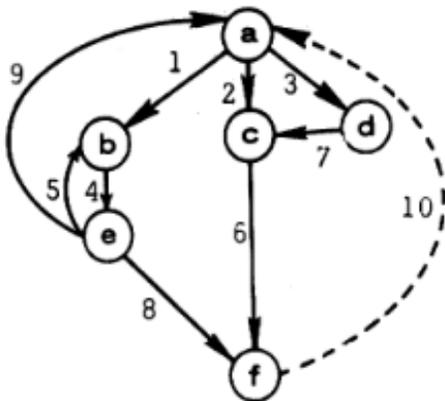


Fig. 6. McCABE example.

In 1976, Thomas McCABE used graph-theory to explain programming complexity [27], this method made it easier to trace the code paths within the program using algebraic expression to solve the infinite backward loops as shown in Fig. 6 as an example for a control graph.

On the other hand, Halstead [28] in 1977 has defined the way that metrics should affect the software implementation or expression despite of the type of the language that the developers have been used, but at the same time it won't affect the platform that has been used on the code execution time. The main idea was to find out a relation between all measurable properties for the software, this will measure the easiness of understanding the software code.

The complexity of coding issues has been raised especially with the appearance of object-oriented programming languages, Java was and still one of the most object-oriented languages that is used especially with the arise of mobile programming, mobile and other Java dependencies like Linux and Unix repositories needs away to find out the reachability

issue for a dead repositories code [29] to reduce the time missed in seeking a dead source.

VI. CONCLUSION

Complexity measures can be used to predict critical information about testability, reliability and maintainability of the software systems from automatic analysis of the source code. There are many code complexity measurements as Lines of Code metrics, McCabe, Halstead Metrics, Maintainability Index, and other code complexity measurements. In this paper, a tool has been developed to analyze the complexity of JAVA code using two complexity measures, Halsted and McCabe. The Halstead and McCabe theories has been explained, and the way which used to analyze code and find the complexity rate. The results show that the presented approach gave very useful and understandable results that can be used for developers assisting.

It has been concluded that this issue is very helpful to distinguish between the program which has a complexity ratio if it is high or not, because if there was less complexity ratio then the program is in its best case, easier to be understood, and easier to re-use and maintenance. Moreover, the focus in the paper has been made on analyzing codes written in Java, however in the future work there is a decision to expand this project to be negotiable on programs written in other languages such as C++, C# and/or any other languages.

In this paper, McCabe and Halstead theories have been only used, there is a hope to extend the program and add other metrics in the future work such as Zage metrics, McClure etc. This program is widely used to help the instructor to check the code, at the university for example, and compare codes written by programmers or students at the university or company. This program can be used by any person using Java to check his work quality and performance. The plan is to make the proposed technique useful for predicting the complexity of the program while designing phase by adding new features and statistical data.

REFERENCES

- [1] Arnold, K., Gosling, J., Holmes, D., & Holmes, D. The Java programming language (Volume 2). Reading: Addison-wesley. 2000.
- [2] Sanchez, S. M., & Lucas, T. W. Exploring the world of agent-based simulations: simple models, complex analyses: exploring the world of agent-based simulations: simple models, complex analyses. In Proceedings of the 34th conference on Winter simulation: exploring new frontiers. 2002. Pages 116-126.
- [3] Shrivastava, S. V., & Shrivastava, V. Impact of metrics based refactoring on the software quality: A case study. In TENCON 2008-2008 IEEE Region 10 Conference. 2008. Pages 1-6.
- [4] Davis, J.S., & LeBlanc, R.J. A Study of the Applicability of Complexity Measures. IEEE Transactions on Software Engineering, Volume 14. Number 9. 1988.
- [5] Sheppard, S. B., Curtis, B., Milliman, P., Borst, M. A., & Love, T. First-year results from a research program on human factors in software engineering. In afips (p. 1021). IEEE. 1899
- [6] Prabhu, Jeevan. Complexity Analysis of Simulink Models to Improve the Quality of Outsourcing in an Automotive Company. Manipal University. 2010.
- [7] Olszewska, Marga. Simulink-Specific Design Quality Metrics. TUCS Technical Report 1002. Turku Centre for Computer Science. 2011.
- [8] Tourlakis G. J. Computability, Reston, Virginia. Volume 12. 1984. Pages 39-42.

- [9] Olabiyisi S.O, Omidiora E. O and Sotonwa K. A. Comparative Analysis of Software Complexity of Searching Algorithms Using Code Based Metrics. *International Journal of Scientific & Engineering Research*. Volume 4. Number 6. 2013.
- [10] Aiguier, Marc et al. Complex Software Systems: Formalization and Applications. *International Journal on Advances in Software*. Volume 2. Number 1. 2009. Pages 47–62.
- [11] Delange, J., Hudak, J., Nichols, W., McHale, J., Nam, M. Y., (2015) Evaluating and Mitigating the Impact of Complexity in Software Models. CMU/SEI-2015-TR-013, Software Engineering Institute Carnegie Mellon University.
- [12] Banker, R. D. et al. A Model to Evaluate Variables Impacting the Productivity of Software Maintenance Projects. *Management Science*. Volume 37. Number 1. 1991. Pages 1–18.
- [13] Jay, Graylin et al. Cyclomatic Complexity and Lines of Code: Empirical Evidence of a Stable Linear Relationship. *Journal of Software Engineering and Applications*. Volume 2. Number 3. 2009. Pages 137–143.
- [14] Nolte, Thomas. Hierarchical Scheduling of Complex Embedded Real-Time Systems. *Ecole d'Été Temps-Réel 2009 (ERT09)*. Paris, France. September 2009.
- [15] Nichols, William R. Plan for Success, Model the Cost of Quality. *Software Quality Professional*. Volume 14. Number 2. 2012. Pages 4–11.
- [16] Obradovic, Alex. Using TSP to Develop and Maintain Mission Critical IT Systems. *TSP Symposium*. 2013.
- [17] Banker, Rajiv D. et al. Software Complexity and Maintenance Costs. *Communications of the ACM*. Volume 36. Number 11. 1993. Pages 81–94.
- [18] Shull, Forrest et al. What We Have Learned About Fighting Defects. *Proceedings of the 8th IEEE Symposium*. 2002. Pages 249–258.
- [19] Zage, W. M. & Zage, D. M. Evaluating Design Metrics on Large-Scale Software. *IEEE Software*. Volume 10. Number 4. July 1993. Pages 75–81.
- [20] Nam, Min-Young. ERACES: Complexity Metrics Tool User Guide. 2015.
- [21] McCabe T.J. A Complexity Measure. *IEEE Transactions on Software Engineering*. Volume 2. Number 4. 1976. Pages 308-320.
- [22] Halstead M.H. *Elements of software science*: Published by North Holland Amsterdam and N.Y. 1977.
- [23] Harrison, W. A., & Magel, K. I. A complexity measure based on nesting level. *ACM Sigplan Notices*, Volume 16. Number 3. 1981. Pages 63-74.
- [24] Meqdadi, O, Alhindawi, N, Collard, ML, Maletic, JI, "Towards understanding large-scale adaptive changes from version histories" in *International Conference on Software Maintenance (ICSM)*, 2013 29th IEEE.
- [25] Alhindawi, N, Alsakran, J, Rodan, A, and Faris, H, "A Survey of Concepts Location Enhancement for Program Comprehension and Maintenance" in: *Journal of Software Engineering and Applications*, 7:5 (2014), pp. 413–421.
- [26] Lee, J, Kang, S, and Lee, D "Survey on software testing practices," in *IET Software*, vol. 6, no. 3, pp. 275-282, June 2012.
- [27] McCABE, T. J. (n.d.). A Complexity Measure - IEEE Xplore Document. Retrieved August 26, 2017.
- [28] Hamer, P. G. (1992). Advertisements. *Environmental Science & Technology*, 26(12).
- [29] Buchsbaum, A, Yih-Farn Chen, Huale Huang, E. Koutsofios, J. Mocenigo, A. Rogers, M. Jankowsky, S. Mancoridis, "Visualizing and analyzing software infrastructures", *Software IEEE*, vol. 18, pp. 62-70, 2001.

An Unsupervised Local Outlier Detection Method for Wireless Sensor Networks

Tianyu Zhang, Qian Zhao, Yoshihiro Shin and Yukikazu Nakamoto

*Graduate School of Applied Informatics University of Hyogo

Computational Science Center Building 5-7F

7-1-28 Minatojima-minamimachi Chuo-ku Kobe Hyogo Japan

Abstract—Recently, wireless sensor networks (WSNs) have provided many applications, which need precise sensing data analysis, in many areas. However, sensing datasets contain outliers sometimes. Although outliers rarely occur, they seriously reduce the precision of the sensing data analysis. In the past few years, many researches focused on outlier detection. However, many of them ignored one factor that WSNs are usually deployed in a dynamic environment that changes with time. Thus, we propose a new method, which is an unsupervised learning method based on mean-shift algorithm, for outlier detection that can be used in a dynamic environment for WSNs. To make our method adapt to a dynamic environment, we define two new distances for outlier detection. Moreover, the simulation shows that our method performed on real sensing dataset has ideal results; it finds outliers with a low false positive rate and has a high recall. For generality, we also test our method on different synthetic datasets.

Keywords—Wireless sensor networks; outliers detection; unsupervised learning; mean-shift algorithm

I. INTRODUCTION

In recent decades, wireless sensor networks (WSNs) have been widely used in various applications to improve people's lives including securing their properties and ensuring their safety. For example, sensor nodes are used in smart houses and other buildings to monitor and regulate the living environments to provide better living comfort and save energy. Sensor nodes are also deployed in vehicle systems to provide data required by system control. However, in such applications of WSNs, the sensing dataset may contain outliers due to, for example, low quality sensor nodes, damage to nodes caused by harsh environments, or malicious attacks from outside. The outliers make the analysis of sensing dataset imprecise, which affects the WSN performance and can even cause serious mistakes that lead to disasters. Therefore, outlier detection methods are very important to guarantee the effectiveness of applications provided by WSNs.

There are many researches [1], [2], [3], [4] about how to automatically detect outliers that need a previously collected sensing dataset. For example, statistic-based methods use a previously collected dataset to estimate a model that is an approximation for the underlying distribution that generates the dataset. After that, they detect outliers with the estimated model. However, the estimated model may become invalid when the environment changes because the underlying distribution changes with the environment as well. Supervised learning based methods have a similar weak point. They need training data where every data point in the dataset is previously

labeled as normal or outlier to estimate a model. Similarly, the labels in training data may also become invalid when the environment changes. Moreover, preparing the training data is very time-consuming and expensive. Therefore, a method that can endure environment changes and automatically pick out outliers is needed in WSNs. In contrast, unsupervised learning based methods use raw data, which does not need to estimate a model from previously collected sensing dataset or prepare training data previously. Hence, unsupervised learning is more adaptable and convenient to WSNs. As a result of this property, we propose an unsupervised learning based method for detecting outliers.

Simply speaking, our outlier detection method first clusters the collected dataset and then uses the clustering result to detect outliers. We use the mean-shift algorithm to cluster the dataset because it can not only cluster the dataset but also find the mode (the mode is the most frequently occurring data point in a dataset) of each cluster as well. Then, the mode of each cluster and the median value of the sensing dataset can be used to detect which clusters are outliers. Moreover, to the best of our knowledge, this work is the first one to use the mean-shift algorithm to detect outliers in WSNs. Moreover, we simulated our method on the real sensor dataset of Intel Berkeley Research Laboratory and some synthetic datasets. We also compared our method with other unsupervised outlier detection methods [5], [6]. Simulation results shows that our method has a low FPR compared with related works, which indicates that our method outperforms than the related works in outlier detection.

The remainder of this paper is organized as follows. In Section 2, we present related researches and classify these researches into two classes that are model based and non-model based methods. Section 3 introduces preliminary knowledge related to our proposed method and the mean-shift algorithm used in our method. In Section 4, we presents the detail of our method. We test our method on real sensing dataset and synthetic dataset, and the results are shown in Section 5. Finally, Section 6 concludes this paper and provides a look at future work of our research.

II. RELATED WORK

There are many surveys about outliers and abnormal detection, such as Y. Zhang et al. [7], Pimentel et al. [8], Chandola et al. [9], Xie et al. [10] and Gupta et al. [11]. In these reviews, outlier detection methods are all based on statistic or machine learning methods. Some of the statistic and machine learning based methods are similar. For example, parametric-based

methods in statistic-based methods are similar to supervised learning in machine learning based methods because both of them estimate a model from a previously collected dataset. Non-parametric-based methods in statistic-based methods are similar to unsupervised learning in machine learning based methods, in that they do not need to estimate a model. Hence, we classify a number of related works into model-based methods and non-model based methods.

A. Model Based Methods

As we described above, model based methods focus, for instance, on estimating a probability model and assume the model generates measured data points. If a data point has a low probability by the estimated model, the data point is considered to be an outlier.

The following three methods are based on statistics to estimate a model. Wu et al. [12] presented a localized algorithm to identify outlying sensors and event in sensor networks. They utilize the spatial relationship of neighbor sensor nodes' readings to detect outlying sensors and event. Bettencour et al. [13] proposed a local outlier detection method to detect outliers in WSNs. They also use the spatio-temporal correlation of measurements between a sensor and its neighbors to build a model. Palpanas et al. [14] proposed using kernel density estimators to estimate a sensing dataset model on the basis of the distance for online deviation detection in streaming data. This is the supervised learning based method that Rajasegarar et al. [15] used, and they presented a method for anomaly detection in WSNs based on a one-class quarter-sphere support vector machine (SVM). They use training data to fit a hyper-surface, which is used to detect outliers.

B. Non-model Based Methods

Non-model based methods do not estimate a model. They use the relationship between data points, such as the distance between data points, and the density of the dataset.

These two methods are statistical non-model based methods. Subramaniam et al. [16] enhanced the work of Palpanas et al. [14] by detecting outliers online by approximating sensing data in a sliding window and using a local metrics-based algorithm to detect outliers in datasets that are hard to distinguish by distance. Sheng et al. [17] proposed a non-parametric-based method based on histogram information to detect outliers in WSNs. The biggest contribution of their method is that it reduces the communication cost by utilizing histogram information.

These are unsupervised learning methods in machine learning. Zhang et al. [5] presented an online local outlier detection method based on an unsupervised centered quarter-sphere SVM for WSN environmental monitoring applications. Fawzy et al. [6] presented a clustering based outlier detection method for WSNs. Similarly, Kiss et al. [18] also presented a clustering based outlier detection method. Other unsupervised learning based techniques include K-means approaches [19] and PCA-based approaches [20].

III. PRELIMINARIES

In this section, we first introduce types of outliers and then introduce the related concepts and assumption in our proposed

method. Finally, we introduce the clustering algorithm that we used in our method: "mean-shift algorithm".

A. Types of Outliers

Outliers are usually categorized as "global outliers" and "local outliers" (Fig. 1). Global outliers significantly deviate from the rest of the data points [21]. They are the simplest type of outliers and can be easily removed with some filters, such as "anchor data", that will be used in our method. On the other hand, local outliers are data points whose pattern significantly deviates from the pattern of the local area, so additional information of neighbor data points is needed for detecting local outliers. Therefore, detecting local outliers is more difficult than detecting global outliers.

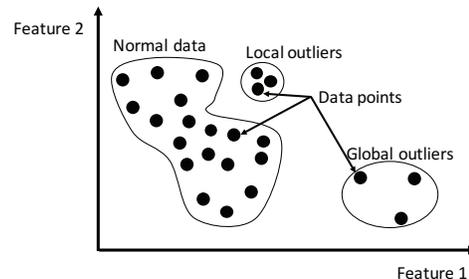


Fig. 1. Global outliers and local outliers.

B. Related Concepts and Assumption

There are three main indexes to show the center of a dataset: "mean value", "median value", and "mode".

The mean value is the average of the set of numbers, which can be easily calculated. However, it is easily affected by outliers because it becomes larger or smaller due to the effect of outliers.

The median value is the middle value in numerical order of a dataset. It is not observably affected by the outliers because if a dataset contains outliers, the median value is still decided by the majority of the non-outlier data points. Hence, most data points of a dataset are around the median value of the dataset.

The mode is a point that corresponds to the maximum probability density of a dataset. Hence, most data points are around the mode, which is similar to the median. However, calculating the mode of the dataset needs a lot of calculations. We can get an approximate value for the mode by using the median of the dataset.

In this paper, we assume that data points from a similar environment are generated by the same probability density function (PDF). Moreover, outliers are generated by other PDFs. The collected sensing dataset is mixed with normal data points and outliers. As stated above, the majority of data points should be around the center of the PDF. Moreover, the probability of outliers occurring is very low [22]. Hence, most of the data points in the dataset can be considered as normal data points, and they are around the center of the PDF. We choose the median value of the dataset to approximately represent the center of the PDF that generated the normal data points.

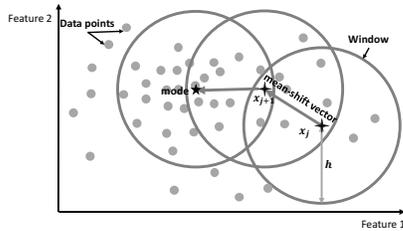


Fig. 2. Mean-shift migration from \mathbf{x}_j to mode.

C. Mean-Shift Algorithm

The mean-shift algorithm [23] is an unsupervised learning based cluster algorithm developed by Fukunaga and Hostetler [24] in 1975. It is an intuitive “mode” seeking method. Cheng et al. [25] showed that the mean-shift algorithm procedure is equivalent to the gradient ascent by kernel density estimation. The result of kernel density estimation is the mode.

First, we introduce the general idea of the mean-shift algorithm. Assuming that a dataset contains N data points in an M -dimension Euclidean space, each data point contains M features, such as $\mathbf{x}_i = (x_{i1}, \dots, x_{iM}), i = (1, \dots, N)$. We now explain the window, radius, mean-shift vector, and mode in the mean-shift algorithm.

The window is a subset of the dataset that has center \mathbf{x}_j and radius h (Fig. 2). It contains data points within a radius of h . The window notation in this paper is $win(\mathbf{x}_j, h)$. Every data point in a dataset can be considered as a center; hence, every data point can generate a window with radius h when initiating a mean-shift algorithm.

The radius h of a window is the only parameter of the mean-shift algorithm. The appropriate radius h is calculated by the standard deviation of the dataset [26]. Moreover, a stable dataset density is needed to get radius h to adapt to the dynamic environment. Hence, we introduce anchor data points (see Section 4(A) for details).

The mean-shift vector is calculated within a window. It decides the distance (length of mean-shift vector) and direction for moving the window from the previous center (\mathbf{x}_j) to the next center (\mathbf{x}_{j+1}). At the next center, the mean-shift repeats to make a new window and calculate the mean-shift vector of the new window. This process will terminate when the length of the mean-shift vector approaches zero. The mean-shift vector is calculated with the density gradient of the kernel density estimator according to Chengs study [25]. We show the derivations in the following subsection.

The mode is the center where a window finally stops moving. Data points swept by the movement of the window are contained in the same cluster because they have the same mode (center). Moreover, if some windows share the same mode (i.e. the modes are very close together), clusters generated by those windows are merged into one cluster. Fig. 2 shows the moving window procedure. The mode window is indicated by $win(\mathbf{c}_l, h)$, where \mathbf{c}_l is called the mode of cluster l .

1) *Kernel Density Estimator for Window*: By referring to Fig. 2, the total kernel density estimation of probability density at window $win(\mathbf{x}_j, h)$ [27] is

$$p(\mathbf{x}_j) = \frac{1}{n^{(j)}h^M} \sum_{i=1}^{n^{(j)}} K\left(\frac{\mathbf{x}_j - \mathbf{x}_i}{h}\right), \quad (1)$$

where, $n^{(j)}$ is the total number of data points in $win(\mathbf{x}_j, h)$.

$K(\bullet)$ is defined as the kernel function. In accordance with the radially symmetric mentioned by Cheng [25], we are only interested in kernel function $K(\mathbf{u})$ that satisfies

$$K(\mathbf{u}) = ck(\|\mathbf{u}\|^2), \quad (2)$$

where, $k(\|\mathbf{u}\|^2)$ is called *profile* of $K(\bullet)$. c is the positive normalization constant that assures kernel function $K(\mathbf{u})$ equals one. By utilizing the profile, we have

$$p(\mathbf{x}_j) = \frac{c}{n^{(j)}h^M} \sum_{i=1}^{n^{(j)}} k\left(\left\|\frac{\mathbf{x}_j - \mathbf{x}_i}{h}\right\|^2\right) \quad (3)$$

This is the kernel density estimator at $win(\mathbf{x}_j, h)$.

2) *Calculating Mean-shift Vector of Window by using Density Gradient*: To calculate the mean-shift vector of a window, we calculate the density gradient of $p(\mathbf{x}_j)$, and we set $g(s) = -k'(s)$.

$$\begin{aligned} \nabla p(\mathbf{x}_j) &= \frac{2c}{h^{M+2}} \sum_{i=1}^{n^{(j)}} (\mathbf{x}_i - \mathbf{x}_j) g\left(\left\|\frac{\mathbf{x}_j - \mathbf{x}_i}{h}\right\|^2\right) \\ &= \frac{2c}{h^{M+2}} \left[\sum_{i=1}^{n^{(j)}} g\left(\left\|\frac{\mathbf{x}_j - \mathbf{x}_i}{h}\right\|^2\right) \right] \times \\ &\quad \left[\frac{\sum_{i=1}^{n^{(j)}} \mathbf{x}_i g\left(\left\|\frac{\mathbf{x}_j - \mathbf{x}_i}{h}\right\|^2\right)}{\sum_{i=1}^{n^{(j)}} g\left(\left\|\frac{\mathbf{x}_j - \mathbf{x}_i}{h}\right\|^2\right)} - \mathbf{x}_j \right] \end{aligned} \quad (4)$$

The second term of (4) is mean-shift vector $\mathbf{m}(\mathbf{x}_j)$ in $win(\mathbf{x}_j, h)$.

$$\mathbf{m}(\mathbf{x}_j) = \frac{\sum_{i=1}^{n^{(j)}} g\left(\left\|\frac{\mathbf{x}_j - \mathbf{x}_i}{h}\right\|^2\right) \mathbf{x}_i}{\sum_{i=1}^{n^{(j)}} g\left(\left\|\frac{\mathbf{x}_j - \mathbf{x}_i}{h}\right\|^2\right)} - \mathbf{x}_j \quad (5)$$

The mean-shift vector always points in the direction of the increasing maximum density as shown in Fig. 2. Since \mathbf{x}_j and the mean-shift vector are known, the next candidate center point of a window is calculated as follows:

$$\begin{aligned} \mathbf{x}_{j+1} &= \mathbf{m}(\mathbf{x}_j) + \mathbf{x}_j \\ &= \frac{\sum_{i=1}^{n^{(j)}} g\left(\left\|\frac{\mathbf{x}_j - \mathbf{x}_i}{h}\right\|^2\right) \mathbf{x}_i}{\sum_{i=1}^{n^{(j)}} g\left(\left\|\frac{\mathbf{x}_j - \mathbf{x}_i}{h}\right\|^2\right)} \end{aligned} \quad (6)$$

Hence, the next window is $win(\mathbf{x}_{j+1}, h)$. Moreover, according to Cheng [25], no matter from which data point the calculation starts, the final result is convergent at the mode of probability density of the observed data.

IV. LOCAL OUTLIER DETECTION METHOD

In this section, we introduce our local outlier detection method. We assume that the WSN in our algorithm is a standard class¹-based WSN. In accordance with the similar environment, the sensor nodes and class head (CH) are distributed into different classes. Sensor nodes communicate with their CH, which transmits the gathered sensing data points to the base station.

Supposing a WSN contains P classes and one class has $W^{(p)}$, ($p \in [1, \dots, P]$) sensor nodes, each sensor node transmits G data points to CH in period t . Hence, each CH receives a set of data points, whose size is $N^{(p)} = W^{(p)} \times G$. One data point \mathbf{x}_i contains M features, $\mathbf{x}_i = (x_{i1}, \dots, x_{iM})$, $i = (1, \dots, N^{(p)})$.

The goal of the method is to cluster collected sensing data points of CH into different clusters and then find which cluster is an outlier in the sensing dataset. We add two main features to accompany the mean-shift algorithm: 1) anchor data points to fix the density of sensing dataset for each period to efficiently utilize the mean-shift algorithm; and 2) a labeling technique to classify the properties of cluster as “normal” or “outliers” in an unsupervised manner. The algorithm is divided into three steps.

A. Step 1: Fixing Density of Sensing Data and Detecting Global Outliers

We define the density of a collected sensing dataset at period t as follows:

$$dens^{(p)} = \frac{N^{(p)}}{\prod_{m=1}^M R_m^{(p)}}, \quad (7)$$

where, $R_m^{(p)}$ is the difference between the maximum and minimum value of the data points' feature m of class p at period t . The value range of feature m of the sensing dataset is different in different periods because the environment is different in different periods. Thus, the density changes along with the period.

Moreover, when the density is changing, it is not appropriate to use the mean-shift algorithm because mean-shift is sensitive to the density of a dataset, and variable density of the sensing dataset reduces the accuracy of the clustering result of the mean-shift algorithm. Furthermore, an incorrect clustering result will reduce the accuracy of outlier detection. To avoid the density changes in such a situation, we define the anchor data points, low anchor $L_m^{(p)}$, and high anchor $H_m^{(p)}$ for each feature m of class p . The low anchor $L_m^{(p)}$ is calculate by the minimum of normal feature m subtract δ_m and the high anchor $H_m^{(p)}$ is calculated by the maximum of normal feature m plus δ_m . The normal range of feature m and the value of

δ_m is decided by users. Thus, a fixed density uses anchor data points as follows:

$$\hat{dens}^{(p)} = \frac{N^{(p)}}{\prod_{m=1}^M (H_m^{(p)} - L_m^{(p)})}, \quad (8)$$

These anchor data points can also remove global, e.g., if a data point is lower than $L_m^{(p)}$ or larger than $H_m^{(p)}$. For example, in an office, the normal temperature range is from 20 °C to 30 °C. We set two anchor data points to 15°C and 35°C. A measurement of 10°C would be a global outlier.

B. Step 2: Clustering with Mean-Shift Algorithm

The purpose of this step is to cluster the collected sensing data of class p at period t into different clusters by mean-shift algorithm. Moreover, we have to update radius h_t at every period to guarantee the accuracy of the clustering result. Algorithm 1 shows the procedure.

Algorithm 1: Mean-Shift based Clustering	
01	for sensing dataset at each t
02	calculating radius h_t at period t
03	for data point \mathbf{x}_i , $i \in (1, \dots, N^{(p)})$, execute the mean-shift algorithm by moving $win(\mathbf{x}_i, h_t)$ to $win(\mathbf{c}_l^{(p)}, h_t)$ data swept by $win(\mathbf{c}_l^{(p)}, h_t)$ is defined as cluster $C_l^{(p)}$
05	if some windows share the same $\mathbf{c}_l^{(p)}$,
06	merge the clusters generated by those windows

As explained in Section 3(B), the mean-shift algorithm can find the mode of a cluster. First, CH calculates radius h_t which is the standard deviation of all the data points in period t . Then, the mean-shift algorithm clusters the sensing dataset by moving $win(\mathbf{x}_i, h_t)$, $i \in (1, \dots, N^{(p)})$ to $win(\mathbf{c}_l^{(p)}, h_t)$, where l indicates the number of clusters. If window $win(\mathbf{x}_j, h_t)$ finally stops at $\mathbf{c}_l^{(p)}$, the data points that are swept by the window are considered as cluster $C_l^{(p)}$. Moreover, if the distance between some modes of clusters is very small, we consider that these clusters share the same mode and merge those clusters. The new mode of merged cluster is the average of mode of each cluster before merging.

C. Step 3: Local Outlier Labeling Technique

We define two distances with the mode of each cluster and the median value of the collected sensing dataset, respectively. WSNs use these two distances to detect outliers. The detail of the two distances and how to detect outliers are as follows.

We define a Euclidean distance of cluster l that is the average distance from the mode $\mathbf{c}_l^{(p)}$ of cluster l to every data point in the collected sensing dataset of class p . We write this Euclidean distance as

$$Dis_l^{(p)} = \frac{\sum_{i=1}^{N^{(p)}} \|\mathbf{x}_i^{(p)} - \mathbf{c}_l^{(p)}\|}{N^{(p)}} \quad (9)$$

$\mathbf{M}_t^{(p)}$ is the median value of the collected sensing dataset of class p at period t . We define another Euclidean distance that is the average distance from $\mathbf{M}_t^{(p)}$ to every data point in the collected sensing dataset of class p . We write it as

¹In WSNs, a group of sensor nodes is called a ‘cluster’. In this paper, we call it a ‘class’ to distinguish it from ‘cluster’ in the mean-shift algorithm.

$$DIS^{(p)} = \frac{\sum_{i=1}^{N^{(p)}} \left\| \mathbf{x}_i^{(p)} - \mathbf{M}_t^{(p)} \right\|}{N^{(p)}} \quad (10)$$

We also find that $Dis_l^{(p)}$ is always larger or equal to $DIS^{(p)}$. The proof is as follows. The sensing dataset contains two parts. $\mathbf{x}_i : i = 1, \dots, N$ is the normal part of the dataset, and $\mathbf{y}_j : j = 1, \dots, n$ is the outlier part of the dataset. \mathbf{M}_t is the median value of the dataset, and $N \gg n$. For the normal part, $\hat{\rho} = E(|\mathbf{x}_i - \mathbf{M}_t|)$ is the average deviation of the normal data points, and $\rho = \max\{|\mathbf{x}_i - \mathbf{M}_t|\}$. For the outlier part, $\hat{R} = E(|\mathbf{y}_j - \mathbf{M}_t|)$ is the average deviation of outliers, and $R = \min\{|\mathbf{y}_j - \mathbf{M}_t|\}$. $\mathbf{c}^{(l)}$ is the mode of cluster l , and the distance from every data point to $\mathbf{c}^{(l)}$ is:

$$\begin{aligned} Dis_l^{(p)} &= \sum_{i=1}^N |\mathbf{x}_i - \mathbf{c}^{(l)}| + \sum_{j=1}^n |\mathbf{y}_j - \mathbf{c}^{(l)}| \\ &\geq \sum_{i=1}^N \left(|\mathbf{c}^{(l)} - \mathbf{M}_t| - |\mathbf{x}_i - \mathbf{M}_t| \right) \\ &\geq N(R - \hat{\rho}) \end{aligned} \quad (11)$$

On the other hand, the distance from every data point to \mathbf{M}_t is:

$$\begin{aligned} DIS^{(p)} &= \sum_{i=1}^N |\mathbf{x}_i - \mathbf{M}_t| + \sum_{j=1}^n |\mathbf{y}_j - \mathbf{M}_t| \\ &= N\hat{\rho} + n\hat{R} \end{aligned} \quad (12)$$

Then, the difference between $Dis_l^{(p)}$ and $DIS^{(p)}$ satisfies:

$$Dis_l^{(p)} - DIS^{(p)} \geq N(R - 2\hat{\rho}) - n\hat{R} \quad (13)$$

We suppose $N(R - 2\hat{\rho}) - n\hat{R} \geq 0$, then:

$$\frac{R - 2\hat{\rho}}{\hat{R}} \geq \frac{n}{N} \quad (14)$$

Since $R \gg \hat{\rho}$ and $N \gg n$, then $\frac{R}{\hat{R}} - 2\frac{\hat{\rho}}{\hat{R}} \gg 0$ and $\frac{R}{\hat{R}} - 2\frac{\hat{\rho}}{\hat{R}} \geq \frac{n}{N}$. Thus, our assumption that $\frac{R - 2\hat{\rho}}{\hat{R}} \geq \frac{n}{N}$ is true. We get $Dis_l^{(p)} \geq DIS^{(p)}$.

According to our assumption that data from a similar environment is generated by the same PDF, the sensing data of every sensor node in the same class has the same PDF because sensor nodes in similar environments are classified into the same class. Hence, the center of every cluster (the mode of each cluster) is similar to the center of the entire sensing dataset (the median value of the entire dataset) of the class. Thus, if cluster l is normal, $Dis_l^{(p)}$ should be close to $DIS^{(p)}$. In other words, the ratio of $Dis_l^{(p)}$ to $DIS^{(p)}$ should be close to 1. Moreover, because $Dis_l^{(p)} \geq DIS^{(p)}$, we set threshold ϵ , which is a very small empirical value, and use discrimination $\frac{Dis_l^{(p)}}{DIS^{(p)}} - 1 \leq \epsilon$ to detect outliers. The algorithm for detecting outliers is as follows.

Algorithm: Outlier detection of cluster	
01	for each cluster $Dis_l^{(p)}$
02	if $\frac{Dis_l^{(p)}}{DIS^{(p)}} - 1 \leq \epsilon$
03	cluster l is labeled as normal
04	else
05	cluster l is labeled as outlier

V. SIMULATIONS

In this section, we show our simulation results based on a real dataset from the Intel Berkeley Research Laboratory and a synthetic dataset. We also compare our simulation results with those of Z. Yang et al. [5] and A. Fawzy et al. [6]. Both of them detected outliers on the basis of unsupervised method, since they used the same real dataset as we did, we compare our method with theirs by using the synthetic dataset generated in the same way for the sake of testing the generality of our method. Moreover, we compare simulation results with and without setting the anchor data since this is an important characteristic of our method.

A. Simulation Results of Real Dataset

In this subsection, we simulate our method on the real dataset from Intel Berkeley Research Laboratory² as shown in Fig. 3. Each sensor node in the WSN records temperature, humidity, light, and voltage once every 31 seconds. We choose the sensor nodes 1, 2, 33, 34, 35, 36, and 37 inside the circle (35 is the CH), and we use two features, the temperature and humidity of 5th March 2004. The normal data ranges and the averages of temperature and humidity are shown in Table I. According to the settings of Table I, we set four types of outliers and anchor data for the real dataset, which are shown in Tables II and III. The four types of outlier cover the cases that outliers are close to or far away from the normal data range, and they are generated by different uniform distributions. For instance, the temperature values of the *outlier1* are generated by uniform distribution in interval [26 ~ 30]. Moreover, we randomly insert outliers into the dataset to respectively generate datasets containing 5%, 10%, 15%, 20%, and 25% outliers for every type of outliers. The anchor data points are set by the rule that the minimum value of normal feature m subtract δ_m and the maximum value of normal feature m plus δ_m , where the δ_m is set to 6 units of a feature, such as 6°C.

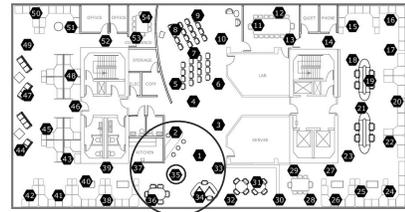


Fig. 3. Sensor nodes deployed in Intel Berkeley Research Laboratory.

- Outlier1 is near the normal data, some outliers are even inside the normal range.

²The dataset can be downloaded from <http://db.csail.mit.edu/labdata/labdata.html>, 2016

TABLE I. NORMAL DATA SETTING

	Range	Average
Temperature ($^{\circ}C$)	21.32~28.14	23.14
Humidity (%)	26.39~44.02	37.69

TABLE II. OUTLIER DATA SETTING

Type of Outlier	Outlier1	Outlier2	Outlier3	Outlier4
Temperature ($^{\circ}C$)	26~30	31~35	22~28	31~35
Humidity (%)	42~46	47~52	47~52	27~44

- Outlier2 is far from the normal data; however, they cannot be removed by anchor data.
- Outlier3 is such that the value of temperature is normal; however, the value of the humidity is abnormal.
- Outlier4 is the opposite setting of Outlier3.

The following terms are used to access our method:

- True Positives (TPs) are true outliers that were detected as outliers by our method.
- False Positives (FPs) are true normal samples that are wrongly detected as outliers.
- True Negatives (TNs) are true normal samples that were detected as outliers.
- False Negatives (FNs) are true outliers that are detected as normal samples.

The false positive rate (FPR) is the ratio of the normal data detected as outliers to the total true normal data, which is $\frac{FP}{FP+TN}$, and it estimates the ability of the algorithm to distinguish outliers and normal data. The FPR of our method is shown in Fig. 4. Moreover, we compare the FPR with Yang's method [5] and Fawzy's method [6]; the result is shown in Table IV. It shows that the performance of our method is acceptable, because the FPR of each case is relatively low comparing with other two related works in Table IV.

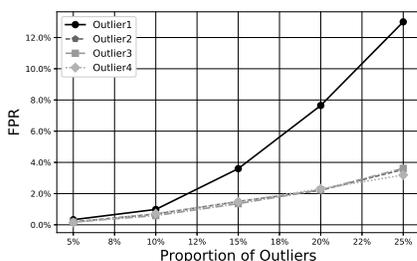


Fig. 4. Simulation results using real dataset of Intel Berkeley Research Laboratory.

In Fig. 4, outlier2 and outlier3 have similar curves so that outlier2 is blocked by outlier3. The FPR of outlier2, outlier3, and outlier4 kept below 3.3% when the outliers' percentage was less than or equal to 20%. Even in extreme conditions where a dataset contains 25% outliers, the worst case (outliers1) in our simulation has an FPR of about 12.8%.

Moreover, outlier2, outlier3, and outlier4 have similar results. Outlier2 can easily be detected as outliers because

TABLE III. ANCHOR DATA SETTING

Type of Anchor Data	Low Anchor	High Anchor
Temperature ($^{\circ}C$)	15.32	34.14
Humidity (%)	20.39	50.02

TABLE IV. COMPARISON OF FPR (%) ON REAL DATASET

Proportion of outlier	5%	10%	15%	20%	25%
Our method	0.20	0.74	1.98	3.60	5.83
Yang's method	1.37	7.12	11.21	18.32	19.10
Fawzy's method	0.31	2.76	4.11	8.54	11.66

its temperature and humidity are both abnormal. Although features of outlier3 and outlier4 are partially normal, we can imagine that the distributions of outlier3 and outlier4 deviated from the normal range in two-dimension. The results of outlier2, outlier3, and outlier4 prove that our method can easily be adapted to different types of outliers.

Another fact (Fig. 4) is that more outliers significantly affect the FPR of our method. In outlier1, with the proportion of outliers increasing, more and more outliers appear in the normal range because some part of outlier1 overlaps the normal range. Hence, a lot of normal data points are easily detected as outliers. Similar results also appear in outlier2, outlier3, and outlier4 because with the proportion of outliers increasing, a great many outliers appear near to the normal range. The FPR of our method decreases when the proportion of outliers increases because normal data points are incorrectly detected as outliers. However, comparing with the other two related works according to Table IV, our method can correctly detect outlier when proportion of outliers increases.

Recall is equal to $\frac{TP}{FN+TP}$ and acts as one estimator that evaluates how many true outliers are correctly detected. The recall of our simulation is shown in Fig. 5.

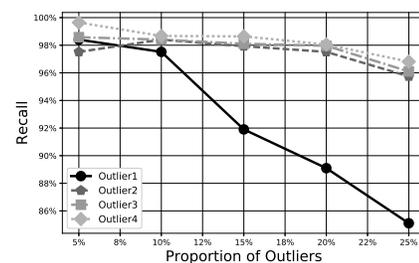


Fig. 5. Simulation results of recall.

This figure shows that all types of outliers have recall near 98% when the proportion of outliers is 5%. The recalls of outlier2, outlier3, and outlier4 are around 96% with increasing proportion of outliers. Even the worst case with outlier1 with 25% outliers, the recall is near 85%. The simulation results of every type show that our method can correctly detect outlier.

B. Simulation Results of Synthetic Datasets

Synthetic sensing data are generated by mixing three Gaussian distributions. The mean μ is randomly selected from

(0.3, 0.35, and 0.45), and the standard deviation is $\sigma = 0.03$. Outliers are generated by uniform distribution, which is distributed in an interval of $[0.5, 1]$. According to the empirical rules of Gaussian, the value range of Gaussian distributions is $\mu \pm 3\sigma$, and the normal range of the synthetic data is $[0.21, 0.54]$. The anchor data is $[0.11, 0.64]$ which is calculated by the normal range of synthetic data ± 0.1 . This synthetic dataset blends all the conditions we discussed in real data, which are outliers overlapping normal data, outliers near to normal data, and partial feature values are normal.

Fig. 6 is the result of FPR of our method. Because the synthetic data blends all types of outliers and the outliers were randomly generated, sometimes more outliers fall into or near the normal range. Thus, we can only control the quantity of outliers; however, we cannot decide where the outliers falls. This leads to the FPR of our method being higher than that of the real data, and this is the reason that the FPR is higher when the proportion of outliers is 15%.

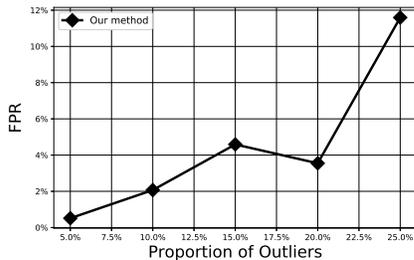


Fig. 6. Simulation results for FPR of proposed algorithm.

The comparison result between our method and Yang’s method and Fawzy’s method is shown in Table V. According to Table V, Yang’s method and Fawzy’s method tends to break down with the number of outlier increasing. Meanwhile, our method keeps a relatively low FPR, so that it can detect the outlier correctly.

TABLE V. COMPARISON OF FPR (%) ON SYNTHETIC DATASET

Method	5%	10%	15%	20%	25%
Our method	0.51	2.06	4.59	3.54	11.59
Yang’s method	2.41	8.51	13.53	19.61	25.01
Fawzy’s method	1.33	5.06	10.79	16.54	21.96

We also calculate the recall of our method performed on the synthetic data to confirm the effect of outliers, which is shown in Fig. 7. The result shows that the recall of our method fluctuates because the randomly generated outliers sometimes fall inside the normal range. When outliers fall inside the normal range, they significantly affect our results. However, the recall of synthetic data has a similar trend, which is decreasing with increasing outliers, with the recall of real data. Moreover, because the probability that outliers occur is low, a dataset that contains 25% outliers is an extreme case. Even in the extreme case, the recall keeps close to around 80% (Fig. 7). Hence, we conclude that our proposed method also has an acceptable performance in the more general cases.

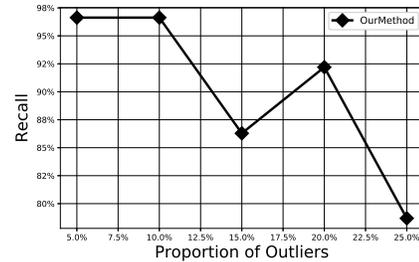


Fig. 7. Simulation results for recall on synthetic datasets.

C. Simulation Results Affected by Anchor Data

As mentioned in Section 4(A), the mean-shift algorithm may cluster the normal data into several clusters because the density of the dataset is changing with time, which leads to normal data being detected as outliers. Since using anchor data points is a feature of this work, to evaluate this aspect, we performed the following simulation where an outlier-free dataset is distributed in a 2-D Gaussian distribution. As shown in Fig. 8(a), two anchor data points were inserted at each point L and H (Fig. 8(b)). The simulation results in Fig. 8(a) show that, without setting anchor data points, the dataset were clustered into four classes, and two of them were determined as outliers. On the other hand, the simulation results in Fig. 8(b) show that, taking advantage of the anchor data points, the normal data were clustered as one class and were correctly determined as “normal.”

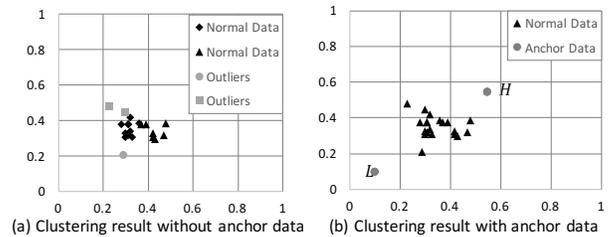


Fig. 8. Clustering results with and without anchor data.

VI. CONCLUSION

In this paper, we described the necessity for detecting outliers in WSNs and presented an unsupervised learning based outlier detection method to solve this problem. In our method, we first fixed the density of the dataset to utilize the mean-shift algorithm efficiently by using anchor data. Then, the mean-shift algorithm was used to cluster the collected sensing dataset into clusters. Finally, we proposed a labeling technique to label those clusters as “normal” or “outliers”; hence, outliers in the sensing dataset can be detected. In the simulations, we showed the performance of our proposed method and compared our work with related work [5], [6]. The results showed that our method has a lower FPR than that of the related work, and when outliers are far away from the normal data, our method obtained an FPR below 3.3%, which is quite low. Moreover, even in datasets where the distributions of outliers are close to the normal data or a substantial number of outliers are in the dataset, our method can still keep FPR at a low rate. The

simulations on synthetic dataset also showed the generality of our method.

From the QoS perspective of WSNs, to keep the WSN working properly, when outliers in the sensing data are discovered, approaches such as how to tolerate the outliers or how to detect outliers on the sensor node side should be considered. Therefore, part of our future work is methods for tolerating outliers and distributed outlier detection in sensor nodes. Moreover, our method can be used for event detection because outliers are an event in the dataset. Based on the current method, we want to improve it, for example, how to reduce computing and using less dataset, which can prolong the life of sensor nodes.

REFERENCES

- [1] A. De Paola, S. Gaglio, G. L. Re, F. Milazzo, and M. Ortolani, "Adaptive distributed outlier detection for wsns," *IEEE transactions on cybernetics*, vol. 45, no. 5, pp. 902–913, 2015.
- [2] E. W. Dereszynski and T. G. Dietterich, "Spatiotemporal models for data-anomaly detection in dynamic environmental monitoring campaigns," *ACM Transactions on Sensor Networks (TOSN)*, vol. 8, no. 1, p. 3, 2011.
- [3] S. Mascaro, A. E. Nicholso, and K. B. Korb, "Anomaly detection in vessel tracks using bayesian networks," *International Journal of Approximate Reasoning*, vol. 55, no. 1, pp. 84–98, 2014.
- [4] X. Li, J. Han, S. Kim, and H. Gonzalez, "Roam: Rule-and motif-based anomaly detection in massive moving object data sets," in *Proceedings of the 2007 SIAM International Conference on Data Mining*. SIAM, 2007, pp. 273–284.
- [5] Z. Yang, N. Meratnia, and P. Havinga, "An online outlier detection technique for wireless sensor networks using unsupervised quarter-sphere support vector machine," in *Proc. IEEE International Conference on Intelligent Sensors, Sensor Networks and Information Processing*. IEEE, 2008, pp. 151–156.
- [6] A. Fawzy, H. M. Mokhtar, and O. Hegazy, "Outliers detection and classification in wireless sensor networks," *Egyptian Informatics Journal*, vol. 14, no. 2, pp. 157–164, 2013.
- [7] Y. Zhang, N. Meratnia, and P. Havinga, "Outlier detection techniques for wireless sensor networks: A survey," *IEEE Communications Surveys & Tutorials*, vol. 12, no. 2, pp. 159–170, 2010.
- [8] M. A. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, "A review of novelty detection," *Signal Processing*, vol. 99, pp. 215–249, 2014.
- [9] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys*, vol. 41, no. 3, p. 15, 2009.
- [10] M. Xie, S. Han, B. Tian, and S. Parvin, "Anomaly detection in wireless sensor networks: A survey," *Journal of Network and Computer Applications*, vol. 34, no. 4, pp. 1302–1325, 2011.
- [11] M. Gupta, J. Gao, C. C. Aggarwal, and J. Han, "Outlier detection for temporal data: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 9, pp. 2250–2267, 2014.
- [12] W. Wu, X. Cheng, M. Ding, K. Xing, F. Liu, and P. Deng, "Localized outlying and boundary data detection in sensor networks," *IEEE transactions on knowledge and data engineering*, vol. 19, no. 8, pp. 1145–1157, 2007.
- [13] L. M. Bettencourt, A. A. Hagberg, and L. B. Larkey, "Separating the wheat from the chaff: Practical anomaly detection schemes in ecological applications of distributed sensor networks," in *International Conference on Distributed Computing in Sensor Systems*. Springer, 2007, pp. 223–239.
- [14] T. Palpanas, D. Papadopoulos, V. Kalogeraki, and D. Gunopulos, "Distributed deviation detection in sensor networks," *ACM SIGMOD Record*, vol. 32, no. 4, pp. 77–82, 2003.
- [15] S. Rajasegarar, C. Leckie, M. Palaniswami, and J. C. Bezdek, "Quarter sphere based distributed anomaly detection in wireless sensor networks," in *Proc. IEEE International Conference on Communications*, vol. 7, 2007, pp. 3864–3869.
- [16] S. Subramaniam, T. Palpanas, D. Papadopoulos, V. Kalogeraki, and D. Gunopulos, "Online outlier detection in sensor data using non-parametric models," in *Proc. International Conference on Very Large Data Bases*. VLDB Endowment, 2006, pp. 187–198.
- [17] B. Sheng, Q. Li, W. Mao, and W. Jin, "Outlier detection in sensor networks," in *Proceedings of the 8th ACM international symposium on Mobile ad hoc networking and computing*. ACM, 2007, pp. 219–228.
- [18] I. Kiss, B. Genge, and P. Haller, "A clustering-based approach to detect cyber attacks in process control systems," in *Industrial Informatics (INDIN), 2015 IEEE 13th International Conference on*. IEEE, 2015, pp. 142–148.
- [19] S. Rajasegarar, C. Leckie, M. Palaniswami, and J. C. Bezdek, "Distributed anomaly detection in wireless sensor networks," in *Proc. IEEE Singapore International Conference on Communication Systems*. IEEE, 2006, pp. 1–5.
- [20] M. A. Livani and M. Abadi, "Distributed pca-based anomaly detection in wireless sensor networks," in *Proc. IEEE International Conference for IEEE Internet Technology and Secured Transactions*. IEEE, 2010, pp. 1–8.
- [21] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.
- [22] K. Publishers, *Anomaly Detection with Data Mining*, Kyoritsu Publishers. Elsevier, 2009.
- [23] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 5, pp. 603–619, 2002.
- [24] K. Fukunaga and L. Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition," *IEEE Transactions on information theory*, vol. 21, no. 1, pp. 32–40, 1975.
- [25] Y. Cheng, "Mean shift, mode seeking, and clustering," *IEEE transactions on pattern analysis and machine intelligence*, vol. 17, no. 8, pp. 790–799, 1995.
- [26] D. Comaniciu, V. Ramesh, and P. Meer, "The variable bandwidth mean shift and data-driven scale selection," in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, vol. 1. IEEE, 2001, pp. 438–445.
- [27] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

Adaptive e-learning using Genetic Algorithm and Sentiments Analysis in a Big Data System

Youness MADANI*
and Jamaa BENGOURRAM

GI laboratory
Departement of Computer Sciences
Faculty of Sciences and Technics
Sultan Moulay Slimane University

*Corresponding author

Mohammed ERRITALI
and Badr HSSINA

TIAD laboratory
Departement of Computer Sciences
Faculty of Sciences and Technics
Sultan Moulay Slimane University

Marouane Birjali
LAROSERI Laboratory
Department of Computer Sciences
Faculty of Sciences, El Jadida
University of Chouaib Doukkali

Abstract—In this article we describe our adaptive e-learning system, which allows the learner to take courses adapted to his profile and to the pedagogical objectives set by the teacher, we use for adaptation the genetic algorithms to give the learner the concepts that must learn in an optimal way by seeking the objectives most adapted to his profile. And after a second level of adaptation using one of the social networks of the learner (twitter, facebook, Google + ...), based on his post on one of these social networks we propose two levels of analysis. The first one is to look for the period of activity which gives us an idea about the period when the learner is active and the second consists of making an analysis of the feelings on the publications that are published during the period of activity and related to education. Our work therefore is to adapt the profile of the learner with the pedagogical objectives by using the genetic algorithm and the notions of the research of information by doing this work in a Big Data system, that is to say we parallelize the search problem using Hadoop with Hadoop distributed file system (HDFS) and the MapReduce programming model, and after using information from a social network of the learner, we look for the period of activity of the learner and the feeling (sentiment analysis) related to the publications of the period of activity.

Keywords—Adaptive E-learning; genetic algorithms; research of information; social network; period of activity; sentiment analysis; parallelize the search problem; big data; Hadoop; MapReduce; Hadoop distributed file system (HDFS)

I. INTRODUCTION

The new information and communication technologies (ICT) profoundly improve our ways of informing, communicating and training us. This technological emergence has revealed a new mode of learning known as e-learning. It is based on access to online training, interactive and sometimes personalized, distributed via a network (Internet or Intranet) or another electronic medium. This access makes it possible to develop the skills of the learners while making the learning process independent of time and place.

The field of research in e-learning is very broad. It is also the object of a prosperous industrial activity and e-learning research issues Could be described as questions about adapting educational practices with today's technology [1].

Since the beginnings of e-learning, Artificial Intelligence (AI) techniques have been tested to increase the learning

experience. The use of AI algorithms in e-learning gives birth to adaptive e-learning.

Adaptive learning aims to propose a learning method that adapts to each learner's profile. This teaching emerged in the 1970 but is gaining momentum as technologies become more powerful and less costly. We are close to the research on artificial intelligence which has made enormous advances [1].

The goal of adaptive e-learning is to give the user of an e-learning platform a pedagogical content customized according to his profile with the use of algorithms like that of artificial intelligence. The idea is to find the pedagogical objective most adapted to the learner's profile because we can find a course with several pedagogical objectives suggested by a teacher.

We find several algorithms used in the literature in adaptive e-learning such as genetic algorithms that transform work into an optimization problem and give the learner the concepts that must learn.

Our work lies in this research axis, an adaptive e-learning system based on genetic algorithms and the notions of information retrieval like the similarity to find the relevant documents for a user request that expresses the user need. As a comparison between our work and information retrieval systems the learner profile will play the role of the query, and the different learning objectives will play the role of the sought documents, so it is as we want to calculate the similarity between the learner profile and the learning objectives to find those that are relevant to his profile and to keep only those objectives that are relevant that will then play the role of the initial population for the genetic algorithm in order to find an optimal objective, without forgetting that this work will be carried out in a parallel way by working with the Hadoop framework, storing the input and the output data of our algorithm in HDFS and using the Hadoop MapReduce programming model.

In addition to searching for an optimal pedagogical content for the learner's profile our work also consists in using one of the most powerful platforms in the world, it is the social networks, our proposal is to give the learner the possibility to connect to our e-learning platform using a social network like facebook, twitter, etc. The idea is to calculate a measure that we call it **period of activity** that will give us an idea about the period of the day when the learner is very active and therefore

will help us to improve the learner's skills, and after making a sentiment classification (motive, demotivate and neutral) of the publications of this period of activity that are related to the field of education to know the feeling of the learner that will help the teacher to define the level of learning.

The reminder paper is organized as follows: In Section 2 we give a brief overview of the Genetic Algorithm(GA) and Information Retrieval System(IRS). Section 3 gives some related work, and in Section 4 we explain how we used social networks in our work. A detailed description of our work is presented and detailed in Section 5, and in Sections 6 and 7 we describe/give some experimental results the parallelization's steps, the Section 8 summarize our work; finally, we give a conclusion and some future works in Section 9.

II. GENETIC ALGORITHM AND INFORMATION RETRIEVAL SYSTEM

A. Information Retrieval System

The search for information tries to solve the following problem: Given a very large collection of objects (mostly documents), find those that respond to a need for information expressed by a user (request). In the Information Retrieval System, we find a request and we want to find the objects (documents) that are relevant to it, the way to evaluate a document if it is relevant or not is to calculate the similarity between the request and that document.

Before the calculation of the similarity it is important to index all the documents and also the request that is to make them in a presentation to facilitate its use in our case we use the vector representation [2], where each element of the vector represents the weight (frequency) of each term or concepts in the document or in the query.

Our corpus in our case contains the documents that represent the learner's objectives, the first thing to do is to extract all the terms or concepts in the corpus, and for each document construct a vector That represents it, if a term exists in the document we calculate its weight and if not we put 0, at the end of this operation we construct a vector for each document to calculate the similarity between the profile of the learner and each pedagogical objective.

The calculation of the weights of terms or concepts in each document is calculated by the following formulas:

$$Poid(t_i, d_j) = TF * IDF \quad (1)$$

avec:

- $$TF = \frac{f(t_i, d_j)}{N} \quad (2)$$

$f(t_i, d_j)$ is the number of occurrences of the term t_i in the document d_j and N is the total number of terms in the document d_j .

- $$IDF = \frac{\log(f(t_i, d_j))}{M} \quad (3)$$

$f(t_i, d_j)$ Is the number of occurrences of the term t_i in the document d_j and M is the total number of documents in the corpus.

The similarity used in our work is the Cosine similarity. This measure uses the complete vector representation, i.e. the frequency of the objects (words). Two objects (documents) are similar if their vectors are confused, the formula is defined by the ratio of the scalar product of the vectors X and Y and the product of the norm of X and Y .

$$Sim_{cos}(X, Y) = \frac{\sum_{i=1}^n x \cdot y}{\sqrt{(\sum_{i=1}^n x^2)} \cdot \sqrt{(\sum_{i=1}^n y^2)}} \quad (4)$$

B. Genetic Algorithm

Genetic algorithms (GAs) are stochastic optimization algorithms based on the mechanisms of Natural selection and genetics, their operation is extremely simple, we leave with a population of potential solutions (chromosomes) initial selected arbitrarily, we evaluate their relative performance (fitness). On the basis of these performances, a new population of potential solutions is created using simple evolutionary operators such as selection, crossing and mutation. This cycle is repeated until a satisfactory solution is found [3].

In our work we use a simple GA, which consists of iterating the following three operations: reproduction, crossing and mutation, the population created during each iteration is called a generation and it's noted P_t .

There has been an increasing interest in the application of GA tools to IR in the last few years. Concretely, the machine learning paradigm, whose aim is the design of a system able to automatically acquire knowledge by themselves, seems to be interesting on this topic.

The first thing in a genetic algorithm is the definition of the initial population (selection operator or evaluation) on which we will apply the treatment as in our case it is to show the documents (educational objectives) relevant to the profile of the learner using the cosine similarity that will play the role of fitness function which is a very important parameter in GA because with it we can decide whether an individual is going to be selected or not. There is a lot of methods to make the selection like the biased lottery, the elitist method or the selection by tournaments.

After applying the selection operator to the initial population, it is the reproduction step with the application of the crossing or crossover operation and the mutation operation.

In the literature we find many works that applies genetic algorithms in the search for information, as in [4] where authors use in their information retrieval system the genetic algorithm to find the relevant documents for a user query, they use the vector representation to present the documents of the search base and the query, they have made comparisons with precision measurements and recall of the system using different fitness functions like Cosine, Dice and Jaccard.

In [5] the researcher explored the problems embedded in this process, attempted to find solutions such as the way of choosing mutation probability and fitness function, and chose

Cranfield English Corpus test collection on mathematics. Such collection was conducted by Cyril Cleverdon and used at the University of Cranfield in 1960 containing 1400 documents, and 225 queries for simulation purposes. The researcher also used cosine similarity and jaccards to compute the similarity between the query and documents and used two proposed adaptive fitness function, mutation operators as well as an adaptive crossover. The process aimed at evaluating the effectiveness of results according to the measures of precision and recall.

Vajitoru [6] Also uses the Genetic Algorithms in the research of information, he proposed a new operation of crossing to improve the research with the genetic algorithm, for that he made a comparison between his proposal and a classic GA and the results show the Effectiveness of its proposal.

Sathya and Simon [7] use the genetic algorithms to improve an information retrieval system and make it effective for obtaining more pages relevant to the user's query and optimize the search time.

In [8] the Researchers present a new fitness function for approximate information retrieval which is very fast and very flexible than cosine similarity.

Fan et al. propose an algorithm for indexing function learning based on GA, whose aim is to obtain an indexing function for the key term weighting of a documentary collection to improve the IR process [9].

III. GENETIC ALGORITHM ON ADAPTIVE E-LEARNING

Several works of artificial intelligence are used in adaptive e-learning to give the learner a content adequate to his profile in the literature we find:

Hawkes and Derry [10] have used the informal fuzzy reasoning in the TAPS system to determine with uncertainty the solution that the student has built among those of the system (models).

Ruiz et al. [11] have modeled an adaptive hypermedia system, called Feijjo.net, based on the learning style. The system uses fuzzy logic to determine the learner's style from the CHAEA questionnaire.

Chrysafiadi and Virvou [12] have proposed a learner model that represents the learner's knowledge through the overlay model (presented concepts that the learner master with "1" or with the word "known" and those that do not master by "0" or unknown), the fuzzy logic allowed to define and update the level of knowledge of each concept, with each interaction with the e-learning system.

Martin and VanLehn [13] have presented OLAE as an assessment tool that collects data from students solving physics problems in college. For each problem, OLAE automatically creates a Bayesian network that calculates the probabilities indicating the rules that the student uses.

Vicari et al. [14] have introduced AMPLIA, an intelligent learning environment used as a training tool in the medical field, the system combines bayesian networks with cognitive.

There are also works that use genetic algorithms for adaptive e-learning, namely:

the work of Romero et al. [15] Which represent a methodology to improve education systems, using grammar based on genetic algorithm techniques and multi-objective optimization to extract prediction rules allowing teachers to select the most appropriate changes to improve the efficiency of the Training.

Chang and Ke [16] Have proposed a customized composition of courses in an adaptive learning system, based on the genetic algorithm (GA), with the aim of specifying the appropriate learning resources for each learner.

In [1] the Researchers describe an adaptive system conceived in order to generate pedagogical paths which are adapted to the learner profile and to the current formation pedagogical objective. They have studied the problem as an "Optimization Problem" using Genetic Algorithms, the system seeks an optimal path starting from the learner profile to the pedagogic objective passing by intermediate courses to prepare the courses for adaptation.

In [17] a genetic algorithm based adaptive learning scheme for context aware e-learning has been described, the Researchers defined a new three level structure for learner's context comprising of the content level, presentation level and media level is defined. The learning path generation algorithm now evolves into a learning scheme generation as it generates a learning path accommodating the entire learner's context.

IV. SOCIAL NETWORK

As we said earlier our adaptive e-learning system uses social networks to extract information about the learner for that we use social authentication.

Social authentication with a social network is an authentication type that allows us to use existing login information of a user to a social network such as Facebook, Twitter or Google+, to connect the user to a third website, instead of creating a new login account specifically for this site. Social login is simple and effective they allow users to authenticate to websites without having to create an additional account, Just a click on a social button authentication is Enough [18].

The authentication button increases the enrollment to a platform. Why? Just because the authentication button removes the need for the user to refill a form, choose a username and secure password. Thus, it now needs to do one click to move from one social network to another.

This module simplifies the registration of a new user to a site. Instead of filling the required fields for registration or login, you can simply click on the button corresponding to the social network, and that's all you are a registered user. With this module the number of registered users on your website increases as well as the potential activity of users.

We use the social authentication in our work to retrieve the learner's publications to analyze them afterwards either to look for the period of activity or to analyze the feelings of the publications that are in relation with education.

A. Period of Activity

The period of activity is a measure proposed in our work to look for the period of the day when the learner publishes a lot of publications in his social network account, we work in

our case with Facebook¹, so the period of activity is the period of the day when the learner publishes a lot in his Facebook account.

We define three periods of the day, from 8 am to noon, from 2 pm to 6 pm and from 6 pm to 10 pm. The period in which we find a large number of learner publications is the **Period of Activity (PA)**.

$$PA = \max P_i \quad (5)$$

Where,

- $i = 1, 2$ ou 3
- P_i is the number of publications in the period number i (1=from 8 am to noon, 2=from 2 pm to 6 pm and 3=from 6 pm to 10 pm)

B. Sentiment Analysis

The second thing to do after the calculation of the period of activity is to recover all the publications that are published in this period and after filtering the publications according to the field of education, the idea is to recover only the Publications that contain terms related to education such as school, education, learn, learning, teacher, teaching, university, faculty, etc.

After we classify the publications obtained according to three classes: motivate, demotivate and neutral, using the dictionary AFINN² and after, the feeling of the learner in the period of activity takes the value of the majority class of publications for example if we have 20 publications that are related to education in the period of activity, and 10 of them are of motivated class, 6 of demotivate class and 4 of neutral class, so the majority class here is the motivate class and then the learner's feeling takes on the value motivate.

V. DESCRIPTION FOR OUR WORK

Our work as mentioned previously consists in proposing a new e-learning system by combining the use of genetic algorithms, the notions of information retrieval systems, social networks, sentiment analysis and Big Data with the Hadoop framework.

In this section, we will present the different stages of our work, how we combine genetic algorithms with the notions of information retrieval and apply the results to give the learner a pedagogical content corresponding to his profile, and how we use the Facebook's publications of the learner for adaptation, this work will be made using the MapReduce programming model and the HDFS (Hadoop Distributed file system).

¹<http://www.Facebook.com>

²AFINN is a dictionary that contains words with weights between -5 and 5 which expresses the sentimental degree of the word

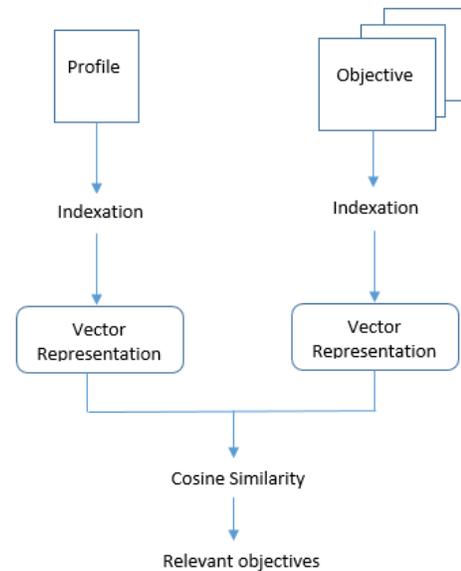


Fig. 1. Step of our information retrieval system.

A. Our Work: Information Retrieval

As we have said earlier the use of the field of information retrieval is important in our work either to evaluate the initial population for the genetic algorithm (GA) or for the reproduction of a new generation when applying the GA.

As our job is to find a pedagogical content for a learner profile the first thing to do is to save the profile as well the different pedagogical objectives for a course in text files, the idea is that we go to start looking for the pedagogical objectives those are relevant to the learner profile, it is like we want to find the relevant documents for a query expressed by a user, in our case the profile of the learner is like a query.

So at the beginning it is like an information retrieval system (IRS) we will index the learner's profile and the pedagogical objectives by creating representative vectors using the vector model [2] where each element is the weight of the term or concepts in the documents (profile or pedagogical objective), we will calculate the similarity between the learner profile and each pedagogical objective using Cosine similarity, Fig. 1 indicates the different steps to find the relevant objectives for the learner's profile.

B. Our Work: Genetic Algorithms

The second step in our work is the application of the genetic algorithm to find a single pedagogical objective that is optimal for the learner profile by pressing on the result of the information retrieval system to find the initial population.

After we calculate the similarity between the learner's profile and each pedagogical objective, we sort the result obtained in decreasing order to keep the first eight objectives that will play the role of the initial population this stage is called the selection stage of individuals who are most adapted to the working environment of the genetic algorithm.

For our problem of the research of the optimal pedagogical objective for the learner profile, individuals are the pedagogical

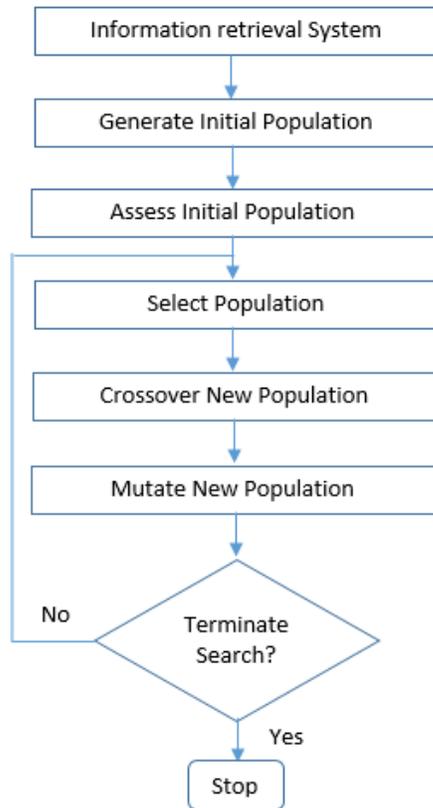


Fig. 3. The process of our genetic algorithm.



Fig. 4. Facebook application.

The RestFB API needs some configuration like creating a Facebook application in the Facebook developers space⁴ this application contains a name, a working domain name must be the same domain name for our platform, an application ID and apps secret ID, the last two parameters are the ones that give us the ability to collect data from the learner account, Fig. 4 illustrates a Facebook application under the name madani:

The second step which is very necessary for using the RestFB API is to install the login button that is to say, make a connection between the facebook application using one of the SDK proposed by Facebook like PHP SDK or JavaScript SDK, in this work we tried to use JavaScript SDK with JavaEE.

After creating the Facebook application and the login button we can retrieve a parameter that is used by RestFB and which allows us to retrieve the learner's publications for using it in the search phase of the period of activity, it is the ACCESS TOKEN.

After all these steps we will have the publications of the learner, and by using the API RestFB and the formula 5 we can easily find the period of activity.

2) *Sentiment Analysis*: After we find the period of activity of the learner (the period of the day when the learner was very active), we collect all the publications that are published in this period and filter them in relation to publications that are related to education, to do this work we use a list of terms related to education such as learning, learn, teacher, school, university, students, faculty, education, teaching, training, etc.

The idea is that among the publications of the period of activity we keep only those are in relation to teaching and education, and afterwards we classify these publications according to three classes: motivate, demotivate or neutral.

For the classification, we use the AFINN dictionary, but before the classification process we have to make text pre-processing on the publications to decrease the noise that will influence the classification, such as:

- **Tokenization**: Which is the phase of splitting the tweet into terms or tokens by removing white spaces, commas and other symbols etc, it's an important step because in our works we focus on individual words to search them is the AFINN dictionary or in WordNet.
- **Removing Stop Word**: It removes The words that have no effect on the classification of tweets like preposition and the article (a, an, the).
- **Removing URL**: The URLs have no effect on the classification so it is important to eliminate them.
- **Removing numbers**
- **Stemming**: Stemming is another very important process. In our work and because we focus on English language we use the porter stemming [19].

After the text pre-processing phase, For a given publication of the learner in its Facebook account we go through its words and we look for their weights from the AFINN dictionary, after we sum the weights of all words of the publication those exist in AFINN, after that we calculate the average of the weights and using a threshold we classify the tweet according to three classes Motivate, demotivate or neutral.

Our proposal is to work with dictionary-based approach, for that, we use the AFINN dictionary which contains words in English with a weight that can take a value between 5 and -5 (strongly positive, mildly, strongly negative, etc.) for example the word "abandoned" has the value -2 and the word "accept" has the value 1, etc. and to look for the feeling of a publication the idea is to browse all the words in it and sum these weights using AFINN and after using a threshold to classify the publication into motivate, demotivate or neutral.

In Table I, we present the classification and error rate with the use of AFINN before and after the application of text pre-processing (TP), and from this table, we remark that the text pre-processing increase the classification rate and decrease the error rate.

⁴<https://developers.facebook.com/>

TABLE I. CLASSIFICATION AND ERROR RATE

Method	Classification rate	Error rate
Without TP	0.57	0.43
With TP	0.76	0.24

VI. PARALLELIZATION OF OUR WORK

In an e-learning platform it is possible to find many learners also many courses and for each course many chapters and concepts, so a large amount of data which makes learning in the platform ineffective because of the time that we can wait to find the result of analyzing a learner's profile, also today everybody publishes in social networks so for a learner it is possible that we find a lot of publication in his facebook account (huge volume of publications) so the analysis of his publications and his profile becomes difficult with the classic means.

Our proposal which is the main work of this article is to parallelize our genetic algorithm to find the optimal pedagogical objective for the profile of the learner, also we parallelize the classification of the publications for finding the sentiment of the learner in the period of activity.

We propose to work in a big data system using HADOOP's solutions; Hadoop Distributed File System (HDFS) and the programming model MapReduce Which is a Java API for writing distributed programs for information retrieval ,The idea is to install a cluster Hadoop (set of machines) and to share the work of the analysis of the learner's publications and the search for the optimal pedagogical objective between the machines of the cluster.

A. Search for the Optimal Pedagogical Objective

1) *Description:* For that we record the profile of the learner as well as the pedagogical objectives for a course in text files and save them in HDFS to parallelize the storage between the different machines of the Hadoop cluster, also we record the result of the analysis of the learner's profile in HDFS as the optimal chromosome.

To optimize the storage volume of the data in HDFS we stored all the pedagogical objectives in the same file instead of giving each objective a file because the file system (HDFS) will give for each file the size of the block which is 64MB therefore for example if a file has a size of 10MB after storage in HDFS its size will be 64MB therefore an increase in the size of the stored data, so we put all the objectives in the same file to optimize the storage size.

After a learner passes a QUIZ for such course, he gives us an idea of the concepts that he does not master and it is these concepts that will form his profile, after we save these concepts in a text file, the same applies to each pedagogical objective we save it in a file with the name and email of the teacher that he puts it on the platform, so in the document of the pedagogical objectives we find each objective with a name of teacher, his email And the concepts that the learner must learn to validate the course.

The different parallelization's steps of our work are the following:

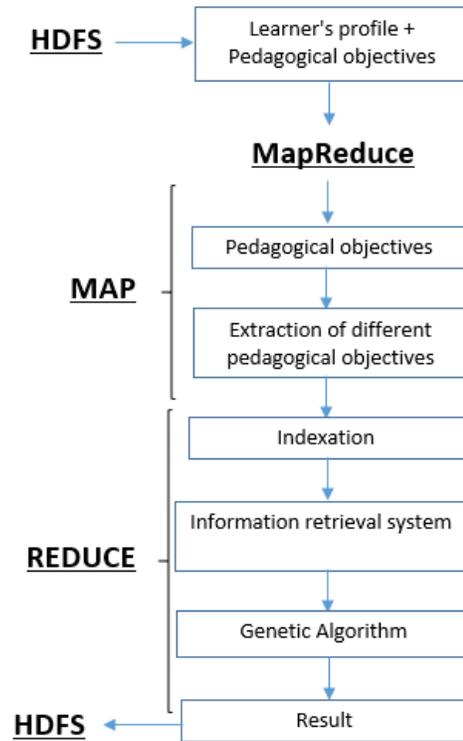


Fig. 5. Steps of parallelization.

- **The recording of the data to be analyzed:**
As I have already said we record the learner profile and the pedagogical objectives (same text files) in HDFS.
- **MapReduce program:**
The parallelization of our work is done with the programming model MapReduce which is a Java API allowing the writing of the distributed programs, it Constitutes of two main operations, the Map operation and the Reduce operation.
In our case, the Map method allows us to browse the file containing the learning objectives and to extract all the pedagogical objectives after sending the result to the Reduce method which allows us to index the learner's profile and Pedagogical objectives, the application of the information retrieval system and the application of the genetic algorithm in a parallel manner.
- **Recording of the analysis result:**
The last step is the recording of the result obtained after the MapReduce operation in HDFS which is the optimal pedagogical objective for the learner's profile in the form of a chromosome formed by genes with either the value 1 meaning that The learner must learn the equivalent concept, or the value 0 otherwise.

2) *Schematization of Parallelization Steps:* Fig. 5 gives the different steps for parallelizing the adaptive learning algorithm.

3) *MapReduce Algorithm:* Our MapReduce algorithm followed To find the optimal goal for a learner is the following:

Class Mapper

```
Method Map(Docid,FileOfObjective)
  For each line ∈ FileOfObjective
    Write(Docid,line)
  End for
```

Class Reducer

```
Method Reduce(Docid,List(line))
  S ← NULL
  For each n ∈ List(line)
    S ← S + n
  End for
  List2 ← Split(s)
  Index(List2 & Learner's profile)
  Result ← BuildingGA()

Write(Result," ")
```

With:

- **Split(s):** Extracts the various pedagogical objectives.
- **Index(List2 & Learner's profile):** A function which makes the vector representation of each pedagogical objective and of the learner's profile.
- **BuildingGA():** The execution of the genetic algorithm

B. Search for Sentiment of the Learner

After we collect the Facebook publications of the period of activity either with the RestFB API or with Apache Flume we store them directly in HDFS to parallelize the classification.

1) *Parallelization Steps:* Fig. 6 shows the different steps to parallelize the classification using HDFS and MapReduce.

2) *Description of Parallelization:* From Fig. 6 the first step of the parallelisation is to store the data set (publications of the learner) to classify in HDFS for sharing the storage between several machines (Hadoop cluster) and after it is the step of classification by applying The AFINN method.

The input of the MapReduce operation at each iteration contains a publication to classify and the output contains the sentiment of the learner, the result of the classification is stored in HDFS.

Our MapReduce algorithm followed for the classification of the publications is the following:

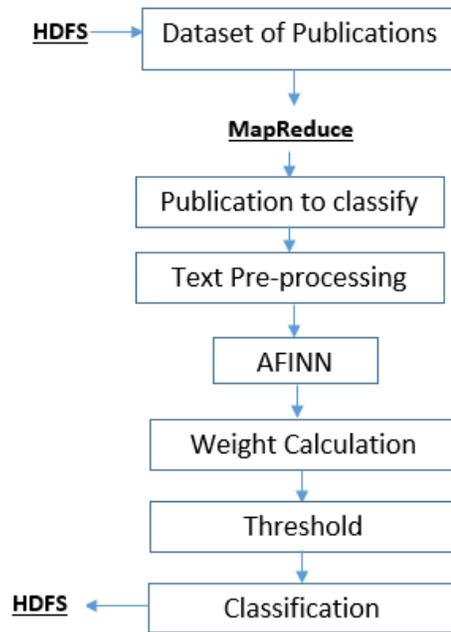


Fig. 6. Parallelization steps.

Class Mapper

```
Method Map(publication)
  S ← 0
  publication ← TextPreProcessing(publication)
  SE[] ← Split(publication)
  For each word ∈ SE
    S ← S+AFINN(word)
  End for
  sentiment ← s/(length(SE))
  write(sentiment,1)
```

Class Reducer

```
Method Reduce(sentiment,ListOfOne(1,1,1,...))
  S ← 0
  For each n ∈ ListOfOne(1,1,1,...)
    S ← S+n
  write(sentiment,S)
```

Where,

- **TextPreProcessing(publication):** Apply different types of text pre-processing
- **AFINN(word):** Calculate the weight of feelings equivalent to word if it is existent in AFINN

After the calculation of the period of activity (PA) which will give us an idea of the part of the day when the learner must learn, the classification of the learner's publications which are published in PA according to three classes (motivate, demotivate, neutral) gives the platform's teacher an idea of the level with which the learner must learn and it depends on



Fig. 7. Configuration of our cluster.

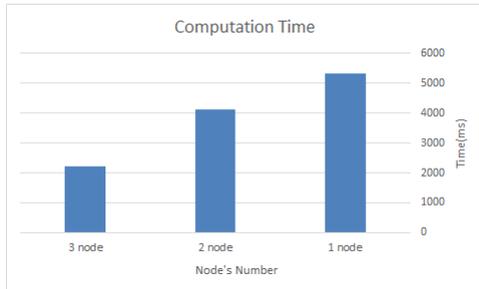


Fig. 8. Calculation time.

his motivation, of course, if he is motivated we can increase the level of study, and the opposite if he is demotivated.

VII. PARALLELISM EFFECT

As we said earlier our work consists of parallelizing the search for an optimal pedagogical content for the learner in a big data system using the Hadoop framework with Hadoop Distributed File System (HDFS) to distribute the data storage necessary for the Analysis (pedagogical objectives, learner profile and the learner publications), and to record the result of the analysis (the optimal chromosome and the sentiment of the learner) to facilitate the interpretation afterwards and also to parallelize the classification of the publication.

The goal of parallelizing the work is to reduce the computation time if we have a large size of the data (profile, objectives and publications), for this and to demonstrate the effect of sharing the work between several machines we decide to work with a variable number of nodes (Hadoop machines).

A. Configuration of our Cluster

Fig. 7 shows the configuration of our cluster (Hadoop machine set) which contains three Hadoop machines, one master machine and two slave machines, so our work will be shared between three machines to reduce the computation time. The nodes used are three nodes with the Linux UBUNTU 15.04 operating system.

According to Fig. 7, our Hadoop cluster contains three UBUNTU machines that are all three enabled with a total storage of 24GB (8GB for each node).

B. Effect of Parallelization

Fig. 8 represents the calculation time of our work using respectively 1, 2 and 3 Hadoop nodes.

From Fig. 8 we notice that by increasing the number of nodes in the cluster the computation time decreases and this is the goal of sharing the work between several machines.

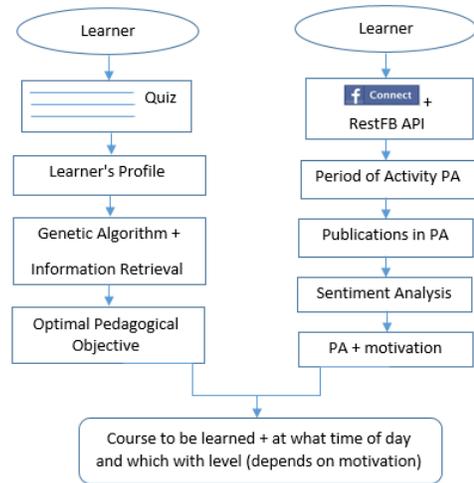


Fig. 9. Different steps of our adaptive e-learning system.

A very important information to note is that a MapReduce program which works with one node works also with three nodes and can also work with a cluster of 100 nodes, so if we want to decrease the calculation time just we need to add another node to the cluster which is a very easy operation.

VIII. SUMMARY OF OUR ADAPTIVE E-LEARNING SYSTEM

Fig. 9 presents the different steps to adapt a course to a learner based on his/her profile, the period of activity and his motivation.

According to Fig. 9, our system is based on two levels of adaptation, the first one is in the form of an optimization problem which makes it possible to find the pedagogical objectives which are optimal in relation to the learner's profile, and the second is to use the data of the learner published in his social networks accounts to find his period of activity and his motivation related to education (publications that are related to education).

At the end of all these processes, we find the courses that the learner must learn, at what period of the day (the period when the learner is active) and with what level (depends on his motivation).

IX. CONCLUSION AND PERSPECTIVES

In this paper, we have presented our system of adaptive learning based on a genetic algorithm, the notions of information retrieval systems, Data from social networks, New measure called period of activity, sentiment analysis and Big Data technologies (Hadoop, HDFS, MapReduce); by searching the pedagogical content that is optimal for a Learner based on his profile and a set of pedagogical objectives set by a teacher, the period of activity (when the learner is active) and with what level the learner must learn (his motivation).

In this work, we have proposed a measure called the period of activity that it helps us to find the period of the day when the learner is active (publish a lot of publications in his social networks accounts).

We proposed also a new approach to parallelize the algorithm which allows to search the optimal pedagogical objective and for finding the motivation of the user by using a Hadoop cluster of three-node (UBUNTU machines) to share the work between several machines.

With the use of a Hadoop cluster the computation time decreases by increasing the number of nodes in the cluster which facilitates the search of the pedagogical content for the learner if we have a large volume of the data to be analyzed, and also facilitate the search of the motivation of the learner if we have a lot of publications in his social network account.

Our next work is to develop a new approach for the sentiment analysis without the use of AFINN dictionary based on the semantic, and also use a new optimization algorithm of artificial intelligence to find a course adapted to the learner's profile.

REFERENCES

- [1] Samia Azough, Mostafa Bellafkih and El Houssine Bouyakhf, Adaptive E-learning using Genetic Algorithms, IJCSNS International Journal of Computer Science and Network Security, VOL.10 No.7, July 2010
- [2] Salton, G. Automatic text processing: the transformation, analysis, and retrieval of information by computer. New York: Addison- Wesley Publishing Co. Inc. 1989.
- [3] David, L. Handbook of Genetic Algorithms. New York : Van Nostrand Reinhold. 1991.
- [4] BANGORN KLABBANKOH, OUEN PINNGERN PH.D., APPLIED GENETIC ALGORITHMS IN INFORMATION RETRIEVAL Md. Abu Kausar, Md. Nasar and Sanjeev Kumar Singh, A Detailed Study on Information Retrieval using Genetic Algorithm, Journal of Industrial and Intelligent Information Vol. 1, No. 3, September 2013, doi: 10.12720/jiii.1.3.122-127
- [5] Laith Mohammad, Qasim Abualigah and Essam S. Hanandeh, APPLYING GENETIC ALGORITHMS TO INFORMATION RETRIEVAL USING VECTOR SPACE MODEL, International Journal of Computer Science, Engineering and Applications (IJCEA) Vol.5, No.1, February 2015
- [6] Vrajitoru, 1998, Crossover improvement for the genetic algorithm in information retrieval, Information Processing and Management: an International Journal Volume 34 Issue 4, July 1, 1998
- [7] Philomina Simon and S. Siva Sathya, Genetic Algorithm for Information Retrieval, International Conference on Intelligent Agent & Multi-Agent Systems, IAMA 2009, doi: 10.1109/IAMA.2009.5228033
- [8] Ahmed A. A. Radwan, Bahgat A. Abdel Latef, Abdel Mgeid A. Ali, and Osman A. Sadek, Using Genetic Algorithm to Improve Information Retrieval Systems, International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol:2, No:5, 2008
- [9] W. Fan, M.D. Gordon and P. Pathak. Personalization of search engine services for effective retrieval and knowledge management, in: Proc. 2000 International Conference on Information Systems (ICIS), Brisbane, Australia, 2000.
- [10] Hawkes, L. W., Derry, S. J. Advances in local student modeling using informal fuzzy reasoning. International journal of human-computer studies, 45(6), 697-722, 1996
- [11] Ruiz, M. D. P. P., Barriales, S. O., Prez, J. R. P., Rodriguez, M. G. Feijoo. net: an approach to personalized E-learning using learning styles. In Web Engineering (pp. 112-115). Springer Berlin Heidelberg, 2003
- [12] Chrysaftadi, K., Virvou, M. Evaluating the integration of fuzzy logic into the student model of a web-based learning environment. Expert Systems with Applications, 39(18), 13127-13134, 2012
- [13] Martin, J., VanLehn, K. Student assessment using Bayesian nets. International Journal of Human-Computer Studies, 42(6), 575-591, 1995
- [14] Vicari, R., Flores, C. D., Seixas, L., Gluz, J. C., and Coelho, H. AMPLIA: A Probabilistic Learning Environment. International Journal of Artificial Intelligence in Education, 18(4), 347-373, 2008
- [15] Romero, C., Ventura, S., and De Bra, P. Knowledge discovery with genetic programming for providing feedback to courseware authors. User Modeling and User-Adapted Interaction, 14(5), 425-464, 2004
- [16] Chang, T. Y., and Ke, Y. R. A personalized e-course composition based on a genetic algorithm with forcing legality in an adaptive learning system. Journal of Network and Computer Applications, 36(1), 533-542, 2013
- [17] Manju Bhaskar, Minu M Das, Dr. T. Chithralekha and Dr. S. Sivasatya, Genetic Algorithm Based Adaptive Learning Scheme Generation For Context Aware E-Learning, Manju Bhaskar et. al. / (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 04, 2010, 1271-1279
- [18] D. Ganesh & V. V. Rama Prasad, Protection of Shared Data Among Multiple Users for Online Social Networks, International Conference on Contemporary Computing and Informatics (IC3I), 2014.
- [19] M.F. Porter, An algorithm for suffix stripping, Originally published in Program, 14 no. 3, pp 130-137, July 1980.

Solving the Free Clustered TSP Using a Memetic Algorithm

Abdullah Alsheddy

College of Computer and Information Sciences
Al Imam Mohammad Ibn Saud Islamic University (IMSIU)
Riyadh, Saudi Arabia

Abstract—The free clustered travelling salesman problem (FCTSP) is an extension of the classical travelling salesman problem where the set of vertices is partitioned into clusters, and the task is to find a minimum cost Hamiltonian tour such that the vertices in any cluster are visited contiguously. This paper proposes the use of a memetic algorithm (MA) that combines the global search ability of Genetic Algorithm with local search to refine solutions to the FCTSP. The effectiveness of the proposed algorithm is examined on a set of TSPLIB instances with up to 318 vertices and clusters varying between 2 and 50 clusters. Moreover, the performance of the MA is compared with a Genetic Algorithm and a GRASP with path relinking. The computational results confirm the effectiveness of the MA in terms of both solution quality and computational time.

Keywords—Combinatorial optimization; clustered travelling salesman problem; memetic algorithm; guided local search; genetic algorithm

I. INTRODUCTION

The travelling salesman problem (TSP) is one of the best known and most widely studied combinatorial optimization problems. Many variants of the TSP have been proposed and solved during the last decades. This paper focuses on the clustered travelling salesman problem (CTSP), a variant of the TSP that was introduced by Chisman [1]. Similar to the TSP, the objective of the CTSP is to construct a Hamiltonian path with minimum distance, visiting all cities exactly once. Cities in the CTSP, however, are partitioned into predefined clusters and all cities belonging to the same cluster should be visited consecutively.

The CTSP has several applications in various fields. Examples of CTSP applications include automated warehouse routing [1], shops and grocery suppliers [2] and emergency vehicle dispatching [3] in the vehicle routing domain; disk fragmentation and computer operations in the IT domain [4]; machine scheduling and production planning [5] in the manufacturing domain; and microscopy (cytology) [4].

Most of the related research addressed the so-called *ordered* CTSP (OCTSP) in which the clusters has to be visited in a prespecified order. Although such a prespecified order is not necessarily defined in real-life applications, there are few algorithms developed for the CTSP without a pre-order [6]. This variant of the CTSP is referred to as the free CTSP (FCTSP).

This paper considers the FCTSP which can be formally defined as follows. Given a complete undirected graph $G = (V, E)$ with vertex set $V = \{v_1, v_2, \dots, v_n\}$, and edge set

$E = \{(v_i, v_j) : v_i, v_j \in V, i \neq j\}$. The vertex set V is partitioned into m predefined clusters: V_1, V_2, \dots, V_m . Assuming that a non-negative cost or distance c_{ij} is associated with each edge $(v_i, v_j) \in E$, the FCTSP consists of determining a least cost Hamiltonian cycle on G such that the vertices of each cluster are visited contiguously and the clusters can be visited in any order. For illustration, Fig. 1 shows examples of several solutions to a FCTSP.

There have been several exact and metaheuristics algorithms developed for the OCTSP. An approximation algorithm and some other heuristics were proposed in [7]. The LBD-COMP is an exact partitioning algorithm proposed in [8]. Other approximation algorithms were developed in [9] and [10]. A hybrid of Tabu Search (TS) and Genetic Algorithm (GA) was developed for solving the OCTSP in [11]. This hybrid algorithm runs multiple TS search threads while periodically applying a phase of diversification using the Edge Recombination crossover operator to generate offspring solutions that will seed the TS search threads again. Ahmed [12] developed a hybrid genetic algorithm using sequential constructive crossover, the 2-opt algorithm and local search for the OCTSP.

For the FCTSP, a genetic algorithm was proposed in [13], which first searches for an optimal inter-cluster edges and then the intra-cluster edges. The two-level Genetic Algorithm (TLGA) is another algorithm developed for the FCTSP in [14]. The TLGA consists of two interrelated levels; the lower level focuses on finding the shortest Hamiltonian cycle for each cluster, whereas the higher level constructs the complete tour by randomly deleting an edge from each cycle and then heuristically connecting the clusters through the intra-cluster edges. As reported in [14], the TLGA performed well in comparison to other GA variants. Later, Mestria et al. [6] developed several path-relinking strategies incorporated to a Greedy Randomized Adaptive Search Procedure (GRASP) for solving the FCTSP. The best performing heuristic was the GRASP that uses path-relinking in each iteration and as a post-optimization strategy, outperforming other GRASP variants and the TLGA [14].

The GRASP algorithm proposed in [6] has two important characteristics: 1) it deals with the whole FCTSP in a single phase without differentiation between the search for the inter-cluster and for the intra-cluster edges, unlike the two phases approach of the TLGA; and 2) the underlying local search procedure implements the well-known 2-Opt heuristic as one of the most effective local search heuristic for the classical TSP. These characteristics suggest the potential of adapting and then applying successful TSP heuristics to the FCTSP.

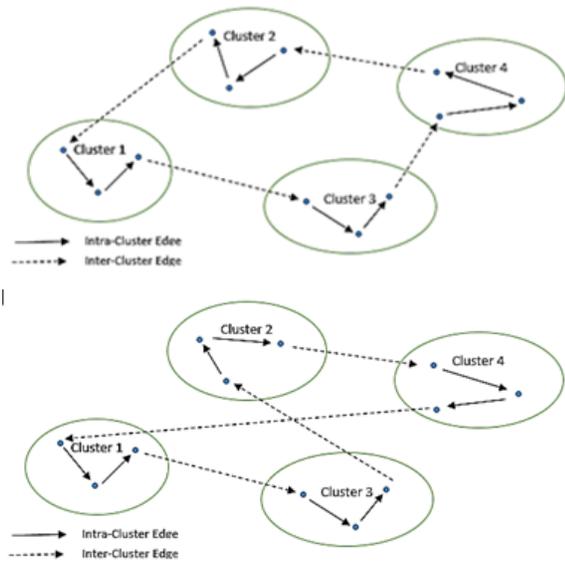


Fig. 1. Examples of two possible solutions to a FCTSP with 4 clusters and 12 vertices.

Thus, we propose in this paper a memetic algorithm that combines the global search ability of Genetic Algorithm with a local-search-based metaheuristic, namely Guided Local Search (GLS) [15], to refine solutions to the FCTSP. GLS, which is a successful algorithm for the TSP [16], incorporates a straight forward extension of the 2-Opt heuristic to handle the clustering constraint of FCTSP. The performance of the proposed memetic algorithm is evaluated through several experiments that include comparisons with the TLGA [14] and the GRASP algorithm [6].

This paper is organized as follows. The proposed approach and its application to the FCTSP is described in section II. Then, the experimental results in comparison with state-of-the-art algorithms are presented in Section III. Finally, concluded remarks and further research are given in Section IV.

II. PROPOSED APPROACH

A. Memetic Algorithms

Local search is the basis of many heuristic methods for combinatorial optimization problems. Starting from an initial solution, local search algorithms iteratively improve the current solution by exploring its neighbourhood for better movements. Although local search algorithms usually return good solutions, these would easily get stuck in local optima, which are typically overcome by an escape mechanism as in Tabu Search, Simulated Annealing and Guided Local Search (GLS) [17]. In GLS, the escaping mechanism is based on augmenting the objective function with penalties. Every time the underlying local search algorithm reaches a local optimal, GLS augments the cost function by adding penalties to selected bad features, and then restarts the local search algorithm while using the augmented function $h(s)$ in the search process instead of the main objective function, $g(s)$:

$$h(s) = g(s) + \lambda \sum_{i \in F} p_i * I_i(s) \quad (1)$$

where, s is a candidate solution and λ is a parameter of the GLS algorithm. F refers to the set of all features that were used to distinguish between solutions with different characteristics. p_i is the penalty of feature i (each p_i is initialized to 0, and incremented by 1 whenever it is selected for penalization), and $I_i(s)$ is an indicator which is equal to 1 only if s exhibits the feature i ; 0 otherwise. GLS penalizes the most costly features in the current solution, weighted by the number of times the feature has been penalized so far.

In contrast to local search algorithms, global search algorithms try to overcome local optima in order to find more globally optimal solutions. Genetic Algorithms are very popular global search algorithms which have been successfully applied to many combinatorial optimisation problems. A GA is a population-based search technique inspired from the biological principals of natural selection and genetic recombination. At every generation, a GA maintains a population of individuals that represent candidate solutions to the problem. This population evolves throughout the optimization process to find global solutions by applying reproduction operators. Each individual is evaluated to give some measure of its “fitness”. Selection of parents for reproduction is based on their fitness. The reproduction operators include crossover and mutation which are both applied with certain probabilities. The evolution of the GA continues until either an optimal solution is found, or some other stopping criteria have been met.

Memetic algorithms [18] denote a broad class of metaheuristics that extend global search methods, such as GA, by incorporating problem-specific knowledge, usually in the form of a local search strategy or through the use of specialised search operators. Thus, a memetic algorithm hybridizes global and local search, such that the individuals of a population in a global search algorithm have the opportunity for local improvement in terms of local search. The applications of MAs are enormous, including areas such as routing, assignment and planning problems [18].

B. Memetic Algorithm for the FCTSP

Our memetic algorithm (MA) is a hybrid of GA for global search and GLS for local search. In GLS, the underlying local search is the 2-Opt heuristic which is a well-known TSP heuristic. GLS is a simple local-search-based metaheuristic with only one parameter to tune. Nevertheless, GLS was shown to be a very effective method for the TSP [16] and other routing and planning problems [15]. In the proposed MA, GLS is complemented by the genetic operators to enhance the exploration of the space of FCTSP solutions.

The proposed MA is outlined in Algorithm 1. It starts by randomly generating an initial population of N solutions. Each individual in this population undergoes local improvement by applying GLS. Then, the algorithm iteratively evolves the population by applying genetic operators and GLS. Since GLS is more computationally expensive than the genetic operators, it is performed in the proposed MA periodically, every 10 generations. A description of the algorithm design for the FCTSP is given below:

C. Guided Local Search (GLS)

In the proposed method, a solution of the FCTSP is a tour that is represented by a permutation (i.e. vector) of cities.

```
MA(Problem instance (problem), population size  
(N), stopping criterion (maxGene))  
  P ← RandomPopoulation(problem, N);  
  P* ← GuidedLocalSearch(P);  
  for generation ← 1 until maxGene do  
    P' ← MatingSelection(P*);  
    P' ← GeneticOperators(P');  
    if generation mod 10 = 0 then  
      P' ← GuidedLocalSearch(P');  
    end if  
    P* ← SurvivorSelection(P ∪ P');  
  end for  
  return P*;
```

Algorithm 1: Memetic Algorithm

The permutation determines the order of the cities in the tour. The 2-Opt heuristic is a well-known and very simple, yet effective local search algorithm for the classical TSP. The 2-Opt heuristic iteratively improves an initial tour by testing all neighbour solutions obtained by applying the 2-exchange neighbourhood operator (i.e. a neighbour is obtained from the current tour by deleting two edges and reconnecting the two resulting paths with the only possible way that yields a new tour). To implement the 2-Opt heuristic for the FCTSP, the feasibility of the newly generated solutions with respect to the clustering constraint is maintained by applying the 2-exchange operator to any two non-adjacent edges if and only if both edges are in the same cluster or are inter-cluster edges.

As reported in [16], GLS can sit on top of the 2-Opt heuristic and guide it to escape local minima in an efficient and effective manner. GLS converges quickly to a close to optimal solution, particularly when it is coupled with Fast Local Search [15]. The latter is a general method that divides a neighbourhood set into sub-groups. Each subgroup is associated with an activation bit, to control which sub-groups will be explored during the search process. The proposed MA incorporates GLS as the local search procedure, and the same algorithm design presented in [16] to implement GLS for the FCTSP are followed in this study.

The key element of GLS is the definition of a set of solution features. A feature should contribute to part of the overall solution cost. For the FCTSP, a tour includes a number of edges, and each edge is associated with a cost (edge length). Therefore, the set of all edges defines the set of features for the FCTSP. Each tour either includes (i.e. exhibits) an edge (i.e. feature) or not.

D. Fitness Function and Mating Selection

For the FCTSP, the fitness of each individual chromosome in the population is the length of the entire tour specified by the chromosome. Mating selection determines the procedure to choose individuals from the current population to undergo the genetic operators. In the proposed MA, the idea is to use a selection strategy that favours exploration. Thus, all individuals in the current population are subjected to undergo the genetic operators. This is attained by randomly ordering parents, and then the genetic operator will use the first and second parents to generate the first offspring, the second and third parents to generate the second offspring, and so on.

E. Genetic Operator

The genetic operators, both crossover and mutation, for the FCTSP can be used at the inter-cluster level by changing the visiting sequence of clusters, or at the intra-cluster level by changing the gene segment for each cluster. In this study, the genetic operator at the inter-cluster level is implemented in order to intensify exploration. The following describe the details of the crossover and mutation operators.

Among the several effective crossover operators that have been proposed for the TSP and its variants is the sequential constructive crossover (SCX). The SCX operator has been modified and applied to the OCTSP in [12]. We follow the same implementation to apply the SCX to the FCTSP, however, with slight modifications. The following procedure describes how the offspring is constructed from *Parent*₁ and *Parent*₂ using the modified SCX:

- Step 1: The first vertex of *Parent*₁ is chosen to be the first gene of the offspring chromosome.
- Step 2: Given the current vertex *v* of the offspring chromosome, and the two candidate vertices *v*₁ and *v*₂ that represent the first legitimate vertices appeared after *v* in the chromosome of *Parent*₁ and *Parent*₂ respectively, the next gene in the offspring chromosome will be *v*₁ if it is nearer to *v*, and *v*₂ otherwise.
- Step 3: Once all vertices in the current cluster are added to the offspring chromosome, move to the next cluster according to the order of clusters in *Parent*₁ and repeat Step 2 until the offspring chromosome is completed.

Mutation plays an important role to help GA avoids establishing a uniform population unable to evolve. It usually modifies the genes of a chromosome selected with a mutation probability. For the FCTSP, the reciprocal exchange mutation operator is implemented, which selects two positions within a chromosome at random and then swaps their contents to produce new chromosomes. The swap is applied to every cluster in the chromosome. This mutation was used in a GA proposed for the OCTSP in [12].

F. Survivor Selection

The survivor selection procedure selects the next generation from parents in the current population and the offspring that are generated by the genetic operators. The proposed MA implements a fitness-biased survivor selection method where all candidate individuals are ranked, and the fitter *N* individuals are chosen to form the population of the next generation.

III. EXPERIMENTS AND RESULTS

This section presents the conducted experiments and their results during the evaluation of the performance of the proposed MA for the FCTSP. The MA was implemented in Java programming language and executed on a PC with 3.40 GHz Intel(R) Core(TM) i7-2600 CPU and 4.00 GB RAM under MS Windows 7 operating system. The algorithm is evaluated on a set of TSPLIB instances¹ as used in [6].

¹<http://comopt.ifl.uni-heidelberg.de/software/TSPLIB95/>

TABLE I. PERFORMANCE OF THE MA VS. PERFORMANCE OF THE GRASP AND THE TLGA ON FCTSP INSTANCES

Instance	TLGA		GRASP		MA	
	%	t_{sec}	%	t_{sec}	%	t_{sec}
5-eil51	8.47	0.4	0	1	0	0.138
10-eil51	2.73	0.4	0	1	0	0.009
15-eil51	7.78	0.4	0	1	0	0.012
5-berlin52	1.44	0.6	0	1	0	0.008
10-berlin52	10.18	0.4	0	1	0	0.007
15-berlin52	14.69	0.4	0	1	0	0.008
5-st70	0.86	1.6	0	2.2	0	0.033
10-st70	1.88	1	0	1.8	0	0.021
15-st70	7.23	0.8	0	1.8	0	0.016
5-eil76	3.94	1.8	0.54	2.4	0	0.298
10-eil76	9.45	1.2	0.71	2.4	0	0.043
15-eil76	2.48	1.2	0.35	2.4	0	0.093
5-pr76	1.78	2	0.92	2.6	0	0.087
10-pr76	1.02	1.2	0.01	2.4	0	0.029
15-pr76	5.95	1.2	0.15	2.4	0	0.127
10-rat99	6.7	3.4	0.24	4.6	0	0.042
25-rat99	23.48	2.4	1.02	4.6	0	0.042
25-kroA100	5.69	2.2	0	4.8	0	0.271
50-kroA100	22.23	2.8	1.02	5	0	0.054
10-kroB100	2.49	3.8	0.07	4.8	0	0.095
50-kroB100	25.36	2.2	0.16	5	0	0.393
25-eil101	6.33	2.2	1.51	4.8	0	2.731
50-eil101	20.34	2.2	2.95	5	0	0.402
25-lin105	19.39	2	0.15	5.2	0	0.199
50-lin105	26.29	3.2	0.54	5.8	0	0.216

A. Parameter Settings

The proposed MA is controlled by a number of parameters that need to be set. The genetic operators include three parameters, namely crossover probability, mutation probability and population size, which are empirically set to 1.0, 0.2 and 20 respectively. The only control parameter of GLS is λ which is calculated as follows: $0.3 \times g^*(s) / |F^*|$, where $g^*(s)$ is the cost of the first local optimal and $|F^*|$ is the average number of features exhibited in a solution. The GLS stops when the best solution is not updated for a $maxIter$ number of consecutive penalizations (i.e. local search calls). The $maxIter$ is set as a function of the problem size, i.e. $maxIter$ is set to the number of the cities in the considered FCTSP instance. The stopping criterion for the MA is controlled by the maximum number of generations ($maxGene$) which is set to 5000.

B. Comparing the MA with State-of-the-Art Techniques

In [6], an algorithm based on GRASP with path-relinking was proposed and compared to the Two-level Genetic Algorithm (TLGA) [14], on a set of small size TSPLIB instances. These algorithms were encoded in the C programming language, and executed on a 2.83 GHz Intel Core 2 Quad with 4 cores and 8 GBs of RAM running the Ubuntu Linux OS (version 4.3.2-1). In order to evaluate the performance of the proposed MA, it is applied to the same set of instances, and make direct comparisons to the results of the GRASP and the TLGA algorithms as presented in [6].

Table I shows this comparative study between MA and both the TLGA and the GRASP methods. It gives, for each combination of algorithm and problem instance, the following performance measures: the average solution quality (%) which represents the mean excess above the best known solution in 20 runs, and the average computational time (t). The reported results for TLGA and GRASP are obtained from [6]. The number that precedes the TSPLIB instance name gives the number of clusters. The results suggest the following remarks:

- In terms of solution quality, the MA solves all the 25 instances to optimality, whereas the GRASP method constantly obtains the optimal solution only on 10 instances, and none of the instances was solved to optimality by the TLGA.
- In terms of the computational time, the comparison cannot be made directly as they were executed in different machines. However, the results show a significant gap between the MA and the other two methods, and the MA is capable to obtain better performance in much less time. On average, the amount of time that the MA requires to solve the considered FCTSP instances to optimality is less than 10% of that of the GRASP algorithm on all of the instances solved by the GRASP algorithm. Even on the unsolved instances by the GRASP method, the MA solves them to optimality in a very short time compared to the GRASP and TLGA methods.

These remarks reveal the outstanding performance of the proposed MA on these FCTSP instances in terms of both solution quality and computational time.

C. Evaluating the MA on Various Instances with Different Clusters

This experiment aims to evaluate the impact of the number and size of the clusters of the FCTSP on the performance of the MA. The experiment is designed as follows:

- A set of 20 TSPLIB instances are selected. They are of various names and sizes up to 318 cities.
- Since the best-known solution for each of these instances is already known, an optimal tour for each instance is obtained, and then the clusters are defined accordingly. Consequently, the best-known solution for the TSPLIB instance will be the same for its FCTSP counterpart.
- Nine FCTSP instances are derived from each TSPLIB instance by using different number of clusters (C) of almost equal sizes, where $C \in \{2, 4, 6, 8, 10, 20, 30, 40, 50\}$.
- On each FCTSP instance, the MA is applied while using similar experimental settings to the previous experiment.

The computational results reveal that the MA always solves to optimality all the FCTSP instances without any sensitivity to the number of clusters. Therefore, only the computational time required by the MA to solve each instance is reported in Table II. The results for selected instances are plotted in Fig. III-C. The results show that the hardness of the FCTSP instance for the MA to solve an instance to optimality grows as the number of clusters shrinks. For example, the complete time required by the MA to find optimal solutions on *kroA200* with 2, 4, 10 and 50 clusters are 16.5, 3.2, 1.9 and 0.4 seconds, respectively.

IV. CONCLUSION

In this study, a new memetic algorithm based on the GA, GLS and 2-Opt algorithms is proposed for solving the free

TABLE II. AVERAGE COMPUTATIONAL TIME REQUIRED BY THE MA TO SOLVE SOME TSPLIB INSTANCES WITH VARIOUS CLUSTERS TO OPTIMALITY

Instance	Clusters								
	2	4	6	8	10	20	30	40	50
kroA100	0.095	0.053	0.037	0.03	0.024	0.028	0.032	0.03	0.028
kroB100	0.063	0.035	0.027	0.022	0.039	0.021	0.024	0.025	0.026
rd100	0.061	0.04	0.037	0.025	0.022	0.02	0.019	0.025	0.027
eil101	0.137	0.033	0.041	0.034	0.023	0.023	0.027	0.034	0.034
lin105	0.186	0.096	0.052	0.042	0.036	0.028	0.031	0.037	0.03
pr107	0.184	0.272	0.112	0.123	0.093	0.105	0.037	0.191	0.153
pr124	0.101	0.051	0.036	0.03	0.028	0.023	0.023	0.026	0.031
bier127	0.234	0.201	0.344	0.301	0.22	0.125	0.18	0.139	0.23
pr144	0.286	0.126	0.24	0.068	0.056	0.038	0.034	0.047	0.063
kroA150	0.854	0.283	0.164	0.105	0.084	0.055	0.064	0.094	0.064
kroB150	1.911	0.718	0.213	0.351	0.273	0.084	0.081	0.122	0.085
ch150	0.756	0.154	0.118	0.085	0.093	0.059	0.07	0.068	0.055
pr152	0.375	0.304	0.273	0.069	0.101	0.074	0.091	0.095	0.085
rat195	1.163	0.853	0.127	0.094	0.096	0.108	0.095	0.086	0.131
kroA200	16.564	3.286	3.991	2.759	1.913	0.917	0.276	0.317	0.416
kroB200	93.54	17.229	3.688	3.603	0.989	0.84	0.592	0.098	0.355
ts225	5.853	1.963	0.65	0.845	0.376	0.141	0.22	0.16	0.182
a280	1.763	0.997	0.533	0.472	0.895	0.264	0.135	0.192	0.252
pr299	69.791	24.579	17.487	3.017	7.634	2.076	0.985	1.715	1.87
lin318	34.536	12.599	7.977	7.249	5.385	1.049	0.701	1.734	1.03

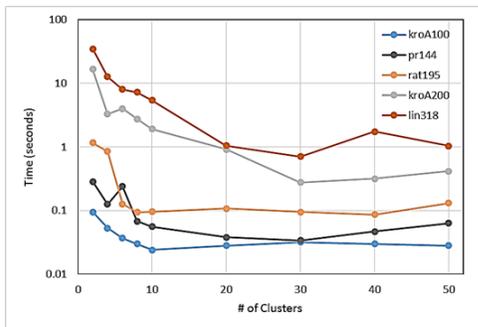


Fig. 2. Average computational time required by the MA as a function of the number of clusters, on selected TSPLIB instances.

clustered travelling salesman problems. In this method, the GA, which implements the sequential constructive crossover and the reciprocal exchange mutation operator, is used for global search; while the GLS algorithm that sits on top of the 2-Opt heuristic is used for local search. The performance of this proposed method is evaluated in terms of solution quality and speed on a set of TSPLIB, with comparison to a GA and GRASP methods. The impacts of the different number of clusters on the performance of the proposed method are also analyzed. The obtained experimental results reveal the outstanding performance of the proposed memetic algorithm which solves to optimality all the instances used in this study in a reasonable amount of time.

REFERENCES

- [1] James A. Chisman. The clustered traveling salesman problem. *Computers and Operations Research*, 2(2):115 – 119, 1975.
- [2] Hassan Ghaziri and Ibrahim H Osman. A neural network algorithm for the traveling salesman problem with backhauls. *Computers and Industrial Engineering*, 44(2):267 – 281, 2003.
- [3] A. Weintraub, J. Aboud, C. Fernandez, G. Laporte, and E. Ramirez. An emergency vehicle dispatching system for an electric utility in chile. *Journal of the Operational Research Society*, 50(7):690–696, 1999.
- [4] G. Laporte and U. Palekar. Some applications of the clustered travelling salesman problem. *Journal of the Operational Research Society*, 53(9):972–976, 2002.

- [5] F.C.J. Lokin. Procedures for travelling salesman problems with additional constraints. *European Journal of Operational Research*, 3(2):135 – 141, 1979.
- [6] Mário Mestria, Luiz Satoru Ochi, and Simone de Lima Martins. GRASP with path relinking for the symmetric euclidean clustered traveling salesman problem. *Computers and Operations Research*, 40(12):3218 – 3229, 2013.
- [7] Michel Gendreau, Gilbert Laporte, and Jean-Yves Potvin. Heuristics for the clustered traveling salesman problem. Technical Report CRT-94-54, Centre de Recherche sur les Transports, Université de Montréal, Montreal, Canada, 1994.
- [8] T Aramgatisiris. An exact decomposition algorithm for the traveling salesman problem with backhauls. *Journal of Research in Engineering and Technology*, 1:151–164, 2004.
- [9] N. Guttmann-Beck, R. Hassin, S. Khuller, and B. Raghavachari. Approximation algorithms with bounded performance guarantees for the clustered traveling salesman problem. *Algorithmica*, 28(4):422–437, 2000.
- [10] Shoshana Anily, Julien Bramel, and Alain Hertz. A 53-approximation algorithm for the clustered traveling salesman tour and path problems. *Operations Research Letters*, 24(12):29 – 35, 1999.
- [11] Gilbert Laporte, Jean-Yves Potvin, and Florence Quilleret. A tabu search heuristic using genetic diversification for the clustered traveling salesman problem. *Journal of Heuristics*, 2(3):187–200, 1997.
- [12] Zakir Hussain Ahmed. The ordered clustered travelling salesman problem: A hybrid genetic algorithm. *The Scientific World Journal*, 2014:1–13, 2014.
- [13] Jean-Yves Potvin and François Guertin. *The Clustered Traveling Salesman Problem: A Genetic Approach*, pages 619–631. Springer US, Boston, MA, 1996.
- [14] Chao Ding, Ye Cheng, and Miao He. Two-level genetic algorithm for clustered traveling salesman problem with application in large-scale tsps. *Tsinghua Science and Technology*, 12(4):459 – 465, 2007.
- [15] Christos Voudouris, Edward Tsang, and Abdullah Alsheddy. Guided local search. *Handbook of metaheuristics*, pages 321–361, 2010.
- [16] Christos Voudouris and Edward Tsang. Guided local search and its application to the traveling salesman problem. *European journal of operational research*, 113(2):469–499, 1999.
- [17] Michel Gendreau and Jean-Yves Potvin. *Handbook of metaheuristics*, volume 146. Springer, 2010.
- [18] Pablo Moscato and Carlos Cotta. *A Modern Introduction to Memetic Algorithms*, pages 141–183. Springer US, Boston, MA, 2010.

Lung Cancer Detection and Classification with 3D Convolutional Neural Network (3D-CNN)

Wafaa Alakwaa
Faculty of Computers & Info.
Cairo University, Egypt

Mohammad Nassef
Faculty of Computers & Info.
Cairo University, Egypt

Amr Badr
Faculty of Computers & Info.
Cairo University, Egypt

Abstract—This paper demonstrates a computer-aided diagnosis (CAD) system for lung cancer classification of CT scans with unmarked nodules, a dataset from the Kaggle Data Science Bowl, 2017. Thresholding was used as an initial segmentation approach to segment out lung tissue from the rest of the CT scan. Thresholding produced the next best lung segmentation. The initial approach was to directly feed the segmented CT scans into 3D CNNs for classification, but this proved to be inadequate. Instead, a modified U-Net trained on LUNA16 data (CT scans with labeled nodules) was used to first detect nodule candidates in the Kaggle CT scans. The U-Net nodule detection produced many false positives, so regions of CTs with segmented lungs where the most likely nodule candidates were located as determined by the U-Net output were fed into 3D Convolutional Neural Networks (CNNs) to ultimately classify the CT scan as positive or negative for lung cancer. The 3D CNNs produced a test set Accuracy of 86.6%. The performance of our CAD system outperforms the current CAD systems in literature which have several training and testing phases that each requires a lot of labeled data, while our CAD system has only three major phases (segmentation, nodule candidate detection, and malignancy classification), allowing more efficient training and detection and more generalizability to other cancers.

Keywords—Lung cancer; computed tomography; deep learning; convolutional neural networks; segmentation

I. INTRODUCTION

Lung cancer is one of the most common cancers, accounting for over 225,000 cases, 150,000 deaths, and \$12 billion in health care costs yearly in the U.S. [1]. It is also one of the deadliest cancers; overall, only 17% of people in the U.S. diagnosed with lung cancer survive five years after the diagnosis, and the survival rate is lower in developing countries. The stage of a cancer refers to how extensively it has metastasized. Stages 1 and 2 refer to cancers localized to the lungs and latter stages refer to cancers that have spread to other organs. Current diagnostic methods include biopsies and imaging, such as CT scans. Early detection of lung cancer (detection during the earlier stages) significantly improves the chances for survival, but it is also more difficult to detect early stages of lung cancer as there are fewer symptoms [1].

Our task is a binary classification problem to detect the presence of lung cancer in patient CT scans of lungs with and without early stage lung cancer. We aim to use methods from computer vision and deep learning, particularly 2D and 3D convolutional neural networks, to build an accurate classifier. An accurate lung cancer classifier could speed up and reduce costs of lung cancer screening, allowing for more widespread

early detection and improved survival. The goal is to construct a computer-aided diagnosis (CAD) system that takes as input patient chest CT scans and outputs whether or not the patient has lung cancer [2].

Though this task seems straightforward, it is actually a needle in the haystack problem. In order to determine whether or not a patient has early-stage cancer, the CAD system would have to detect the presence of a tiny nodule (< 10 mm in diameter for early stage cancers) from a large 3D lung CT scan (typically around $200 \text{ mm} \times 400 \text{ mm} \times 400 \text{ mm}$). An example of an early stage lung cancer nodule shown in within a 2D slice of a CT scan is given in Fig. 1. Furthermore, a CT scan is filled with noise from surrounding tissues, bone, air, so for the CAD systems search to be efficient, this noise would first have to be preprocessed. Hence our classification pipeline is image preprocessing, nodule candidates detection, malignancy classification.

In this paper, we apply an extensive preprocessing techniques to get the accurate nodules in order to enhance the accuracy of detection of lung cancer. Moreover, we perform an end-to-end training of CNN from scratch in order to realize the full potential of the neural network i.e. to learn discriminative features. Extensive experimental evaluations are performed on a dataset comprising lung nodules from more than 1390 low dose CT scans.

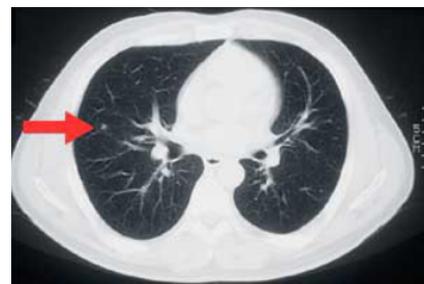


Figure 1: 2D CT scan slice containing a small (5mm) early stage lung cancer nodule.

The paper's arrangement is as follows: Related work is summarized briefly in Section II. Dataset for this paper is described in Section III. The methods for segmentation are presented in section IV. The nodule segmentation is introduced in Section V based on U-Net architecture. Section VI presents 3D Convolutional Neural Network for nodule classification and

patient classification. Our discussion and results are described in details in Section VII. Section VIII concludes the paper.

II. RELATED WORK

Recently, deep artificial neural networks have been applied in many applications in pattern recognition and machine learning, especially, Convolutional neural networks (CNNs) which is one class of models [3]. Another approach of CNNs was applied on ImageNet Classification in 2012 is called an ensemble CNNs which outperformed the best results which were popular in the computer vision community [4]. There has also been popular latest research in the area of medical imaging using deep learning with promising results.

Suk et al. [5] suggested a new latent and shared feature representation of neuro-imaging data of brain using Deep Boltzmann Machine (DBM) for AD/MCI diagnosis. Wu et al. [6] developed deep feature learning for deformable registration of brain MR images to improve image registration by using deep features. Xu et al. [7] presented the effectiveness of using deep neural networks (DNNs) for feature extraction in medical image analysis as a supervised approach. Kumar et al. [8] proposed a CAD system which uses deep features extracted from an autoencoder to classify lung nodules as either malignant or benign on LIDC database. In [9], Yaniv et al. presented a system for medical application of chest pathology detection in x-rays which uses convolutional neural networks that are learned from a non-medical archive. that work showed a combination of deep learning (Decaf) and PiCodes features achieves the best performance. The proposed combination presented the feasibility of detecting pathology in chest x-ray using deep learning approaches based on non-medical learning. The used database was composed of 93 images. They obtained an area under curve (AUC) of 0.93 for Right Pleural Effusion detection, 0.89 for Enlarged heart detection and 0.79 for classification between healthy and abnormal chest x-ray.

In [10], Suna W. et al., implemented three different deep learning algorithms, Convolutional Neural Network (CNN), Deep Belief Networks (DBNs), Stacked Denoising Autoencoder (SDAE), and compared them with the traditional image feature based CAD system. The CNN architecture contains eight layers of convolutional and pooling layers, interchangeably. For the traditional compared to algorithm, there were about 35 extracted texture and morphological features. These features were fed to the kernel based support vector machine (SVM) for training and classification. The resulted accuracy for the CNN approach reached 0.7976 which was little higher than the traditional SVM, with 0.7940. They used the Lung Image Database Consortium and Image Database Resource Initiative (LIDC/IDRI) public databases, with about 1018 lung cases.

In [11], J. Tan et al. designed a framework that detected lung nodules, then reduced the false positive for the detected nodules based on Deep neural network and Convolutional Neural Network. The CNN has four convolutional layers and four pooling layers. The filter was of depth 32 and size 3,5. The used dataset was acquired from the LIDC-IDRI for about 85 patients. The resulted sensitivity was of 0.82. The False positive reduction gotten by DNN was 0.329.

In [12], R. Golan proposed a framework that train the weights of the CNN by a back propagation to detect lung nodules in the CT image sub-volumes. This system achieved sensitivity of 78.9% with 20 false positives, while 71.2% with 10 FPs per scan, on lung nodules that have been annotated by all four radiologists

Convolutional neural networks have achieved better than Deep Belief Networks in current studies on benchmark computer vision datasets. The CNNs have attracted considerable interest in machine learning since they have strong representation ability in learning useful features from input data in recent years.

III. DATA

Our primary dataset is the patient lung CT scan dataset from Kaggle's Data Science Bowl (DSB) 2017 [13]. The dataset contains labeled data for 1397 patients, which we divide into training set of size 978, and test set of size 419. For each patient, the data consists of CT scan data and a label (0 for no cancer, 1 for cancer). Note that the Kaggle dataset does not have labeled nodules. For each patient, the CT scan data consists of a variable number of images (typically around 100-400, each image is an axial slice) of 512×512 pixels. The slices are provided in DICOM format. Around 70% of the provided labels in the Kaggle dataset are 0, so we used a weighted loss function in our malignancy classifier to address this imbalance.

Because the Kaggle dataset alone proved to be inadequate to accurately classify the validation set, we also used the patient lung CT scan dataset with labeled nodules from the Lung Nodule Analysis 2016 (LUNA16) Challenge [14] to train a U-Net for lung nodule detection. The LUNA16 dataset contains labeled data for 888 patients, which we divided into a training set of size 710 and a validation set of size 178. For each patient, the data consists of CT scan data and a nodule label (list of nodule center coordinates and diameter). For each patient, the CT scan data consists of a variable number of images (typically around 100-400, each image is an axial slice) of 512×512 pixels.

LUNA16 data was used to train a U-Net for nodule detection, one of the phases in our classification pipeline. The problem is to accurately predict a patient's label ('cancer' or 'no cancer') based on the patient's Kaggle lung CT scan. We will use accuracy, sensitivity, specificity, and AUC of the ROC to evaluate our CAD system's performance on the Kaggle test set.

IV. METHODS

Typical CAD systems for lung cancer have the following pipeline: image preprocessing, detection of cancerous nodule candidates, nodule candidate false positive reduction, malignancy prediction for each nodule candidate, and malignancy prediction for overall CT scan [15]. These pipelines have many phases, each of which is computationally expensive and requires well-labeled data during training. For example, the false positive reduction phase requires a dataset of labeled true and false nodule candidates, and the nodule malignancy prediction phase requires a dataset with nodules labeled with malignancy.

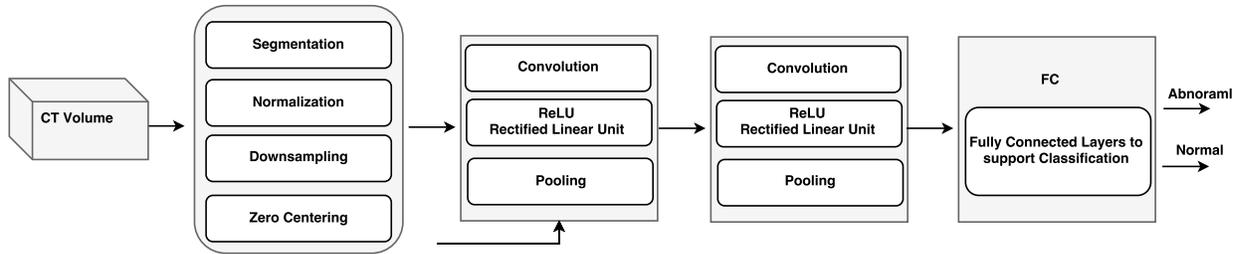


Figure 2: 3D convolutional neural networks architecture.

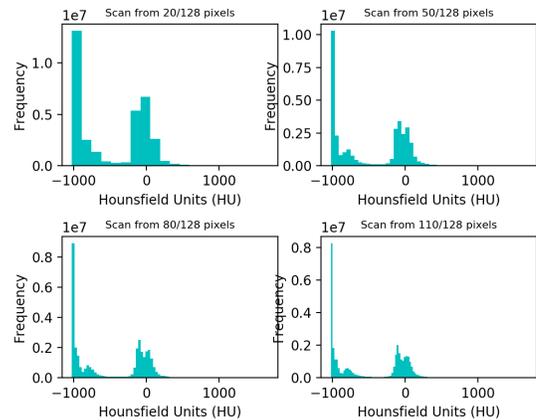
True/False labels for nodule candidates and malignancy labels for nodules are sparse for lung cancer, and may be nonexistent for some other cancers, so CAD systems that rely on such data would not generalize to other cancers. In order to achieve greater computational efficiency and generalizability to other cancers, the proposed CAD system has shorter pipeline and only requires the following data during training: a dataset of CT scans with true nodules labeled, and a dataset of CT scans with an overall malignancy label. State-of-the-art CAD systems that predict malignancy from CT scans achieve AUC of up to 0.83 [16]. However, as mentioned above, these systems take as input various labeled data that is not used in this framework. The main goal of the proposed system is to reach close to this performance.

The proposed CAD system starts with preprocessing the 3D CT scans using segmentation, normalization, downsampling, and zero-centering. The initial approach was to simply input the preprocessed 3D CT scans into 3D CNNs, but the results were poor. So an additional preprocessing was performed to input only regions of interests into the 3D CNNs. To identify regions of interest, a U-Net was trained for nodule candidate detection. Then input regions around nodule candidates detected by the U-Net was fed into 3D CNNs to ultimately classify the CT scans as positive or negative for lung cancer. The overall architecture is shown in Fig. 2, all details of layers will be described in the next sections.

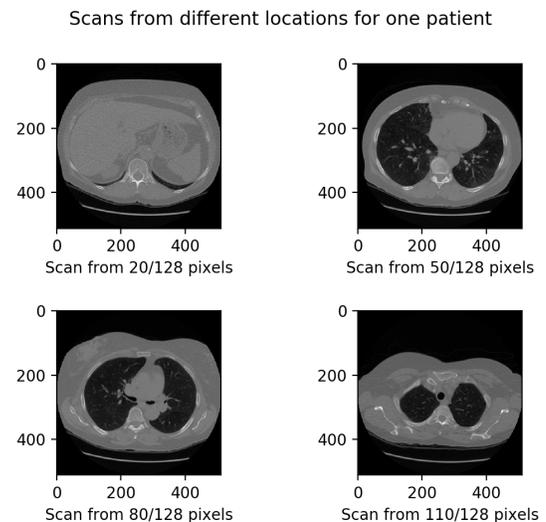
A. Preprocessing and Segmentation

For each patient, pixel values was first converted in each image to Hounsfield units (HU), a measurement of radiodensity, and 2D slices are stacked into a single 3D image. Because tumors form on lung tissue, segmentation is used to mask out the bone, outside air, and other substances that would make data noisy, and leave only lung tissue information for the classifier. A number of segmentation approaches were tried, including thresholding, clustering (Kmeans and Meanshift), and Watershed. K-means and Meanshift allow very little supervision and did not produce good qualitative results. Watershed produced the best qualitative results, but took too long to run to use by the deadline. Ultimately, thresholding was used.

After segmentation, the 3D image is normalized by applying the linear scaling to squeezed all pixels of the original unsegmented image to values between 0 and 1. Spline interpolation downsamples each 3D image by a scale of 0.5 in each of the three dimensions. Finally, zero-centering is performed on data by subtracting the mean of all the images from the training set.



(a) Histograms of pixel values in HU for sample patients CT scan at various slices.



(b) Corresponding 2D axial slices.

Figure 3: 3a Histogram of HU values at 3b corresponding axial slices for sample patient 3D image at various axial.

1) *Thresholding*: Typical radiodensities of various parts of a CT scan are shown in Table I. Air is typically around -1000 HU, lung tissue is typically around -500, water, blood, and other tissues are around 0 HU, and bone is typically around 700 HU, so pixels that are close to -1000 or above -320 are masked

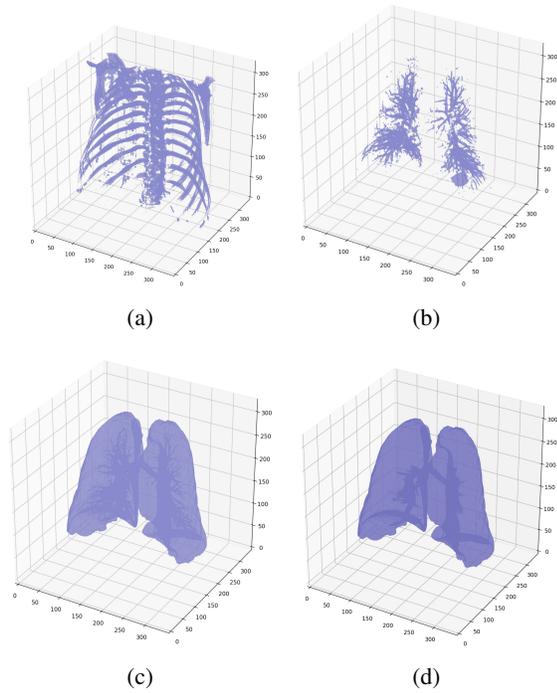


Figure 4: (4a) Sample patient 3D image with pixels values greater than 400 HU reveals the bone segment, (4b) Sample patient bronchioles within lung, (4c) Sample patient initial mask with no air, and (4d) Sample patient final mask in which bronchioles are included.

out to leave lung tissue as the only segment. The distribution of pixel Hounsfield units at various axial slices for a sample patient are shown in Fig. 3. Pixels thresholded at 400 HU are shown in Fig. 3a, and the mask is shown in Fig. 3b. However, to account for the possibility that some cancerous growth could occur within the bronchioles (air pathways) inside the lung, which are shown in Fig. 4c, this air is included to create the finalized mask as shown in Fig. 4d.

Table I: Typical Radiodensities in HU of Various Substances in a CT Scan

Substance	Radiodensity (HU)
Air	-1000
Lung tissue	-500
Water and Blood	0
Bone	700

2) *Watershed*: The segmentation obtained from thresholding has a lot of noise. Many voxels that were part of lung tissue, especially voxels at the edge of the lung, tended to fall outside the range of lung tissue radiodensity due to CT scan noise. This means that our classifier will not be able to correctly classify images in which cancerous nodules are located at the edge of the lung. To filter noise and include voxels from the edges, we use Marker-driven watershed segmentation, as described in Al-Tarawneh et al. [17]. An original 2D CT slice of a sample patient is given in Fig. 5a. The resulting 2D slice of the lung segmentation mask created by thresholding is shown

in Fig. 5b, and the resulting 2D slice of the lung segmentation mask created by Watershed is shown in Fig. 5d. Qualitatively, this produces a much better segmentation than thresholding. Missing voxels (black dots in Fig. 5b) are largely re-included. However, this is much less efficient than basic thresholding, so due to time limitations, it was not possible to preprocess all CT scans using Watershed, so thresholding is used instead.

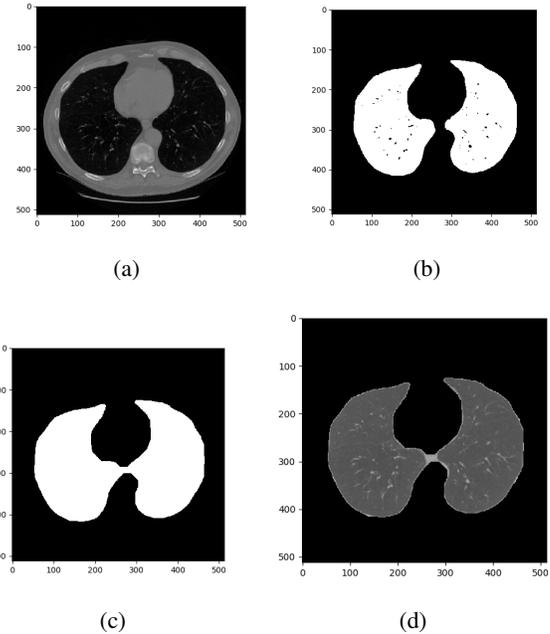


Figure 5: (5a) Original 2D slice of sample patient, (5b) Lung segmentation mask by thresholding of sample patient, (5c) Final watershed segmentation mask of sample patient, and (5d) Final watershed lung segmentation of sample patient.

V. U-NET FOR NODULE DETECTION

Feeding the entire segmented lungs into malignancy classifiers made results very poor. It was likely the case that the entire image was too large search space. Thus feeding smaller regions of interest instead of the entire segmented 3D image is more convenient. This was achieved by selecting small boxes containing top cancerous nodule candidates. To find these top nodule candidates, a modified version of the U-Net was trained as described in Ronneberger et al. on LUNA16 data [18]. U-Net is a 2D CNN architecture that is popular for biomedical image segmentation. A stripped-down version of the U-Net is designed to limit memory expense. A visualization of the U-Net architecture is included in Fig. 6 and is described in detail in Table II. During training, the modified U-Net takes as input 256×256 2D CT slices, and labels are provided (256×256 mask where nodule pixels are 1, rest are 0).

The model is trained to output images of shape 256×256 where each pixel of the output has a value between 0 and 1 indicating the probability the pixel belongs to a nodule. This is done by taking the slice corresponding to label one of the softmax of the final U-Net layer. Corresponding U-Net inputs, labels, and predictions on a patient from the LUNA16 validation set is shown in Fig. 7a, 7b, and 7c, respectively.

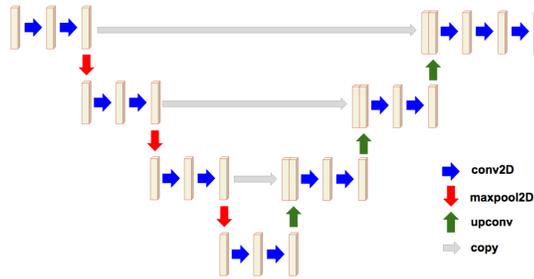


Figure 6: Modified U-Net architecture.

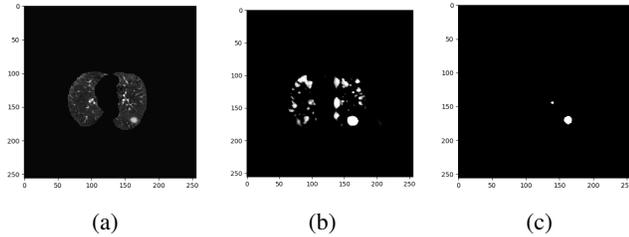


Figure 7: (7a) U-Net sample input from LUNA16 validation set. Note that the above image has the largest nodule from the LUNA16 validation set, which we chose for clarity-most nodules are significantly smaller than the largest one in this image, (7b) U-Net predicted output from LUNA16 validation set, (7c) U-Net sample labels mask from LUNA16 validation set showing ground truth nodule location.

Most nodules are much smaller. A weighted softmax cross-entropy loss calculated for each pixel, as a label of 0 is far more common in the mask than a label of 1. The trained U-Net is then applied to the segmented Kaggle CT scan slices to generate nodule candidates.

VI. MALIGNANCY 3D CNN CLASSIFIERS

Once the U-Net was trained on the LUNA16 data, it is ran on 2D slices of Kaggle data and stacked the 2D slices back to generate nodule candidates¹. Ideally the output of U-Net would give the exact locations of all the nodules, and it would be able to declare images with nodules as detected by U-Net are positive for lung cancer, and images without any nodules detected by U-Net are negative for lung cancer. However, as shown in Fig. 7c, U-Net produces a strong signal for the actual nodule, but also produces a lot of false positives, so we need an additional classifier that determines the malignancy.

Because U-Net generates more suspicious regions than actual nodules, the top 8 nodule candidates are located ($32 \times 32 \times 32$ volumes) by sliding a window over the data and saving the locations of the 8 most activated (largest L2 norm) sectors. To prevent the top sectors from simply being clustered in the brightest region of the image, the 8 sectors were not permitted to overlap with each other. Then these sectors are combined

¹Preprocessing and reading of LUNA16 data code based on <https://www.kaggle.com/arnavkj95/candidate-generation-and-luna16-preprocessing>

Table II: U-Net Architecture (Dropout with 0.2 Probability after each ‘a’ Conv. Layer during Training, ‘Up’ Indicates Resizing of Image via Bilinear Interpolation, Adam Optimizer, Learning Rate = 0.0001)

Layer	Params	Activation	Output
Input			$256 \times 256 \times 1$
Conv1a	$3 \times 3 \times 32$	ReLu	$256 \times 256 \times 32$
Conv1b	$3 \times 3 \times 32$	ReLu	$256 \times 256 \times 32$
Max Pool	2×2 , stride 2		$128 \times 128 \times 32$
Conv2a	$3 \times 3 \times 80$	ReLu	$128 \times 128 \times 80$
Conv2b	$3 \times 3 \times 80$	ReLu	$128 \times 128 \times 80$
Max Pool	2×2 , stride 2		$64 \times 64 \times 80$
Conv3a	$3 \times 3 \times 160$	ReLu	$64 \times 64 \times 160$
Conv3b	$3 \times 3 \times 160$	ReLu	$64 \times 64 \times 160$
Max Pool	2×2 , stride 2		$32 \times 32 \times 160$
Conv4a	$3 \times 3 \times 320$	ReLu	$32 \times 32 \times 320$
Conv4b	$3 \times 3 \times 320$	ReLu	$32 \times 32 \times 320$
Up Conv4b	2×2		$64 \times 64 \times 320$
Concat	Conv4b,Conv3b		$64 \times 64 \times 480$
Conv5a	$3 \times 3 \times 160$	ReLu	$64 \times 64 \times 160$
Conv5b	$3 \times 3 \times 160$	ReLu	$64 \times 64 \times 160$
Up Conv5b	2×2		$128 \times 128 \times 160$
Concat	Conv5b,Conv2b		$128 \times 128 \times 240$
Conv6a	$3 \times 3 \times 80$	ReLu	$128 \times 128 \times 80$
Conv6b	$3 \times 3 \times 80$	ReLu	$128 \times 128 \times 80$
Up Conv6b	2×2		$256 \times 256 \times 80$
Concat	Conv6b,Conv1b		$256 \times 256 \times 112$
Conv6a	$3 \times 3 \times 32$	ReLu	$256 \times 256 \times 32$
Conv6b	$3 \times 3 \times 32$	ReLu	$256 \times 256 \times 32$
Conv7	$3 \times 3 \times 3$		$256 \times 256 \times 2$

into a single $64 \times 64 \times 64$ image, which will serve as the input to classifiers, which assign a label to the image (cancer or not cancer).

A 3D CNN is used as linear classifier. It uses weighted softmax cross entropy loss (weight for a label is the inverse of the frequency of the label in the training set) and Adam Optimizer, and the CNNs use ReLU activation and dropout after each convolutional layer during training. The network is shrunk to prevent parameter overload for the relatively small Kaggle dataset. The 3D CNN architecture is described in detail in Table III.

Convolutional neural network consists of some number of convolutional layers, followed by one or more fully connected layers and finally an output layer. An example of this architecture is illustrated in Fig. 8.

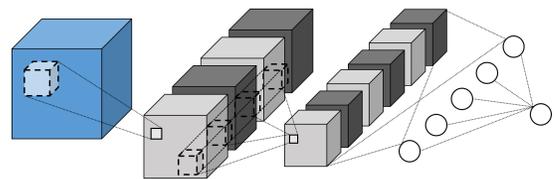


Figure 8: An example architecture of a 3D convolutional neural network used here. On the left is the input 3D volume, followed by two convolutional layers, a fully connected layers and an output layer. In the convolutional layers, each filter (or channel) is represented by a volume.

Formally, we denote the input to layer m of the network by $I^{(m)}$. The input to a 3D convolutional layer m of a neural network is a $n_1^{(m-1)} \times n_2^{(m-1)} \times n_3^{(m-1)}$ 3D object with $n_c^{(m-1)}$

so $I^{(m-1)} \in (\mathbb{R}^{n_1^{(m-1)} \times n_2^{(m-1)} \times n_3^{(m-1)}})$ and its elements are denoted by $I_{i,j,k}^{(m,\ell)}$ where i, j , and k index the 3D volume and ℓ selects the channel. The output of a convolutional layer m is defined by its dimensions, i.e., $n_1^{(m)} \times n_2^{(m)} \times n_3^{(m)}$ as well as the number of filters or channels it produces $n_c^{(m)}$. The output of layer m is a convolution of its input with a filter and is computed as

$$I_{i,j,k}^{(m,\ell)} = f_{\tanh}(b^{(m,\ell)} + \sum_{\tilde{i}, \tilde{j}, \tilde{k}, \tilde{\ell}} I_{\tilde{i}, \tilde{j}, \tilde{k}}^{(m-1, \tilde{\ell})} W_{i-\tilde{i}, j-\tilde{j}, k-\tilde{k}, \ell-\tilde{\ell}}^{(m,\ell)}) \quad (1)$$

where, $W^{(m,\ell)}$ and $b^{(m,\ell)}$ are the parameters which define the ℓ th filter in layer m . The locations where the filters are evaluated (i.e., the values of i, j, k for which $I_{i,j,k}^{(m,\ell)}$ is computed) and the size of the filters (i.e., the values of $W^{(m,\ell)}$ which are non-zero) are parameters of the network architecture. Finally, we use a hyperbolic tangent activation function with $f_{\tanh}(a) = \tanh(a)$.

Convolutional layers preserve the spatial structure of the inputs, and as more layers are used, build up more and more complex representations of the input. The output of the convolutional layers is then used as input to a fully connected network layer. To do this, the spatial and channel structure is ignored and the output of the convolutional layer is treated as a single vector. The output of a fully connected is a 1D vector $I^{(m)}$ whose dimension is a parameter of the network architecture. The output of neuron i in layer m is given by

$$I_i^{(m)} = f_{ReLU} \left(b^{(m,i)} + \sum_j I_j^{(m-1)} W_j^{(m,i)} \right) \quad (2)$$

where, $W^{(m,i)}$ and $b^{(m,i)}$ are the parameters of neuron i in layer m and the sum over j is a sum over all dimensions of the input. The activation function $f_{ReLU}(\cdot)$ here is chosen to be a Rectified Linear Unit (ReLU) with $f_{ReLU}(a) = \max(0, a)$. This activation function has been widely used in a number of domains [19], [20] and is believed to be particularly helpful in classification tasks as the sparsity it induces in the outputs helps create separation between classes during learning.

The last fully connected layer is used as input to the output layer. The structure and form of the output layer depends on the particular task. Here we consider two different types of output functions. In classification problems with K classes, a common output function is the softmax function:

$$f_i = \frac{\exp(I_i^{(o)})}{\sum_j \exp(I_j^{(o)})} \quad (3)$$

$$I_i^{(o)} = b^{(o,i)} + \sum_{k=1}^K W_k^{(o,i)} I_k^{(N)} \quad (4)$$

where, N is the index of the last fully connected layer, $b^{(o,i)}$ and $W^{(o,i)}$ are the parameters of the i th output unit and $f_i \in [0, 1]$ is the output for class i which can be interpreted as the probability of that class given the inputs. We also consider a variation on the logistic output function:

$$f = a + (b - a) \left(1 + \exp(b^{(o)} + \sum_j W_j^{(o)} I_j^{(N)}) \right)^{-1} \quad (5)$$

which provides a continuous output f which is restricted to lie in the range (a, b) with parameters $b^{(o)}$ and $W^{(o)}$. We call this the *scaled logistic* output function. We note that when considering a ranking-type multi-class classification problem like predicting the malignancy level this output function might be expected to perform better.

Table III: 3D CNN Architecture (Dropout with 0.2, Adam Optimizer, Learning Rate = 0.0001)

Layer	Params	Activation	Output
Input			$28 \times 28 \times 28$
Conv1	$5 \times 5 \times 5$	ReLU	$28 \times 28 \times 28 \times 7$
Max Pool	$1 \times 1 \times 1$, stride $2 \times 2 \times 4$		$14 \times 14 \times 7 \times 7$
Conv2	$5 \times 5 \times 3$	ReLU	$14 \times 14 \times 7 \times 17$
Max Pool	$2 \times 2 \times 2$, stride $1 \times 1 \times 0$		$6 \times 6 \times 3 \times 17$
Dense			256
Dense			2

A. Training

Given a collection of data and a network architecture, the main goal is to fit the parameters of the network to that data. To do this we will define an objective function and use gradient based optimization to search for the network parameters which minimize the objective function. Let $D = n_i, y_{i=1}^D$ be the set of D (potentially augmented) training examples where n is an input (a portion of a CT scan) and y is the output (the malignancy level or a binary class indicating benign or malignant) and Θ denote the collection of all weights W and biases b for all layers of the network. The objective function has the form

$$E(\Theta) = \sum_{i=1}^D L(y_i, f(n_i, \Theta)) + \lambda E_{prior}(\Theta) \quad (6)$$

where, $f(n_i, \Theta)$ is the output of the network evaluated on input n with parameters Θ , $L(y_i, f(n_i, \Theta))$ is a loss function which penalizes differences between the desired output of the network y and the prediction of the network \hat{y} . The function $E_{prior}(\Theta) = \|W\|^2$ is a weight decay prior which helps prevent over-fitting by penalizing the norm of the weights and λ controls the strength of the prior.

We consider two different objective functions in this paper depending on the choice of output function. For the softmax output function we use the standard cross-entropy loss function $L(y_i, \hat{y}) = -\sum_{k=1}^K y_k \log(\hat{y}_k)$ where y is assumed to be a binary indicator vector and \hat{y} is assumed to be a vector of probabilities for each of the K classes. A limitation of a cross-entropy loss is that all class errors are considered equal, hence mislabeling a malignancy level 1 as a level 2 is considered just as bad as mislabeling it a 5. This is clearly problematic, hence for the scaled logistic function we use the squared error loss function to capture this. Formally, $L(y_i, \hat{y}) = (y - \hat{y})^2$ where we assume y and \hat{y} to be real valued.

Given the objective function $E(\Theta)$, the parameters Θ are learned using stochastic gradient descent (SGD) [21]. SGD

operates by randomly selecting a subset of training examples and updating the values of the parameters using the gradient of the objective function evaluated on the selected examples. To accelerate progress and reduce noise due to the random sampling of training examples we use a variant of SGD with momentum [22]. Specifically, at iteration t , the parameters are updated as

$$\Theta_{t+1} = \Theta_t + \Delta\Theta_{t+1} \quad (7)$$

$$\Delta\Theta_{t+1} = \rho\Delta\Theta_t - \epsilon\nabla E_t(\Theta_t) \quad (8)$$

where, $\rho = 0.9$ is the momentum parameter, $\Delta\Theta_{t+1}$ is the momentum vector, ϵ_t is the learning rate and $\nabla E_t(\Theta_t)$ is the gradient of the objective function evaluated using only the training examples selected at iteration t . At iteration 0, all biases are set to 0 and the values of the filters and weights are initialized by uniformly sampling from the interval $[-\sqrt{\frac{6}{f_{an_in} + f_{an_out}}}, \sqrt{\frac{6}{f_{an_in} + f_{an_out}}}]$ as suggested by [23] where f_{an_in} and f_{an_out} respectively denote the number of nodes in the previous hidden layer and in the current layer. Given this initialization and setting $\epsilon_t = 0.01$, SGD is running for 2000 epochs, during which ϵ_t is decreased by 10% every 25 epochs to ensure convergence.

VII. SIMULATION RESULTS

The experiments are conducted using DSB dataset. In this dataset, a thousand low-dose CT images from high-risk patients in DICOM format is given. The DSB database consists of 1397 CT scans and 248580 slices. Each scan contains a series with multiple axial slices of the chest cavity. Each scan has a variable number of 2D slices (Fig. 9), which can vary based on the machine taking the scan and patient. The DICOM files have a header that contains the necessary information about the patient id, as well as scan parameters such as the slice thickness. It is publicly available in the Kaggle [13]. Dicom is the de-facto file standard in medical imaging. This pixel size/coarseness of the scan differs from scan to scan (e.g. the distance between slices may differ), which can hurt performance of our model.

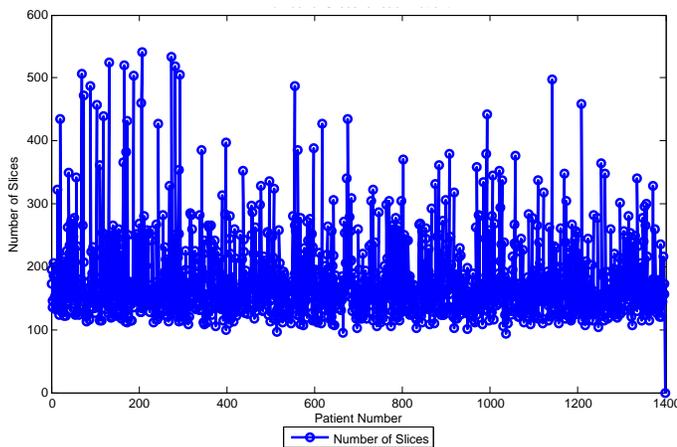


Figure 9: Number of slices per patient in data science bowl dataset.

The experiments are implemented on computer with CPU i7, 2.6 GHz, 16 RAM, Matlab 2013b, R-Studio, and Python. Initially speaking, the nodules in DSB dataset are detected and segmented using thresholding and U-Net Convolutional Neural Network. The diameters of the nodules range from 3 to 30 mm. Each slice has 512×512 pixels and 4096 gray level values in Hounsfield Unit (HU), which is a measure of radiodensity.

In the screening setting, one of the most difficult decisions is whether CT or another investigation is needed before the next annual low-dose CT study. Current clinical guidelines are complex and vary according to the size and appearance of the nodule. The majority of nodules were solid in appearance. For pulmonary nodule detection using CT imaging, CNNs have recently been used as a feature extractor within a larger CAD system.

For simplicity in training and testing we selected the ratings of a single radiologist. All experiments were done using 50% training set, 20% validation set and 30% testing set. To evaluate the results we considered a variety of testing metrics. The accuracy metric is the used metric in our evaluations. In our first set of experiments we considered a range of CNN architectures for the binary classification task. Early experimentation suggested that the number of filters and neurons per layer were less significant than the number of layers. Thus, to simplify analysis the first convolutional layer used seven filters with size $5 \times 5 \times 5$, the second convolutional layer used 17 filters with $5 \times 5 \times 3$ and all fully connected layers used 256 neurons. These were found to generally perform well and we considered the impact of one or two convolutional layers followed by one or two fully connected layers. The networks were trained as described above and the results of these experiments can be found in Table I. Our results suggest that two convolutional layers followed by a single hidden layer is one of the optimal network architecture for this dataset. The average error for training is described in Fig. 10.

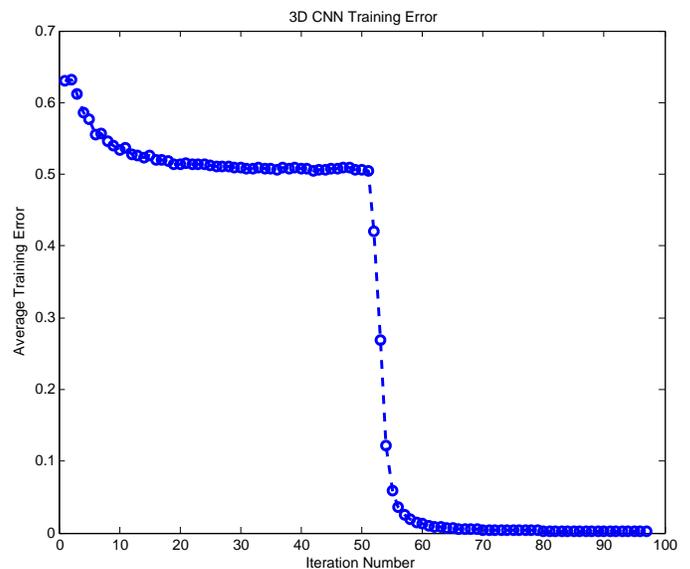


Figure 10: Average training error in 3D CNN.

Another important parameter in the training of neural networks is the number of observations that are sampled

at each iteration, the size of the so-called minibatch. The use of minibatches is often driven in part by computational considerations but can impact the ability of SGD to find a good solution. Indeed, we found that choosing the proper minibatch size was critical for learning to be effective. We tried minibatches of size 1, 10, 50 and 100. While the nature of SGD suggests that larger batch sizes should produce better gradient estimates and therefore work better, our results here show that the opposite is true. Smaller batch sizes, even as small as 1, produce the best results. We suspect that the added noise of smaller batch sizes allows SGD to better escape poor local optima and thus perform better overall.

The recognition results are shown by confusion matrix achieved on the DSB dataset with 3D CNN as shown in Table IV. As shown from the Table IV, Accuracy of model is 86.6%, Mis-classification rate is 13.4%, False positive rate is 11.9%, and False Negative is 14.7%. Almost all patients are classified correctly. Additionally, there is an enhancement on accuracy due to efficient U-Net architecture and segmentation.

Table IV: Confusion Matrix of 3D CNN using 30% Testing

Actual	Predicted	
	Abnormal	Normal
Abnormal	0.853	0.147
Normal	0.119	0.881

VIII. CONCLUSION

In this paper we developed a deep convolutional neural network (CNN) architecture to detect nodules in patients of lung cancer and detect the interest points using U-Net architecture. This step is a preprocessing step for 3D CNN. The deep 3D CNN models performed the best on the test set. While we achieve state-of-the-art performance AUC of 0.83, we perform well considering that we use less labeled data than most state-of-the-art CAD systems. As an interesting observation, the first layer is a preprocessing layer for segmentation using different techniques. Threshold, Watershed, and U-Net are used to identify the nodules of patients.

The network can be trained end-to-end from raw image patches. Its main requirement is the availability of training database, but otherwise no assumptions are made about the objects of interest or underlying image modality.

In the future, it could be possible to extend our current model to not only determine whether or not the patient has cancer, but also determine the exact location of the cancerous nodules. The most immediate future work is to use Watershed segmentation as the initial lung segmentation. Other opportunities for improvement include making the network deeper, and more extensive hyper parameter tuning. Also, we saved our model parameters at best accuracy, but perhaps we could have saved at other metrics, such as F1. Other future work include extending our models to 3D images for other cancers. The advantage of not requiring too much labeled data specific to our cancer is it could make it generalizable to other cancers.

REFERENCES

- [1] W.-J. Choi and T.-S. Choi, "Automated pulmonary nodule detection system in computed tomography images: A hierarchical block classification approach," *Entropy*, vol. 15, no. 2, pp. 507–523, 2013.
- [2] A. Chon, N. Balachandar, and P. Lu, "Deep convolutional neural networks for lung cancer detection," tech. rep., Stanford University, 2017.
- [3] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision.," in *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 253–256, IEEE, 2010.
- [4] K. Alex, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25 (NIPS 2012)* (F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds.), pp. 1097–1105, 2012.
- [5] H. Suk, S. Lee, and D. Shen, "Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis," *NeuroImage*, vol. 101, pp. 569–582, 2014.
- [6] G. Wu, M. Kim, Q. Wang, Y. Gao, S. Liao, and D. Shen, "Unsupervised deep feature learning for deformable registration of mr brain images.," *Medical Image Computing and Computer-Assisted Intervention*, vol. 16, no. Pt 2, pp. 649–656, 2013.
- [7] Y. Xu, T. Mo, Q. Feng, P. Zhong, M. Lai, and E. I. Chang, "Deep learning of feature representation with multiple instance learning for medical image analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pp. 1626–1630, 2014.
- [8] D. Kumar, A. Wong, and D. A. Clausi, "Lung nodule classification using deep features in ct images," in *2015 12th Conference on Computer and Robot Vision*, pp. 133–138, June 2015.
- [9] Y. Bar, I. Diamant, L. Wolf, S. Lieberman, E. Konen, and H. Greenspan, "Chest pathology detection using deep learning with non-medical training," *Proceedings - International Symposium on Biomedical Imaging*, vol. 2015-July, pp. 294–297, 2015.
- [10] W. Sun, B. Zheng, and W. Qian, "Computer aided lung cancer diagnosis with deep learning algorithms," in *SPIE Medical Imaging*, vol. 9785, pp. 97850Z–97850Z, International Society for Optics and Photonics, 2016.
- [11] J. Tan, Y. Huo, Z. Liang, and L. Li, "A comparison study on the effect of false positive reduction in deep learning based detection for juxtapleural lung nodules: Cnn vs dnn," in *Proceedings of the Symposium on Modeling and Simulation in Medicine, MSM '17*, (San Diego, CA, USA), pp. 8:1–8:8, Society for Computer Simulation International, 2017.
- [12] R. Golan, C. Jacob, and J. Denzinger, "Lung nodule detection in ct images using deep convolutional neural networks," in *2016 International Joint Conference on Neural Networks (IJCNN)*, pp. 243–250, July 2016.
- [13] Kaggle, "Data science bowl 2017." <https://www.kaggle.com/c/data-science-bowl-2017/data>, 2017.
- [14] LUNA16, "Lung nodule analysis 2016." <https://luna16.grand-challenge.org/>, 2017.
- [15] M. Firmino, A. Morais, R. Mendoa, M. Dantas, H. Hekis, and R. Valentim, "Computer-aided detection system for lung cancer in computed tomography scans: Review and future prospects," *BioMedical Engineering OnLine*, vol. 13, p. 41, 2014.
- [16] S. Hawkins, H. Wang, Y. Liu, A. Garcia, O. Stringfield, H. Krewer, Q. Li, D. Cherezov, R. A. Gatenby, Y. Balagurunathan, D. Goldgof, M. B. Schabath, L. Hall, and R. J. Gillies, "Predicting malignant nodules from screening ct scans," *Journal of Thoracic Oncology*, vol. 11, no. 12, pp. 2120–2128, 2016.
- [17] M. S. AL-TARAWNEH, "Lung cancer detection using image processing techniques," *Leonardo Electronic Journal of Practices and Technologies*, pp. 147–158, June 2012.
- [18] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015.
- [19] M. D. Zeiler, M. Ranzato, R. Monga, M. Mao, K. Yang, Q. V. Le, P. Nguyen, A. Senior, V. Vanhoucke, J. Dean, and G. E. Hinton, "On rectified linear units for speech processing," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3517–3521, May 2013.

- [20] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML*, vol. 30, 2013.
- [21] L. Bottou, *Large-Scale Machine Learning with Stochastic Gradient Descent*, pp. 177–186. August 2010.
- [22] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proceedings of the 30th International Conference on International Conference on Machine Learning*, ICML'13, pp. 1139–1147, JMLR.org, 2013.
- [23] H. Han, L. Li, H. Wang, H. Zhang, W. Moore, and Z. Liang, "A novel computer-aided detection system for pulmonary nodule identification in ct images," in *Proceedings of SPIE Medical Imaging Conferenc*, vol. 9035, 2014.

Multiple Vehicles Semi-Self-driving System Using GNSS Coordinate Tracking under Relative Position with Correction Algorithm

Heejin Lee, Hiroshi Suzuki, Takahiro Kitajima, Akinobu Kuwahara and Takashi Yasuno
Graduate School of Tokushima University,
2-1 Minamijosanjima, Tokushima, 770-8506, Japan

Abstract—This paper describes a simple and low-cost semi-self-driving system which is constructed without cameras or image processing. In addition, a position correction method is presented by using a vehicle dynamics. Conventionally, self-driving vehicle is operated by various expensive environmental recognition sensors. It results in rise in prices of the vehicle, and also the complicated system with various sensors tends to be a high possibility of malfunction. Therefore, we propose the semi-self-driving system with a single type of global navigation satellite system (GNSS) receiver and a digital compass, which is based on a concept of a preceding vehicle controlled by a human manually and following vehicles which track to the preceding vehicle automatically. Each vehicle corrects coordinate using current velocity and heading angle from sensors. Several experimental and simulation results using our developed small-scale vehicles demonstrate the validity of the proposed system and correction method.

Keywords—Self-driving, positioning; global navigation satellite system (GNSS); Global Positioning System (GPS); GLONASS

I. INTRODUCTION

Recently, the self-driving technology has been developed rapidly in Intelligent Technology (IT) and automotive industries. It is well known that Google shows remarkable performances of the self-driving system which is configured by cameras, radars, the Global Navigation Satellite System (GNSS) device and the Google digital map. Generally, the self-driving vehicle operates based on the image processing information using cameras, radars and GNSS for positioning and obstacle detection [1], [2]. However, it is difficult to use in situations of bad weather and the camera covered by obstacles. In addition, the use of the camera and the radar cause to increase the vehicle producing cost [3], [4].

If the GNSS technology, such as Global Positioning System (GPS, USA), GLONASS (RUSSIA) and GALILEO (EU), is used for measuring a correct position of the vehicle, there is no need to measure the vehicle position by using the camera. The GPS is most useful system for the present. The GPS system supplies two service, Precise Positioning Service (PPS) and the Standard Positioning Service (SPS) [5]. The PPS provides reliable positioning which enough to implement the self-driving system [6]. However, the PPS could be used only for military or authorized user. The SPS is operated by only the GPS receiver independently, the position is calculated by receiving information from the satellite. The SPS could be used

without authorization. however, the positioning reliability of the SPS mode is not enough for practical use.

Fig. 1 shows an example of GNSS positioning reliability indicating method in two-dimension. The GNSS positioning reliability is quantified to accuracy and precision [7]. The accuracy means that how close to the absolute coordinate from each of measured coordinate. The error of each positioning is quantified to accuracy which indicated in distance unit. The precision means that how concentrated each measured coordinates, and is indicated in the distribution of each measured coordinates. In this paper, the accuracy has considered significantly for positioning reliability evaluation and position error indicating.

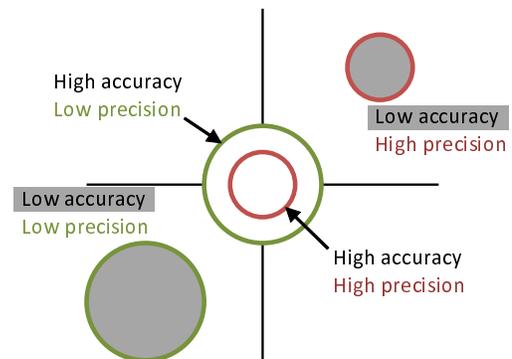


Fig. 1. Accuracy and the precision in two dimension.

Approximately 95% of position errors as accuracy are occurred lower than 3.351 m in the SPS mode of the GPS [8], [9]. Therefore, The self-driving using the SPS is unable practically. On the other hand, an error correction algorithm which is possible to increase a positioning accuracy in about several centimeters is applied under the Differential Global Navigation Satellite System (DGNSS) mode in the SPS mode. In the DGNSS mode, two types could be applied for position correction. One of the type is Real Time Kinematic (RTK) and the other is Satellite Based Augmentation System (SBAS). In the RTK type, Radio Technical Commission for Maritime Service (RTCM) correction data is transmitted to GNSS receiver from a base station [10]. Therefore, in order to use the RTK type correction, it is necessary to install the

base station in advance, and its available range is limited by the base station performances. Thus, the RTK type is impossible to use under the general road environment and also difficult to use for the self-driving vehicle. In the SBAS type, correction data is transmitted to GNSS receiver from the SBAS satellite periodically. Generally, position error is several meters under the SBAS correction. Wide Area Augmentation System (WAAS) of USA or MTSAT Satellite Augmentation System (MSAS) of JAPAN is the famous for the SBAS [11]-[13]. The SBAS is operated with only SBAS compatible receiver which is not required the base station. Therefore, although the position accuracy measured by the SPS or the SBAS is not enough for practical use, the SPS or the SBAS mode is available for the self-driving vehicle using the general road environment [14].

It is well known that the position errors occur a satellite orbit error, an atomic clock error, an ionosphere passage, a troposphere passage, a multi-path, due to the arrangement of satellite geometric error. The largest component of reason in the position error is the time delay by the ionospheric passage [15]. Fig. 2 shows correlation between refraction and propagation delay on the ionospheric passage.

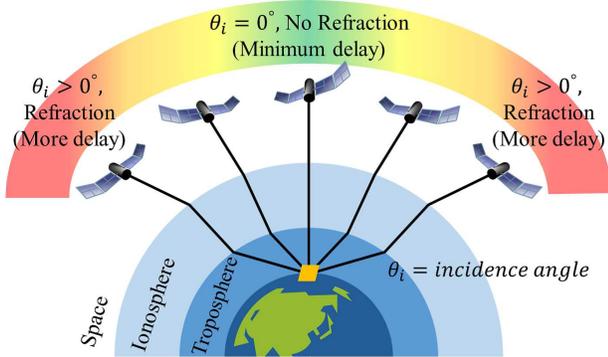


Fig. 2. Propagation delay by refraction on the ionospheric passage.

Radio wave refraction occurs when the radio wave passes through the ionosphere and the path of radio wave become longer. However, if two SPS mode GNSS receivers are located closely each other, the trends of each position error are similar. The reason why there is a greater possibility that two SPS mode GNSS receivers receive the signal from the same satellites located near in 10 km approximately [16]. That is, it causes a similar time delay, since signals received by two SPS mode GNSS receivers are passing through the similar path of ionosphere passage. Therefore, position errors have similar trends between two receivers [17].

In this paper, we propose a simple and low-cost semi-self-driving system which is constructed without cameras or image processing on the basis of a new concept using multiple vehicles that one is a preceding vehicle and the others are following vehicles. Relative positions of two vehicles are used for the proposed semi-self-driving system. The preceding vehicle controlled by the driver always transmits the position information to the following vehicle. The following vehicle is controlled automatically with tracking to the preceding vehicle. In addition, in order to improve stability and reliability

of positioning, vehicle dynamics based position correction algorithm is applied to the system.

In following sections, system algorithm, implementation and evaluation method are described in detail. Several experimental and simulation results using developed small-scale vehicles demonstrate the validity of the proposed system and correction method.

II. DEVELOPED SEMI-SELF-DRIVING SYSTEM

Fig. 3 shows the concept of the developed semi-self-driving system. The system is configured by the preceding and following vehicles. As mentioned above, the SPS or SBAS of GNSS position errors of vehicles are similar when vehicles are located very closely. Therefore, it is possible to ensure reliable accuracy by using the relative position of vehicles. The preceding vehicle transmits own position, heading angle and speed to the first following vehicle. The first following vehicle drives with tracking a coordinate of the preceding vehicle, and the second following vehicle drives with tracking a coordinate of the first following vehicle. The preceding vehicle is controlled by the driver manually. Each vehicle updates own position periodically.

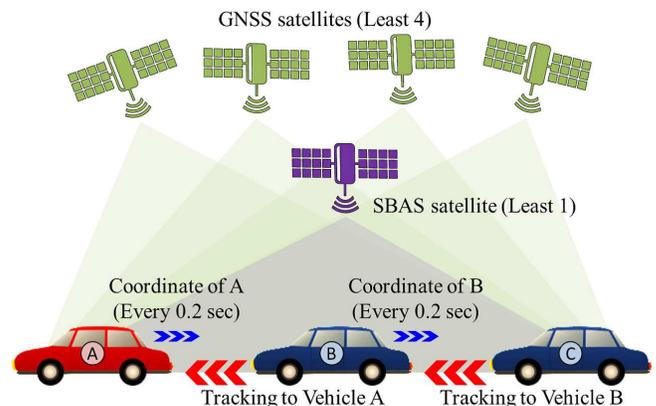


Fig. 3. Concept of developed semi-self-driving system.

A. Grouping Algorithm

The preceding vehicle makes a group driving profile, and broadcasts profile to all of the following vehicles in the group. The profile is composed of number of vehicles, serial number of vehicle, sequence, limit velocity, minimum safety distance. The following vehicle checks the profile, and makes a decision to accept or reject of joining in the group. When the following vehicle leaves the group, the following vehicle transmits a request message to preceding vehicle. The preceding vehicle makes the group profile again without requested following vehicle. When the preceding vehicle leaves the group, the first following vehicle be a new preceding vehicle.

B. Coordination and Tracking Algorithm

Fig. 4 shows a process flow of the proposed system for the preceding vehicle. Time and position of the preceding vehicle is received from the GNSS receiver module in the DGNSS mode which is the SBAS type. The GNSS receiver

output several messages of the National Marine Electronics Association (NMEA) 0183 format. Time, position and fix related data of the receiver, the GNGGA message is used for calculation in the algorithm. On the other hand, current heading angle is measured by the digital compass module. The heading angle is corrected -7.3 degrees due to influence of local magnetic field in the Tokushima prefecture, JAPAN [18]. Coordinate from the GNSS receiver is converted to the degree format. Vehicle velocity and throttle is checked by the system. All status of the preceding vehicle is transmitted to the following vehicle and monitoring server. The system updates position and control command every 0.2 second periodically.

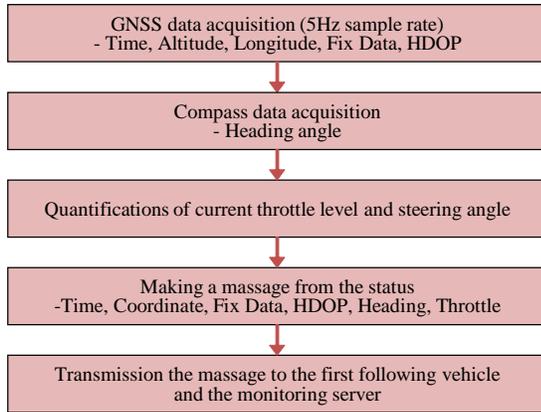


Fig. 4. Process flow of the preceding vehicle.

Fig. 5 shows a process flow of the proposed system for the following vehicle. The current time and position of the following vehicle are received from the GNSS receiver module in the DGNS mode which is the SBAS type. Simultaneously, The status message of the preceding vehicle is received via a wireless data networking device. In the same way as the preceding vehicle, current heading angle is calculated by the digital compass module. Relative distance and angle are calculated on the bases of each position of two vehicles. The relative distances is calculated by (1).

$$s = \sqrt{\left\{ \left(\frac{2\pi r}{360} x_{p-f} \right) \cdot \cos \left(\frac{\pi}{180} y_f \right) \right\}^2 + \left(\frac{2\pi r}{360} y_{p-f} \right)^2} \quad (1)$$

where, s is the relative distance which can be calculated from difference in longitude x_{p-f} and latitude y_{p-f} . y_f is the latitude of the following vehicle. r is the radius of the earth as approximately 6378.137 km. α denotes the relative angle which can be calculated by (2).

$$\alpha = \left(\frac{180}{\pi} \right) \cdot \tan^{-1} \left\{ \frac{y_{p-f}}{x_{p-f} \cdot \cos \left(\frac{\pi}{180} y_f \right)} \right\} \quad (2)$$

If the relative distance is longer than the target distance, system increases velocity by throttle control. Otherwise, system reduces velocity until reaching at the target distance. The steering angle is controlled by comparing target angle with current angle. Update period for the position is set to 0.2 second. Calculation period for direction angle and vehicle velocity is set to 0.01 second approximately.

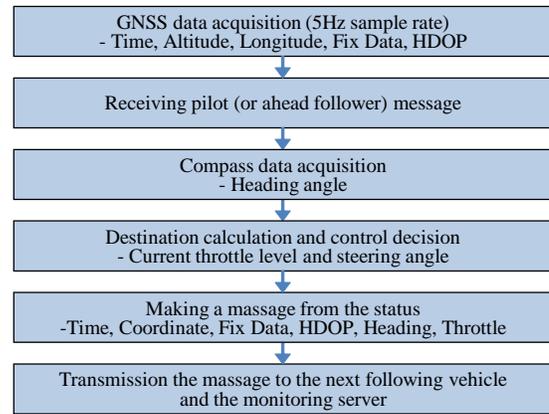


Fig. 5. Process flow of the following vehicle.

C. Driving Termination Algorithm

The group driving could be terminated by the GNSS receiver fault, relative distance, relative heading angle. If the GNSS operation mode in DGNS SBAS, positioning data is available, and used to calculation for driving. On the other hand, in the case of the GNSS operation mode in 2D or below, the positioning data is not available. The system converts current latitude to 99 degree and current longitude to 199 degree for processing coordinate to invalid. Therefore, the driving is terminated by invalid coordinate. If the relative distance is over 110 m, the system terminates driving. The system limits maximum velocity to 200 km/h. In this case, the safety relative distance between preceding vehicle and following vehicle is approximately 110 m [19], [20]. Thus, considering the safety distance and many type of vehicle, the maximum relative distance is limited to approximately 200 m. If the relative heading angle is over 90 degrees, system terminates the driving. The reason why, the vehicle drives in reverse or the reverse direction with high probability.

III. POSITION CORRECTION METHOD

In order to improve positioning reliability and stability, each of the position is corrected by the proposed position correction algorithm which is based on vehicle dynamics [21]. Current velocity and heading angle are measured by velocity and heading angle feedback of the system. Thus, next coordinate prediction is possible by measured absolute values and current coordinate. When vehicle driving, the system compares the GNSS value and system feedback value. Therefore, the impossible movement is corrected by the algorithm. When vehicle stopping, the correction algorithm averages coordinates, and current coordinate is changed to the averaged coordinate that in order to stabilize positioning. The correction algorithm is simulated that in order to evaluate usability.

A. Correction Algorithm in Driving

When the vehicle driving, the system calculates current velocity and heading angle from the system status feedback. Simultaneously, current velocity and heading angle are calculated by trajectory which is recorded by the GNSS. Subsequently, the system compares the calculated values by the GNSS with the system feedback. If the difference of heading angle is over 17.5 degrees, coordinate is corrected by correction (3).

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} x_p \\ y_p \end{bmatrix} + k \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} x_{c-p} \\ y_{c-p} \end{bmatrix} \quad (3)$$

The reference angle is determined by the turning characteristic of vehicle dynamics. In addition, distance of the trajectory is corrected by velocity comparison. Where, x and y are the longitude and the latitude. x' and y' are the corrected coordinate which can be calculated by the coordinate rotation equation in triangular function. x_p and y_p are the latest coordinate which excepts current measured coordinate by the GNSS. x_{c-p} and y_{c-p} are the longitude and latitude variations between current and past coordination by the GNSS without correction. θ denotes the difference angle. k denotes the velocity rate which is divided value of absolute velocity with calculated velocity by the GNSS.

B. Correction Algorithm in Stopping

When the vehicle in stopping, absolute coordinate never be changed. Thus, the system averages 100 coordinates for position stabilization during 20 seconds with 0.2 second sampling rate. The reason why, conventional waiting time for traffic signal is least 20 seconds in normal traffic [22], [23]. Furthermore, the coordinate is changed maximum 10 cm in one sampling of the GNSS receiver module which is implemented in the system. In that case, if averaging over 100 samples, coordinate is changed less than 1 mm at the last averaging. Therefore, the correction algorithm averages latest 100 coordinates, and current coordinate is changed to the averaged coordinate. If the vehicle stopping under 20 seconds, the algorithm averages every coordinates until vehicle started to moving.

IV. EXPERIMENTAL SETUP

A. Constructed Small Scaled Vehicles

Fig. 6 shows appearance of constructed small-scale vehicles implemented the proposed semi-self-driving system. In the experiment, the vehicle A is used for preceding vehicle, B and C are used for following vehicle. The vehicle is constructed by a main circuit, a GNSS antenna, a GNSS receiver module, a digital compass, servo motors for the steering wheel and throttle control, a DC motor for driving wheels, a DC motor driver and a 7.2V Ni-MH battery. The main circuit is constructed by microprocessor, wireless communication module and power supply. Configuration of the preceding vehicle and following vehicles is the same.

Fig. 7 shows a hardware configuration. The GNSS active antenna (SAN JOSE DS-28) is installed on the aluminum roof of the body. The GNSS receiver module (u-blox CAM-M8) based on u-blox M8 chipset is performed with 5 Hz position updating cycle. The digital compass module (HONEYWELL HMC6352) has a precision performance of 0.1 degree increments. For the wireless communication module, MAX STREAM XBee Pro is installed. Communication distance of the module is approximately 1 km under the clear channel condition. The microprocessor (ATMEL Atmega64L) calculates the vehicle speed and steering angle. Position and status of the vehicles are transmitted to following vehicle and the monitoring PC simultaneously.

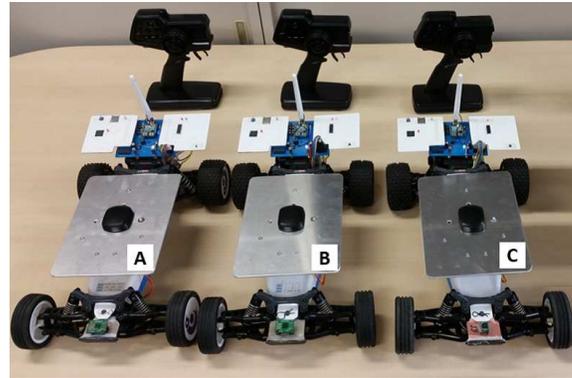


Fig. 6. Constructed small-scale vehicles.

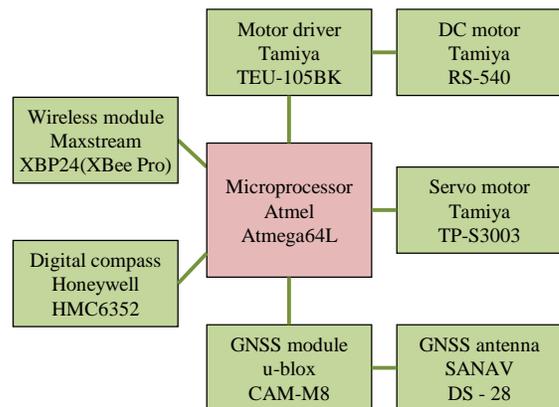


Fig. 7. Hardware configuration of the small-scale vehicle.

B. Setup for Tracking and Approaching Test

1) *Environment*: Fig. 8 shows the experimental environment where a track is constructed in the form of vertical and horizontal 20 m square, and each corner and the center position are marked with orange rubber corns. The experiment was carried out on the ground under the sunny day with open sky condition. Here, the digital compass is calibrated before the experiment.

2) *Procedure of the Tracking Test*: The preceding vehicle is manipulated to the point 1-2-3-4-1 constant velocity. The following vehicle automatically drives with tracking to the preceding vehicle continuously. Trajectory of each vehicle is recorded by the monitoring server.

3) *Procedure of the Approaching Test*: First, the first following vehicle drives with tracking to the preceding vehicle, and the second following vehicle tracks to the first following vehicle. Second, the preceding vehicle stops on the center point of the track. If the relative distance under the GNSS data is shorter than 30 cm, following vehicles are stopped immediately. At that situation, the actual relative distance is measured. An example of actual relative distance measurement is shown in Fig. 9. The reference of the measurement is center of the GNSS antenna on vehicle. Simultaneously, relative distance by the GNSS is recorded in monitoring server. The position error is calculated by these two distances. The experiment was conducted in total of ten times.

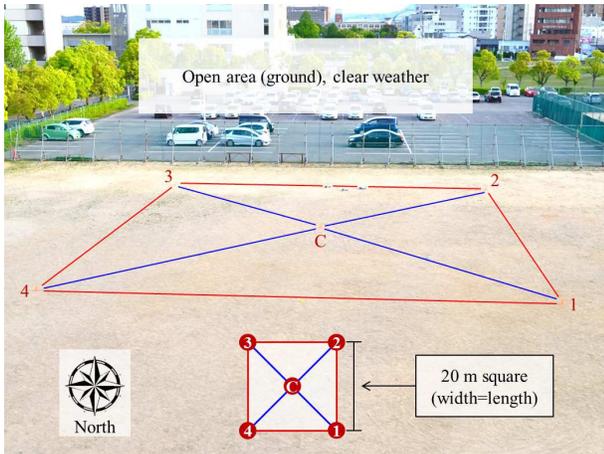


Fig. 8. Experimental environment and track setting.



Fig. 10. Triangular point in the Tokushima prefecture, Japan.

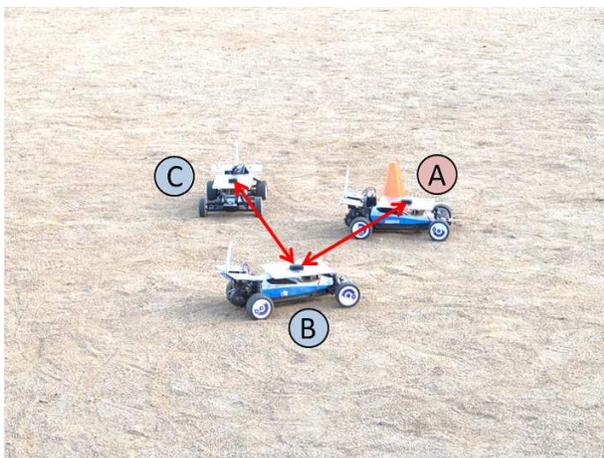


Fig. 9. Measurement of actual distance in approaching test.

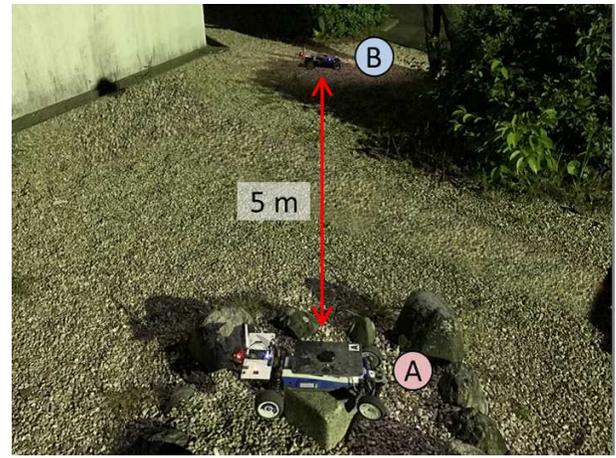


Fig. 11. Position error measurement.

C. Setup for Position Error Measurement Test

1) *Environment*: In order to setup the reference absolute position, a triangular point is used in position error measurement. The triangular point is located on the mountain Bizan in the Tokushima prefecture in JAPAN. Fig. 10 shows the triangular point. The measurement was carried out under the clear weather with open sky condition.

2) *Procedure of Position Error Measurement*: Fig. 11 shows the setup of position error measurement. The vehicle A is allocated on the center of the triangular point. The other vehicle is allocated on 5 m away point from the triangular point. Positions of both vehicles are measured for 3 hours. The position error is calculated by one dimension Root Mean Square (RMS) method [24]. It is possible to evaluate the accuracy of relative position.

V. EXPERIMENTAL RESULTS

A. Tracking Test

Fig. 12 and 13 show tracking performances for the moving preceding vehicle. The following vehicle B could track on the trajectory of the moving preceding vehicle A under the

GNSS data. However, it is noted that the absolute trajectory has positioning errors more or less than 100 cm against the GNSS data.

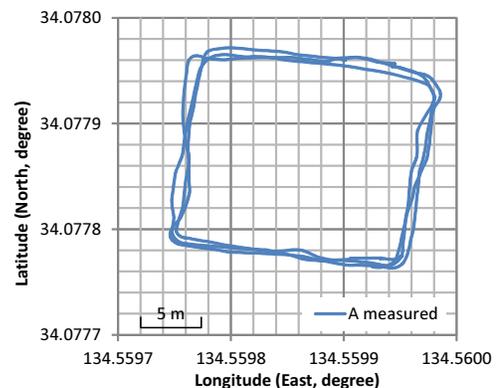


Fig. 12. Trajectory of the preceding vehicle in the tracking test.

B. Approaching Test

Position errors between the relative distance calculated by the GNSS and the actual relative distance for each test are

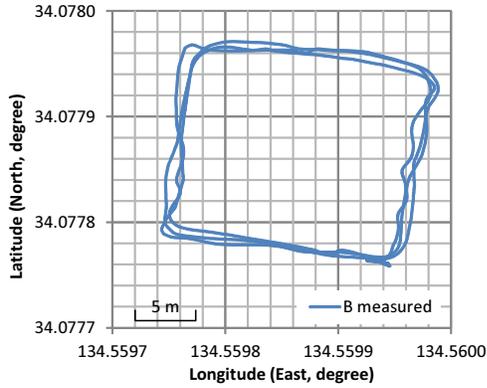


Fig. 13. Trajectory of the following vehicle in the tracking test.

listed in Table I. Maximum, minimum and average distance errors between the preceding vehicle as A and the first following vehicle as B are 54 cm, 0 cm, and 29.8 cm, respectively. Maximum, minimum and average distance errors between the first following vehicle as B and the second following vehicle as C are 54 cm, 3 cm, and 30.4 cm, respectively. Test result of two cases are similar each other. Note that the test was conducted under the GNSS receiving status which Horizontal Dilution Of Precision (HDOP) is average 0.6 approximately.

C. Position Error Measurement

Fig. 14, 15 and Table II show the measurement results of the position error. Average position error in independent measurement is approximately 160 cm. On the other hand, average position error in relative measurement is decreased to 53 cm (-68.8%). The RMS 1σ means that 68.3% position errors are occurred smaller than the 1σ value. The RMS 1σ error is approximately 215 cm in independent mode, and 98 cm (-54.4%) in relative mode. The RMS 2σ means that 95.5% position errors are occurred smaller than the 2σ value. The RMS 2σ error is approximately 266 cm in independent mode, and 139 cm (-47.7%) in relative mode. From these result, significant improvement of the GNSS position accuracy has confirmed in the relative positioning mode.

TABLE I. POSITION ERRORS OF APPROACHING TEST

Test No.	\overline{AB}	\overline{BC}	\overline{AB}	\overline{BC}	\overline{AB}	\overline{BC}	
	GNSS	GNSS	Actual	Actual	Error	Error	
1.	70 cm	104 cm	32 cm	60 cm	38 cm	44 cm	
2.	14 cm	91 cm	68 cm	88 cm	54 cm	3 cm	
3.	72 cm	30 cm	60 cm	84 cm	12 cm	54 cm	
4.	40 cm	23 cm	40 cm	76 cm	0 cm	53 cm	
5.	49 cm	40 cm	100 cm	36 cm	51 cm	4 cm	
6.	72 cm	80 cm	99 cm	105 cm	27 cm	25 cm	
7.	79 cm	85 cm	44 cm	43 cm	35 cm	42 cm	
8.	96 cm	57 cm	95 cm	42 cm	1 cm	15 cm	
9.	94 cm	18 cm	65 cm	41 cm	29 cm	23 cm	
10.	86 cm	21 cm	35 cm	62 cm	51 cm	41 cm	
					Minimum error	0 cm	3 cm
					Maximum error	54 cm	54 cm
					Average error	29.8 cm	30.4 cm
					Average HDOP	0.614	0.617

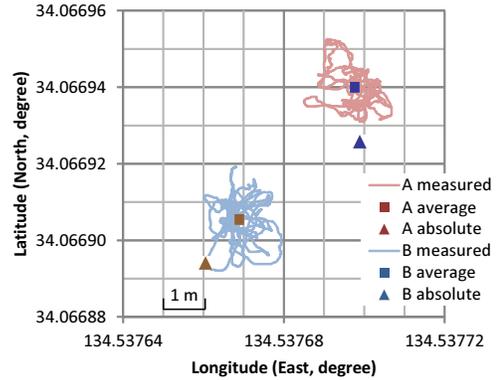


Fig. 14. Result of the position error measurement.

TABLE II. RMS ERROR OF THE MEASUREMENT

Type	Mean	RMS 1σ	RMS 2σ
Independent (A)	163 cm	224 cm	279 cm
Independent (B)	157 cm	207 cm	254 cm
Relative (A,B)	53 cm	98 cm	139 cm

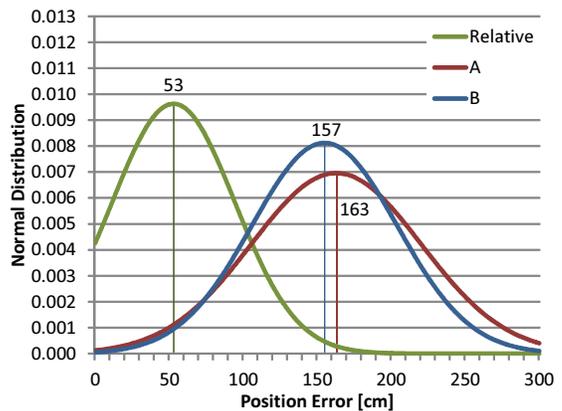


Fig. 15. Normal distribution of the measurement result.

VI. SIMULATION RESULTS OF CORRECTION ALGORITHM

A. Result of Correction for Driving

The simulation results of correction algorithm for driving are listed in Table III. Least 24.24% of coordinates are corrected by the algorithm. Correction distances are distributed 0.17 cm to 42.68 cm, and average correction distance is 14.13 cm in the one sampling.

TABLE III. CORRECTION RATE AND PERFORMANCE

	A	B	Average
Number of total coordinate	883	883	883
Number of corrected coordinate	214	261	237.5
Correction rate	24.24%	29.56%	26.90%
Minimum correction distance	0.17 cm	0.28 cm	0.23 cm
Maximum correction distance	40.64 cm	42.68 cm	41.66 cm
Average correction distance	14.59 cm	13.67 cm	14.13 cm

Fig. 16 shows the correction simulation result of vehicle A as the preceding vehicle. In the result, when the vehicle drives in curve course, the more coordinates are corrected than straight line course.

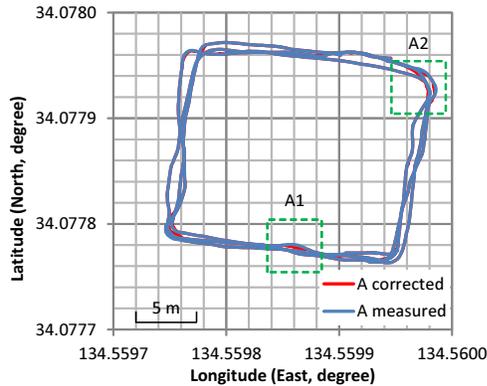


Fig. 16. Correction result of the vehicle A in driving.

Fig. 17 and 18 are show detailed view of A1 point and A2 point in Fig. 16. The side draft is corrected by the algorithm in the straight line course at the point of A1. The overshoot is corrected by the algorithm in the curve course at the point of A2.

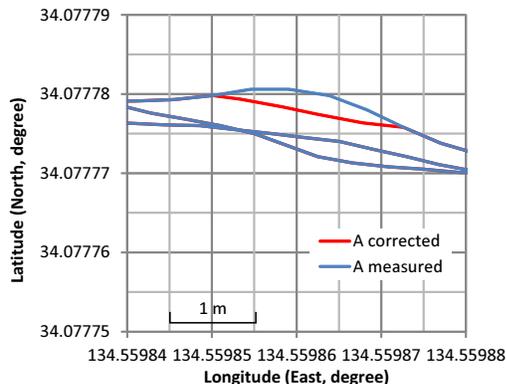


Fig. 17. Coordinate correction of vehicle A in the straight line course A1.

Fig. 19 shows the correction simulation result of vehicle B as the following vehicle. Likewise with vehicle A, when

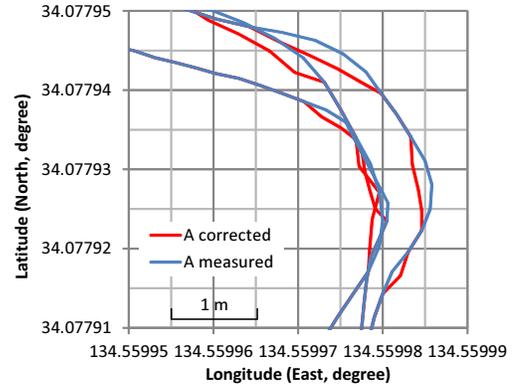


Fig. 18. Coordinate correction of vehicle A in the curve course A2.

the vehicle drives in curve course, the more coordinates are corrected than straight line course.

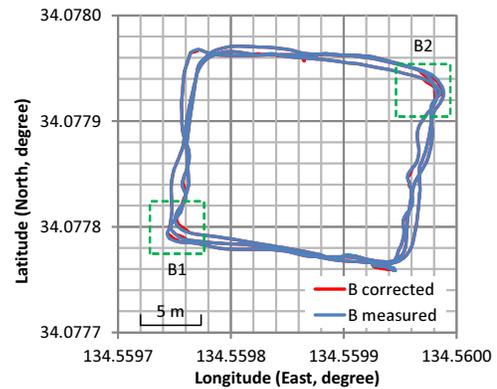


Fig. 19. Correction result of the vehicle B in driving.

Fig. 20 and 21 are show detailed view of B1 point and B2 point in Fig. 19. The overshoot is corrected by the algorithm in the curve course at the point of B1 and B2.

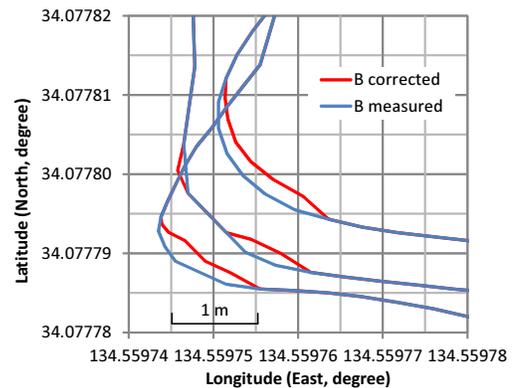


Fig. 20. Coordinate correction of vehicle B in the curve course B1.

B. Correction Result for Stop

The simulation result of correction algorithm in stopping is listed in Table IV. Mean, 1σ , 2σ values of the one dimension RMS errors are decreased to 45 cm (-15.09%), 84 cm (-14.29%), 116 cm (-16.55%) in relative mode by correction

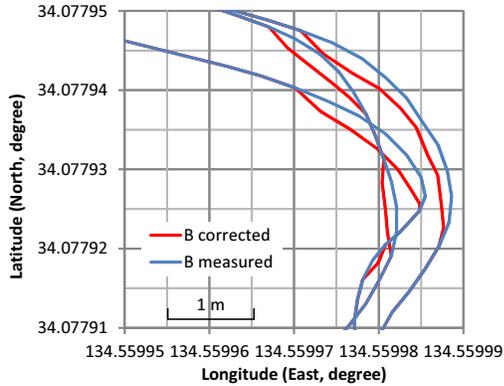


Fig. 21. Coordinate correction of vehicle B in the curve course B2.

algorithm which averages latest 100 coordinate samples in stopping. However, the mean values of the single mode are slightly decreased or not decreased. From these result, improvement of position accuracy has confirmed by correction algorithm for stopping in the relative mode.

TABLE IV. RMS ERROR (100 COORDINATES AVERAGE)

Type	100 avg. Mean	100 avg. RMS 1 σ	100 avg. RMS 2 σ
Independent (A)	163 cm (0%)	215 cm (-4.02%)	267 cm (-4.3%)
Independent (B)	153 cm (-2.55%)	196 cm (-5.31%)	233 cm (-8.27%)
Relative (A,B)	45 cm (-15.09%)	84 cm (-14.29%)	116 cm (-16.55%)

Fig. 22 shows normal distributions comparison of normal positioning and positioning with correction algorithm in stopping. The distribution curve of corrected positioning shows more narrow distribution band than normal positioning. It means that the positioning precision is improved by the correction method.

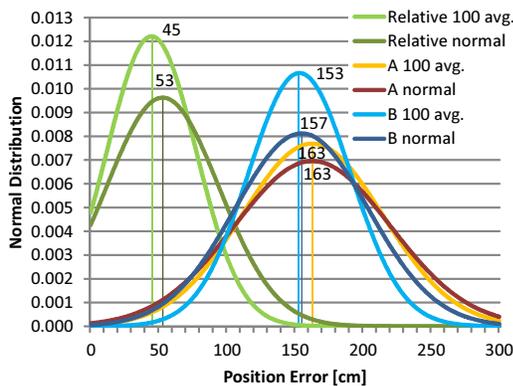


Fig. 22. Distributions of RMS error by correction algorithm in stopping.

VII. CONCLUSION

This paper proposed a semi-self-driving system with GNSS using a relative distance between preceding and following vehicles. In addition, vehicle dynamics based on correction method has applied to system. To ensure the validity of the proposed system, several experiments and simulations

were conducted. As a result, the proposed self-driving system implemented in the follower vehicle performed successfully for both approaching to several fixed destinations and tracking to moving preceding vehicle. Furthermore, according to the result of simulation, the positioning accuracy and precision has improved by correction algorithm which is based on the vehicle dynamic. Although the remaining absolute position error of the GNSS data, the measurement accuracy of the relative distance between the vehicles was improved.

In the future, we will improve the performance of position correction algorithm to realize a complete self-driving system with autonomous preceding vehicle.

REFERENCES

- [1] Gregory Kehoe, Hossein Jula, Fariba Ariaei, "Developing Successful Autonomous Ground Vehicles: Lessons Learned from DARPA Challenges", *12th IFAC Symposium on Transportation Systems, Year Organization*, pp. 399-406, 2009. 8
- [2] Byung-Hyun Lee, Jong-Hwa Song, Jun-Hyuck Im, Sung-Hyuck Im, Moon-Beom Heo, Gyu-In Jee, "GPS/DR Error Estimation for Autonomous Vehicle Localization", *Sensors 2015,15, 20779-20798*, pp. 20779-20798, 2015.
- [3] H. Jung, "An efficient lane detection algorithm for lane departure detection", *Intelligent Vehicles Symposium (IV), 2013 IEEE*, pp. 976-981, 2013. 06
- [4] Johan Bengtsson, "Adaptive Cruise Control and Driver Modeling", *Lund Institute of Technology*, 2001.11
- [5] Bradford W. Parkinson, Penina Axelrad, "Global Positioning System: Theory and Applications, Volume U", *American Institute of Aeronautics and Astronautics*, pp. 3-4, 1996.
- [6] Marcus G. Ferguson, "Global Positioning System (GPS) error source prediction", *American Institute of Aeronautics and Astronautics*, pp. 8-10, 2000. 3
- [7] Peter Steigenberger, Manuela Seitz, Sarah Bockmann, Volker Tesmer, Urs Hugentobler, "Precision and accuracy of GPS-derived station displacements", *Physics and Chemistry of the Earth 53-54*, pp. 72-79, 2010. 8
- [8] Aly M. El-naggar, "Enhancing the accuracy of GPS point positioning by converting the signal frequency data to dual frequency data", *Alexandria Engineering Journal*, Volume 50, Issue 3, pp. 237-243, 2011.08
- [9] J. William, "Global positioning system standard positioning service performance analysis report", *FAA GPS Performance Analysis Report*, 2014. 7
- [10] Mohamad Hosein Refan, Adel Dameshghi, Mehrnoosh Kamarzarrin, "Improving RTDGPS accuracy using hybrid PSOSVM prediction model", *Aerospace Science and Technology 37*, pp. 55-69, 2014. 5
- [11] T.H. Witte, A.M. Wilson, "Accuracy of WAAS-enabled GPS for the determination of position and speed over ground", *Journal of Biomechanics 38*, pp. 1717-1722, 2004. 7
- [12] Lisa L. Arnold, Paul A. Zandbergen, "Positional accuracy of the Wide Area Augmentation System in consumer-grade GPS units", *Computers & Geosciences 37*, pp. 883-892, 2011. 3
- [13] Atsushi Shimamura, "MSAS (MTSAT Satellite-based Augmentation System) Project Status", *Global Positioning System by ION, Volume VI*, 1999. 6
- [14] Wang Bing, Sui Lifen, Xiao Gurui, Duan Yu, Qi Guobin, "Comparison of attitude determination approaches using multiple Global Positioning System (GPS) antennas", *Geodesy and Geodynamics*, Volume 4, Issue 1, pp. 16-22, 2013.02
- [15] John A. Klobuchar, "Ionospheric Effects on GPS", *Air Force Geophysics Laboratory, GPS world*, 1991. 4
- [16] Paulo de Oliveria Camargo, Joao Francisco Galera Monico, Luiz Danilo Damasceno Ferreira, "Application of ionospheric corrections in the equatorial region for L1 GPS users", *Earth Planets Space, 52*, pp. 1083-1089, 2000.

- [17] Johta Awano, Takahiko Ikeda, Michito Imae, "The Transistor Technology(In Japanese)", *CQ publishing company, ISSN 0040-9413*, pp. 41-88, 2016.02
- [18] Chulliat, A., S. Macmillan, P. Alken, C. Beggan, M. Nair, B. Hamilton, A. Woods, V. Ridley, S. Maus, A. Thomson, "The US/UK World Magnetic Model for 2015-2020", *National Geophysical Data Center, NOAA* 2015.
- [19] Ponlathap Lertworawanich, "Safe-Following Distances Based On The Car-Following Model", *PIARC International Seminar on Intelligent Transport System (ITS) In Road Network Operations*, 2006. 8
- [20] Tiefang Zou, Zhi Yu, Ming Cai, Jike Liu, "Analysis and application of relationship between post-braking-distance and throw distance in vehicle?pedestrian accident reconstruction", *Forensic Science International* 207, pp. 135-144, 2010. 10
- [21] Shin Takehara, "The first step of vehicle dynamics (In Japanese)", *Morikita Publishing Co., Ltd.*, pp. 99-111, 2014. 10
- [22] S. Marisamynathan, P. Vedagiri, "Modeling Pedestrian Delay at Signalized Intersection Crosswalks under Mixed Traffic Condition", *Procedia - Social and Behavioral Sciences* 104, pp. 708-717, 2013. 11
- [23] Tom Urbanik, Alison Tanaka, Bailey Lozner, Eric Lindstrom, Kevin Lee, Shaun, Quayle, Scott Beaird, Shing Tsoi, Paul Ryus, Doug Gettman, Srinivasa Sunkari, Kevin Balke, Darcy Bullock, "Signal Timing Manual, Second Edition", *NCHRP Report 812, Transportation Research Board* 2015, 2015
- [24] Bernhard Hofmann-Wellenhof, Herbert Lichtenegger, Elmar Wasle, "GNSS-Global Navigation Satellite Systems GPS, GLONASS, Galileo and more", *SpringerWienNewYork*, pp. 272-276, 2008.

An Efficient Scheme for Real-time Information Storage and Retrieval Systems: A Hybrid Approach

Syed Ali Hassan*, Imran Ul Haq*, Muhammad Asif*, Maaz Bin Ahmad[†] and Moeen Tayyab[‡]

*Department of Computer Science and Information Technology

Lahore Leads University, Lahore, Pakistan

[†] COCIS, PAF Karachi Institute of Economics and Technology, Karachi, Pakistan

[‡] International Islamic University, Islamabad, Pakistan

Abstract—Information storage and retrieval is the fundamental requirement for many real-time applications. These systems demand that data should be sorted all the time, real-time insertion, deletion and searching should be supported and system must support dynamic entries. These systems require search operations to be performed from massive databases implemented by various data structures. The common data structures used by these systems are stack, queue or linked list all having their own limitations. The biggest advantage of using stack is that binary search can be performed on it easily while on the other hand insertion and deletion of nodes involves more processing overhead. In linked list, insertion and deletion of nodes is easier but searching operation involves more processing overhead as binary search cannot be performed efficiently on it. In this paper, a hybrid solution is presented for such systems, which provides efficient insertion, deletion and searching operations. Results show the effectiveness of the proposed approach as it outperforms the existing techniques used by these systems.

Keywords—Insertion; deletion; array; linked list; binary search; linear search

I. INTRODUCTION

The efficient information retrieval, insertion and searching is the basic need for most of the applications of this modern computing era. These applications require efficient data structures to store and retrieve large amount of information. Normally the information is either stored in arrays or linked list. In arrays, searching can be done efficiently using binary search technique. As binary search is less computationally intensive as compared to the linear search especially when the data set is too large, so it is the desired searching technique used by many real-time applications. But the problem using array is that insertion and deletion of nodes requires more shifting operations which becomes a hurdle to use it in real-time scenarios. Linked list efficiently resolve this issue of real-time insertion and deletion of nodes as it requires only updating the pointers values, so it seems more appropriate to use it in real-time applications. But the main problem using linked list is that binary search cannot be implemented on it directly because we cannot search a node without traversing all the previous nodes. This is because the memory allocation of linked list is not contiguous and is allocated at run time while in arrays the nodes reside on contiguous memory locations.

Linear search algorithm searches a node from array or linked list by inspecting each of the nodes in it and comparing it with the search node. In linear search, time required to find a node directly depends on the total number of elements in

the array or linked list. So, the complexity of linear search is $O(N)$ [1], [2] as in Big-O notation. This search technique has the simplest implementation, so it can be applied to array list and all types of linked lists. But it is not efficient when the size of the list is too large. It is useful only when the size of an array or a linked list is small. Binary search is more efficient searching technique and is quite suitable when the number of nodes is more in an array list. The requirement of binary search algorithm is that the elements of an array must be in sorted form [1], [2]. Every iteration of this algorithm makes half the search interval of its previous iteration, so lesser number of comparisons is required to search a node. The complexity of binary search algorithm is $O(\log_2 N)$. So if we can manage to apply binary search efficiently on linked list, it would become an ideal data structure for supporting real-time insertion, deletion and searching. It may enhance the performance of many real-time applications like vehicle exit-control system.

In this paper, a hybrid solution is presented in order to facilitate the real-time applications in terms of efficient insertion, deletion and searching. In a linked list is used to store nodes data and a combination of linear and binary searching techniques are used to efficiently find a node. The proposed technique outperforms the existing solutions for these kinds of applications.

The rest of the paper is organized as follows. Section II presents the related work. The proposed methodology is presented in Section III. Section IV presents the experimental analysis. Finally, the conclusion is drawn in Section V.

II. RELATED WORK

As discussed in the previous section, binary search algorithm can only be applied to sorted array whether it is static or dynamic. This algorithm cannot be directly implemented to linked list [2]. Although the advantages of binary search can be obtain through organization of array elements in non linear data structure tree [3]. Binary search tree searches an element in equal amount of time as taken by binary search $O(\log_2 N)$ [1], [4]. But it is difficult to maintain and manipulate binary search tree.

The second option to implement binary search on linked list is to copy all the elements of linked list into either sorted array or a binary search tree [5]. This option is again not practical for maintenance of the data as each time searching will be faster but require more processing of creating and copying elements.

Extra overload will be faced by processor in obtaining the benefits of binary search. Several other researchers worked in this domain. Kumar et al. [6] discussed that linear search is efficient than binary search if we add sorting time also in case of binary search. The argument can be nullified in applications where sorted data is a requirement. Arora et al. [7] presented a two way linear search approach through which searching is performed from both ends of the list. It is efficient only in cases where the node to be searched belongs to the second half of the list. Chadah et al. [8] tried to reduce the worst case time of binary search algorithm by increasing number of comparisons in each iteration. Naidu et al. [9] targeted the large memory requirements of doubly linked list and proposed an implementation of single linked list to achieve the benefits of doubly linked list. The Ex-Or operation was used in single list in order to traverse in both direction.

The third option is to derive a new methodology that can perform computationally efficient searching in linked list. This may help to develop a real-time information storage and retrieval systems that allows searching, insertion and deletion operations.

III. PROPOSED SOLUTION

Let us consider a doubly linked list structure that consists of a set of sequentially linked records called nodes. Each node contains 'info' and 'links' fields. The 'info' field stores the information and reference or pointer to the previous and next node in the linked list containing 'links' field. In doubly linked list, navigation is possible in both forward and backward ways easily as compared to single linked list. According to the proposed solution, linked list is organized in order of 'info' field in an ascending order. After that, a track of few selected key nodes is maintained in a separate array of pointers. This pointer array is known as sparse array.

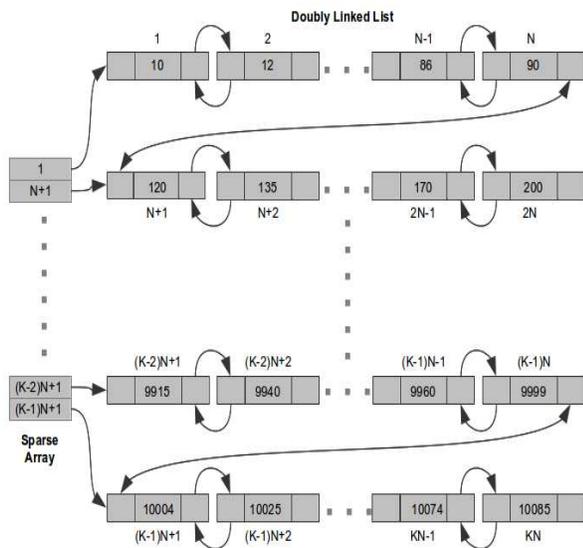


Fig. 1. Arrangement of the doubly linked list and sparse array.

Fig. 1 shows the arrangement of the sorted doubly linked list and sparse array. The sorted linked list consists of $K * N$ number of elements and the first node of each block is taken as key node. The address of each key node is stored in the sparse

array. In Fig. 1 there are K numbers of key nodes and N is the number of entries between two key nodes. In this work, sparse array is used to perform the searching, insertion and deletion operations in the linked list. The following sections describe the searching, insertion, deletion and up gradation of sparse array.

A. Searching Operation

To search the desired pattern in the linked list, a hybrid binary linear search technique (HST) is proposed based on the sparse array. In this technique, initially binary search is performed using key nodes sparse array. If the desired pattern is present in the key nodes than output of binary search is its exact location i.e. for searched pattern 10, 120, 9915 and 10004. On the other hand, if the data we are looking for is not located in the key nodes than the outcome of binary search are two key nodes ' K_i ' and ' K_f ' between whom the desired pattern can be laid. In this case linear search is performed on linked list records between ' K_i ' and ' K_f ' to find the exact pattern location. For example, if we want to search pattern '9960' than outcome of binary search will be key node ' $(K - 2)N + 1$ ' and ' $(K - 1)N + 1$ ' using these key nodes linear search can be performed on linked list. The block diagram of the proposed searching algorithm is shown in Fig. 2.

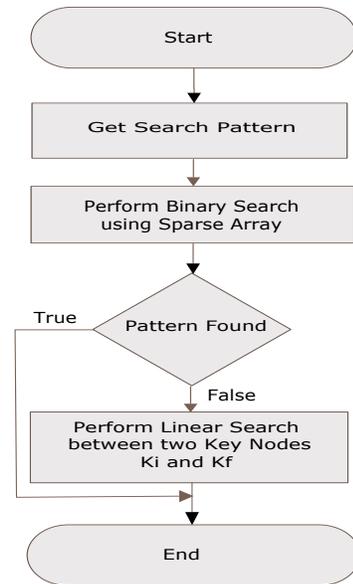


Fig. 2. Block diagram for the proposed search technique.

In the proposed HST, binary search helps to reduce the search time by either giving exact match or by reducing the search space by giving the address of two key nodes as a linear search starting and ending point. The pseudo code for binary and linear search using sparse array is presented in Algorithm 1 and 2, respectively.

Algorithm 1: SparseArrayBinarySearch(S,n, Pattern,N)

- Input: Sparse array S, n is the total number of elements in S, Pattern is the value to be search and N is the number of records between two key nodes
- Output: Position k such that $S[k] \rightarrow \text{info} = \text{Pattern}$, or two key nodes K_i and K_f for linear search if desired

pattern is not found.

- 1) $i \leftarrow 0, j \leftarrow n-1, k \leftarrow (i+j)/2, K_i \leftarrow -1, K_f \leftarrow -1,$ and $r \leftarrow 0$
- 2) while ($i \leq j$)
- 3) do
- 4) if($S[k] \rightarrow \text{info} > \text{Pattern}$) then $j \leftarrow k-1$ and $r \leftarrow 1$
- 5) else if($S[k] \rightarrow \text{info} < \text{Pattern}$) then $i \leftarrow k+1$ and $r \leftarrow -1$
- 6) else return k // successful search
- 7) if($i > j$)
- 8) $K_i \leftarrow k$
- 9) if($r > 0$) then $K_i \leftarrow K_i-1$
- 10) if($K_i > N-1$) then $r = N-1$
- 11) else if($K_i < 0$) then $r = 0$
- 12) else $K_f \leftarrow K_i+N$ return K_i and K_f //end if($i > j$)
- 13) $k \leftarrow (i+j)/2$ //end while

Algorithm 2: SparseArrayLinearSearch(S,Pattern,K_i, K_f)

- Input: Sparse array S, Pattern to be search, linear search starting and ending points K_i and K_f
- Output: Position i such that S[i] → info = Pattern, -1 if desired pattern is not found

- 1) $i \leftarrow K_i$ and $LN \leftarrow S[i]$
- 2) while ($i \leq K_f$)
- 3) do
- 4) if($LN \rightarrow \text{info} = \text{Pattern}$) then return i
- 5) else $LN = LN \rightarrow \text{next}$
- 6) $i \leftarrow i + 1$ //end while
- 7) if ($i < K_f$) then return i
- 8) else return -1

B. Insertion Operation

To insert a given pattern in a sorted linked list, following five steps are involved. First, a new node is allocated and input pattern is stored in the 'info' field of the node. Second, to concatenate the new node with sorted linked list, insertion location is determined by using HST. Third, 'link' fields of new node; store the addresses of its previous and next node in sorted linked list. Fourth, the 'link' fields of the previous and next node are modified according to new node. Finally, up-gradation of sparse array is made which is necessary for the further operations. Fig. 3 shows the block diagram for node insertion operation.

C. Deletion Operation

This operation is used to delete a specific node from the sorted linked list. The node deletion is a four step process. First, a node whose 'info' field contains the desired pattern is determined using proposed HST. Second, redirection of 'links' is performed in the previous and next nodes of the node to be deleted. Third deleted node is De-allocated. Finally, up-gradation of sparse array is performed according to modified linked list for further operations. Fig. 4 shows the block diagram for node deletion operation.

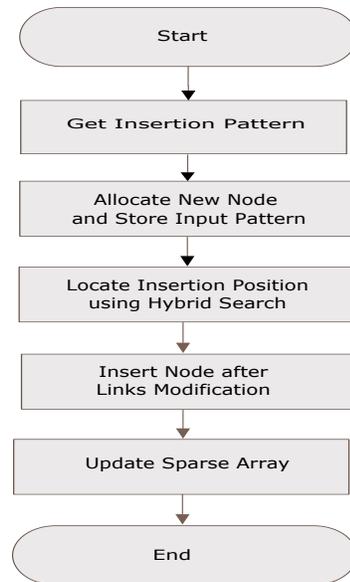


Fig. 3. Block diagram for node insertion operation.

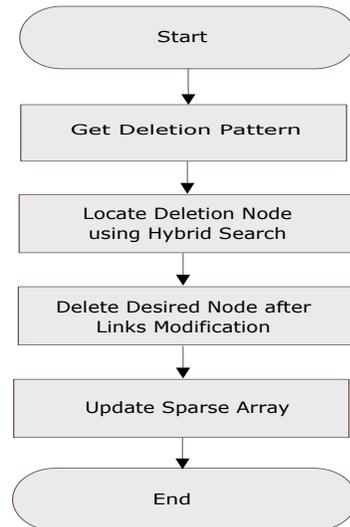


Fig. 4. Block diagram for node deletion operation.

D. Updating of Sparse Array

The insertion and deletion of nodes from the linked list demand up-gradation of the sparse array. The up-gradation is performed according to deletion or insertion operation. In case of insertion operation, all the key nodes that exist beyond the insertion location will be replaced with their corresponding next node in the linked list. On the other hand, after deletion operation all the key nodes that exist beyond the deletion location will be replaced with their corresponding previous node. The pseudo code for up-gradation of the sparse array is given in Algorithm 3.

Algorithm 3: UpDateSparseArray(S,k,m,n)

- Input: S is sparse array, k denotes the index from where S is needed to be upgrade, m indicates the operation insertion or deletion after which up-gradation is required and n is the numbers of entries in S.

- Output: Upgraded sparse array S
- 1) $i \leftarrow K$
 - 2) while ($i \leq n$)
 - 3) do
 - 4) if ($m = 0$) then $S[i] = (S[i] \rightarrow \text{previous})$ // in case of deletion
 - 5) else $S[i] = (S[i] \rightarrow \text{next})$ // in case of insertion
 - 6) $i \leftarrow i + 1$ //end while

IV. EXPERIMENTAL ANALYSIS

The experiments are conducted on PC with intel-core i3-2100 CPU @ 3.1 GHz and 2 GB RAM. The proposed system is a combination of linked list and sparse array based hybrid search (HS-LL) technique. The comparison of the proposed information storage and retrieval system is done with two possible scenarios. First one is based on array using binary search (BS-AR) and second is based on the linked list and linear search (LS-LL) methodology. The experiments are performed on sorted array and linked list having different range of entries between 5000 and 100,000.

Tables I, II and III list the experimental results in terms of the time taken T_s to search, insert and delete the entries using three information storage and retrieval systems, respectively. The experiments are performed by considering the boundary cases i.e, by performing searching, inserting and deletion operations at the start and end position (index or node) in both array and linked list.

Table I indicates that the data searching from sorted array using binary search (BS-AR) performs equally well in both cases either data to be search is located at the start or end of the array. It is observed that data searching in linked list using linear search technique (LS-LL) is worse when desired data is located at the end of the linked list. In this case, searching time is directly related with the number of entries in the linked list. Table I shows that proposed solution (HS-LL) almost perform equally well as that of binary search in array. Fig. 5 shows the performance analysis of three techniques in terms of time taken to search a node or entry located at the middle of linked list and array. It depicts that HS-LL and BS-AR perform equally well and have better performance than LS-LL technique.

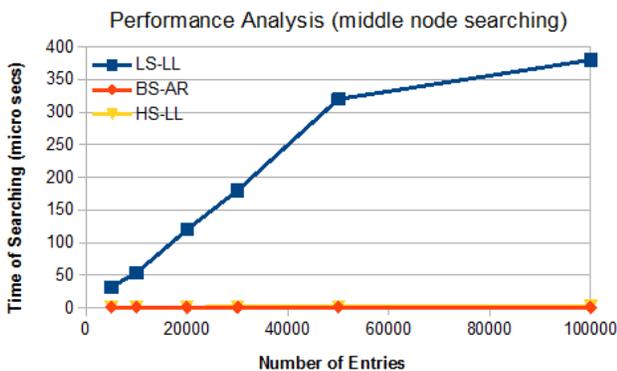


Fig. 5. Data searching using LS-LL, BS-AR and HS-LL.

Table II illustrates that data insertion at the start of the sorted array using BS-AR takes significant time due to shifting

of huge amount of data. Contrarily, array insertion is efficient when element is inserted at the last index because no shifting operations are required to sort the array data. Table II also demonstrates that linked list insertion using linear search (LS-LL) at the start is efficient because insertion position is found at the first attempt during linear searching. On the other hand, node insertion using linear search at the end gives the worse performance because huge processing time is consumed to search the node insertion position. Table II depicts that the linked list insertion with proposed hybrid search technique (HS-LL) perform equally well in both cases either data is inserted at the start or end of the sorted linked list. Fig. 6 depicts the performance analysis of three techniques in terms of time taken to insert a node or entry at the middle of linked list and array. It shows that HS-LL have better performance than BS-AR and LS-LL technique.

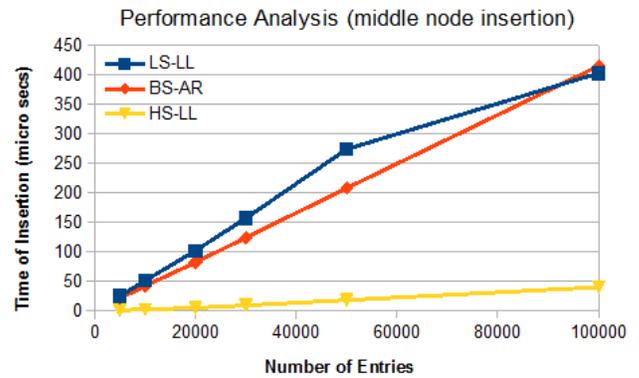


Fig. 6. Data insertion using LS-LL, BS-AR and HS-LL.

Table III shows that for deletion operation behavior of all the techniques is similar to that of insertion operation. The data deletion from the start of the sorted array takes significant time due to shifting of all array elements. On the other hand, it takes small time when element is deleted from the end of array because shifting operations are not required in this case. Table III lists that linked list deletion using linear search from the start of the sorted linked list is efficient because desired node is found at the first attempt during linear searching. Contrarily, last node deletion takes maximum time due to huge searching time to locate the desired node. Table III depicts that the linked list deletion with the proposed hybrid search technique perform equally well in both cases either data is deleted from the start or end of the sorted linked list. Fig. 7 shows the performance analysis of three techniques in terms of time taken to delete a node or entry from the middle of linked list and array. It shows that HS-LL have better performance than BS-AR and LS-LL technique.

The experimental results indicate that the proposed HS-LL solution performs equally well for data searching, insertion and deletion either at start or end. It also outperforms the rest of the two possible scenarios.

It should be noted that the performance of proposed solution is highly correlated with the size of the sparse array. For the large size of sparse array, huge shifting operations are required for its up-gradation. Contrarily, small size reduces the overhead related to sparse array up-gradation process at the cost of increase in linear search. Therefore, size selection of

TABLE I. DATA SEARCHING TIME (μ SEC)

No. of Entries	5000		10000		20000		30000		50000		100000	
	Start	End	Start	End								
BS-AR	1	1	1	1	1	1	1	1	1	1	2	1
LS-LL	1	75	1	109	1	248	1	348	1	720	1	771
HS-LL	1	1	1	1	1	1	1	2	2	2	3	3

TABLE II. DATA INSERTION TIME (μ SEC)

No. of Entries	5000		10000		20000		30000		50000		100000	
	Start	End	Start	End								
BS-AR	43	1	85	1	164	1	247	1	415	2	831	2
LS-LL	1	51	1	102	1	204	1	312	1	545	1	804
HS-LL	2	1	3	1	8	3	12	6	24	11	58	24

TABLE III. DATA DELETION TIME (μ SEC)

No. of Entries	5000		10000		20000		30000		50000		100000	
	Start	End	Start	End								
BS-AR	46	1	91	1	181	2	273	1	458	1	913	2
LS-LL	1	60	1	125	1	251	1	396	1	633	1	768
HS-LL	2	1	3	2	7	3	13	6	22	10	55	22

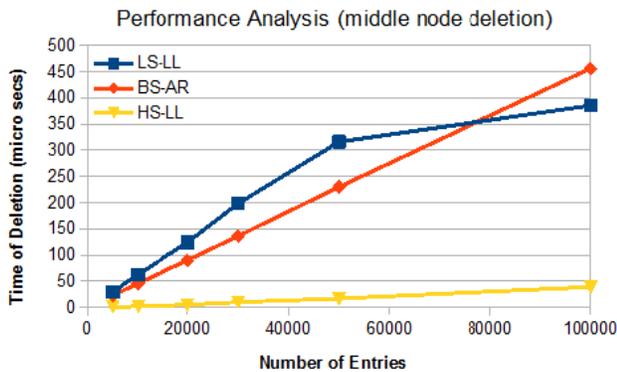


Fig. 7. Data deletion using LS-LL, BS-AR and HS-LL.

sparse must be optimum in such a way that it neither increases the shifting operations nor hurts the searching performance. In this experimental analysis the size of sparse array is taken as 64.

V. CONCLUSION

This paper presented an efficient information storage and retrieval system to facilitate the real-time applications. In this solution, linked list is used to store the information and a hybrid linear binary search technique based on sparse array is proposed to perform efficient data insertion, deletion and searching operations. The experimental results reveal that the proposed methodology outperforms the existing techniques for such kinds of applications.

REFERENCES

- [1] Jean-Paul Tremblay and P. G. Sorenson: "An Introduction to data structures with applications", Mcgraw Hill Computer Science Series, 2nd Edition.
- [2] A. Oommen and C. Pal: "Binary Search Algorithm", Journal Of Innovative Research In Technology, 1(5), 800-803, 2014.
- [3] S. Pushpa1 and P. Vinod: "Binary Search Tree Balancing Methods: A Critical Study", International Journal of Computer Science and Network Security, 7(8),2007.
- [4] P. P. Thwe and L. L. W. Kyi: "Modified Binary Search Algorithm for Duplicate Elements", International Journal of Computer and Communication Engineering Research (IJCCER), 2(2), 77-81, 2014.
- [5] P. Das and P. M. Khilar: "A Randomized Searching Algorithm and its Performance analysis with Binary Search and Linear Search Algorithms", The International Journal of Computer Science and Applications (TIJCSA), 1(11), 11-18, 2013.
- [6] D. Kumar and M. Sharma: "Binary search is faster than the linear search", International Journal of Innovative Research in Technology, 1(5), 796-799, 2014.
- [7] N. Arora, G. Bhasin and N. Sharma: "Two way Linear Search Algorithm", International Journal of Computer Applications, 107(21), 6-8, 2014.
- [8] A. R. Chadha, R. Misal and T.Mokashi: "Modified Binary Search Algorithm", International Journal of Applied Information Systems (IJ AIS), 7(2), 37-40, 2014.
- [9] D. Naidu and A. Prasad: "Implementation of Enhanced Singly Linked List Equipped with DLL Operations: An Approach towards Enormous Memory Saving", International Journal of Future Computer and Communication, 3(2), 2014.

Exploiting Temporal Information in Documents and Query to Improve the Information Retrieval Process: Application to Medical Articles

Jihen MAJDOUBI

Department of Computer Science
College of Science and Humanities at AlGhat
Majmaah University, P.O. Box 66, Majmaah 11952
Kingdom of Saudi Arabia

Ahlam Nabli

Department of Computer Science
College of bandaq
Albaha university
Kingdom of Saudi Arabia

Abstract—In the medical field, scientific articles represent a very important source of knowledge for researchers of this domain. But due to the large volume of scientific articles published on the web, an efficient detection and use of this knowledge is quite a difficult task. In this paper, we propose a novel method for semantic indexing of medical articles by using the semantic resource MeSH (Medical Subject Headings) and the temporal information provided in the documents. The proposed indexing approach was evaluated by intensive experiments. These experiments were conducted on document test collections of real world clinical extracted from scientific collections, namely, CISMEF and CLEF. The results generated by these experiments demonstrate the effectiveness of our indexing approach.

Keywords—Biomedical information retrieval; semantic indexing; temporal criteria; Medical Subject Headings (MeSH) thesaurus

I. INTRODUCTION

The WWW becomes a very vast repository of data and the volume of information generated in this digital world is increasing day by day. This, however, would be wasted if necessary information could not be found, analyzed, and exploited. The goal of any Information Retrieval System (IRS) is to retrieve relevant information to a users query.

This goal is quite a difficult task with the rapid and increasing development of the Internet. Indeed, web information retrieval becomes more and more complex for the user who IRS provides a lot of information. However the user often fails, to find the best information in the context of his/her information need.

The problem in searching over documents is that documents are time-dependent and accumulated over time which results in a large number of irrelevant documents in a set of retrieved documents. Therefore, the users have to spend more time in finding the documents that are satisfying his/her information need. Traditional Information Retrieval approaches based on topic similarity alone is not sufficient for the search in growth document collections. Much research is going on the field of temporal information retrieval to improve the retrieval results. The time criterion has already been the core concept of recent IR ranking models, given that most of documents include a high level of temporal information [23]. Indeed, several

works show that a large amount of web documents become time-dependent [2], [21]. The authors in [13] have argued that about 7% of queries have implicit temporal intent, while other studies show that only 1.5% of queries are explicitly provided with temporal intent.

In this paper, we are interested in the temporal information and its impact in the process of medical article indexing. Our motivation is that timeliness is one of the key aspects that determine a documents credibility besides relevance, accuracy, objectivity and coverage.

The treatment of medical information has made the interest of several research works and a lot of solutions have been proposed so far, based on context query, online ranking model, semantic model. However, to the best of our knowledge, there is no prior attempts dealing with the use of the temporal criteria in the biomedical IR field. In this paper, we propose a novel method for conceptual indexing of medical articles by using the semantic resource MeSH (Medical Subject Headings) and the temporal information provided in the documents. Specifically, both temporal relevance and topic similarity are needed for efficient retrieval. The remainder of this paper is organized as follows. In the next section, we attempt to prove the effectiveness of exploiting temporal criteria in the information retrieval process. Section 3 summarizes our context and motivations. In Section 4, we review the related work. Section 5 details our conceptual indexing approach. An experimental evaluation and comparison results are discussed in Sections 6 and 7. Finally, Section 8 presents some conclusions and future work directions.

II. USING TEMPORAL INFORMATION TO IMPROVE THE RETRIEVAL RESULTS

The notion of using time as an important factor becomes more important for a large number of searches. In the following, we attempt to prove the effectiveness of exploiting temporal criteria in the information retrieval process.

Consider an example of historian interested in knowing about the Tunisia revolution that occurs in past years. He searches in the news archives expecting to retrieve the details of the event- not necessarily the latest news, but a report on

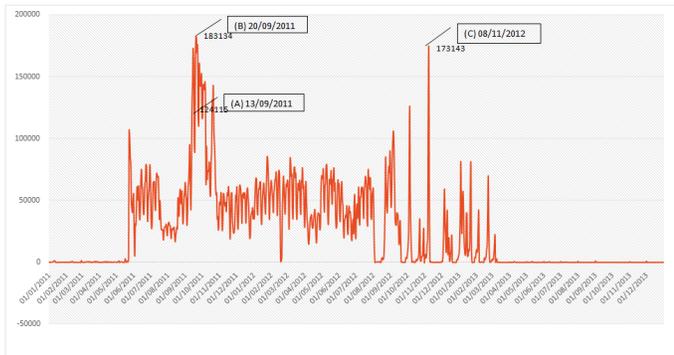


Fig. 1. Number of documents containing the term “barrack obama”.

the latest news about that query is retrieved. The most of the relevant documents for that query is for the time period of 2010 – 2011 or associated with the time that event happened. This example shows clearly that timeliness is one of the key aspects that determine a documents credibility besides relevance, accuracy, objectivity and coverage. Both temporal relevance and topic similarity are needed for efficient retrieval.

In order to better prove this, we try to analyze the frequency of the term “barrack obama” in the “LivingKnowledge sub collection” over the time. This collection spans from May 2011 to March 2013 and contains around 3.8M documents collected from about 1500 different blogs and news sources. The data is split into 970 files, named after the date of that day and some information about its sources (there might be more than one file per day). We plot in Fig. 1 curve representing the number of documents containing the term “barrack obama” in the “Living Knowledge news and blogs” corpus.

The *x-axis* represents time in months (From 01/01/2011 to 01/12/2013), and the *y-axis* indicates the number of documents containing the term “barrack obama” over the corpus.

Fig. 1 clearly shows that the number of documents containing the term “barrack obama” increases significantly during specific time periods.

For example as highlighted at Fig. 1 in 13/09/2011, we have 124116 blogs and news sources discuss about “obama”. This number has grown to reach a value of 183134 in 20/09/2011. By referring to the timeline of the presidency of Barack Obama in 2011 (see Fig. 2), we can remark that peaks (A) and (B) presented in Fig. 1 are well-lined up with the timeline of most actions made by Barack Obama to revive the American economy.

Also, the third peak (C) off the graphic appearing at 08/11/2011 with 173143 news corresponds to date of re-election Barack of Obama (November6, 2012). This is mainly due to the fact that people tend to talk about the Obama’s news mainly during or slightly after time periods when the action was held and number of documents created beyond these time periods increases significantly.

On the basis of examples presented in this section we can confirm that time dimension must be exploited as a highly important relevance criterion to improve the retrieval effectiveness of document ranking models.

- **September 8** – President Obama presents the American Jobs Act, his plan to create jobs and revive the American economy.
- **September 12** – The President delivers a speech in the White House Rose Garden to promote his American Jobs Act.
- **September 16** – President Obama signs the America Invents Act, (H.R. 1249), a major overhaul of the U.S. patent system, into law.
- **September 19** – The President releases his debt reduction plan and the Buffett Rule.

Fig. 2. Timeline of the presidency of Barack Obama in 2011.

III. CONTEXT AND MOTIVATIONS

Each year, the rate of publication of biomedical literature grows, making it increasingly harder for researchers to keep up with novel relevant published work. In recent years several researches have been devoted to attempt to manage effectively this huge volume of information.

In [14], the authors proposes a tool called MAIF (MeSH Automatic Indexer for French) which is developed within the CISMef team. To index a medical resource, MAIF follows three steps: analysis of the resource to be indexed, translation of the emerging concepts into the appropriate controlled vocabulary (MeSH thesaurus) and revision of the resulting index.

In [15], the authors proposed the MTI (MeSH Terminology Indexer) to index English resources. MTI results from the combination of two MeSH Indexing methods: MetaMap Indexing (MMI) [16] and a statistical, knowledge-based approach called PubMed Related Citations (PRC) [9].

The conceptual indexing strategy proposed by [8] involves three steps. First they compute for each concept MeSH C its similarity with the document D . After that, the candidate concepts extracted from step 1 are re-ranked according to a correlation measure that estimates how much the word order of a MeSH entry is correlated to the order of words in the document. Finally the content based similarity and the correlation between C and the document D are combined in order to compute the overall relevance score. The N top ranked concepts having the highest scores are selected as candidate concepts of the document D .

The indexing approach presented in this paper differs from previous works. In this paper, we are interested in integrating temporal information in the process of medical article indexing. Our motivation is that Temporal information is crucial in biomedical information systems. Healthcare providers normally record the progress of a disease or a hospital course chronologically in text, and procedures and laboratory tests are stored in databases with time-stamps. Therefore, automatically reasoning about temporal information can help us understand the dynamics of medical phenomena and may potentially improve the quality of patient care.

IV. RELATED RESEARCH WORKS

Temporal Information Retrieval has started to be considered as a subdivision of the field of information retrieval. In this section, we provide a comprehensive and a comparative overview of most important work on both time and IR.

Li and Croft [11] defined two types of time-based queries in TREC collections that contain news archives. The first favors the most recent documents and the second is shown to have relevant documents within a specific period in the past. To incorporate time information into retrieval models, they proposed a time-based language model using a prior based on an exponential or a normal distribution depending on the types of recency queries.

In [19], the authors proposed an extension to the Query Likelihood Model that incorporates query-specific information to estimate rate parameters. They also introduced a temporal factor into language model smoothing and query expansion using pseudo-relevance feedback. These extensions were evaluated using a Twitter corpus and two newspaper article collections. Results showed that, compared to prior approaches, models proposed are more effective at capturing the temporal variability of relevance associated with some topics.

In [12], the authors proposed a query expansion model for microblogs, which selects terms temporally closer to the query submission time. Their model is supposed to work well for finding documents related to events currently happening but, not as well for past events.

In [10], the authors suggested a general language model that incorporates time into the ranking model in a principled manner. For a given time-sensitive query over a news archive, the approach automatically identifies significant time intervals for the query and uses them to adjust the document relevance scores by boosting the scores of documents published within the important intervals. They presented an extensive experimental evaluation, including TREC as well as an archive of news articles, and showed that proposed techniques improve the quality of search results, compared to the existing state-of-the-art algorithms.

In [26], the authors presented an adaptive temporal query modeling for blog feed retrieval, in that they analyzed the top retrieved documents in terms of temporal histogram to find the bursts. They used documents with the highest scores from the bursts for query expansion and weighted each feedback document with the distance from the peak that contains most documents.

In [13], [24], the authors proposed a temporal query expansion method for microblogs based on the temporal co-occurrence of terms in a timespan. They first performed pseudo-relevant timespan retrieval for an event query and used those timespans for query expansion. Although their goal was retrieving a ranked list of historical event summaries, the temporal query expansion method showed that selecting relevant timespan is crucial for query expansion for microblog documents.

The state-of-the-art presented in this section shows that temporal information retrieval has shown its performance in many scopes. In this paper, we try to exploit temporal information in medical documents to improve the information retrieval

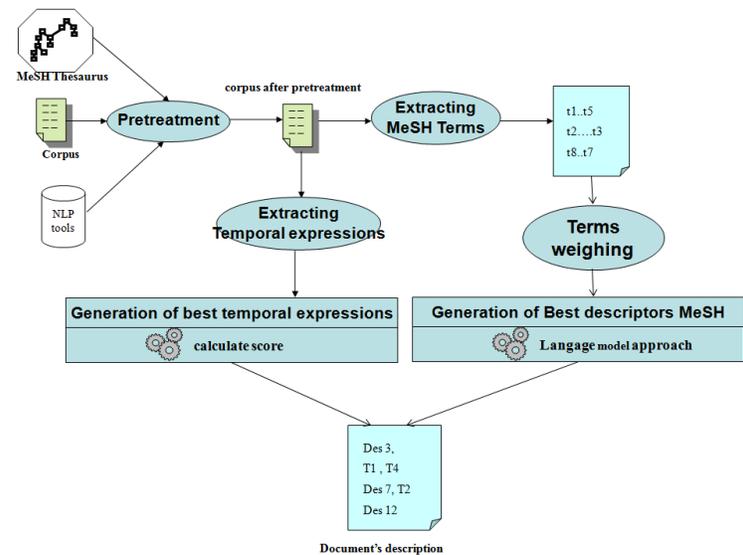


Fig. 3. Architecture of proposed indexing approach.

process. Our choice is motivated by the fact that temporal information is crucial in biomedical information systems and procedures and laboratory tests are stored in databases with time-stamps.

V. PROPOSED APPROACH

In [5], we have proposed an approach for conceptual indexing of medical articles by using the MeSH (Medical Subject Headings) thesaurus. More precisely, we have tried to determine for each document, the most representative MeSH descriptors. For this reason, we deduced a language model for each document and rank Mesh descriptor according to our probability of producing each one given that model. The proposed indexing approach was evaluated by intensive experiments in [6], [7]. These experiments were conducted on document test collections of real world clinical extracted from scientific collections, namely PUBMED and CLEF. The results generated by these experiments demonstrated the effectiveness of proposed indexing approach.

To improve these results, we integrate the time criteria in the indexing process. we made an assumption that the Time plays important roles in medical articles because healthcare providers normally record the progress of a disease or a hospital course chronologically in text.

Our indexing methodology as schematized in Fig. 3, consists of five steps: 1) Pretreatment; 2) Extracting MeSH concepts; 3) Extracting temporal expressions; 4) Generation of Best descriptors MeSH; and 5) Generation of best temporal expressions. We describe below the structure of MeSH vocabulary and then we detail the steps of proposed indexing method.

A. MeSH Thesaurus

The language of biomedical texts, like all natural language, is complex and poses problems of synonymy and polysemy.

Therefore, many terminological systems have been proposed and developed such as Galen ¹, UMLS ² and MeSH ³.

In our context, we have chosen MeSH because it meets the aims of medical librarians and it is a widely used tool for indexing literature.

The structure of MeSH is centered on descriptors, concepts, and terms.

- Each term can be either a simple or a composed term.
- A concept is viewed as a class of synonymous terms, one of them (called Preferred term) gives its name to the concept.
- A descriptor class consists of one or more concepts closely related to each other in meaning. Each descriptor has a Preferred Concept. The descriptor's name is the name of the preferred Concept. Each of the subordinate concepts is related to the preferred concept by a relationship (broader, narrower).

Cardiomegaly [Descriptor]
Cardiomegaly [Concept, Preferred]
Cardiomegaly [Term, Preferred]
Enlarged Heart [Term]
Heart Enlargement [Term]
Cardiac Hypertrophy [Concept, Narrower]
Cardiac Hypertrophy [Term, Preferred]
Heart Hypertrophy [Term]

Fig. 4. An example of MeSH.

As shown by Fig. 4, the descriptor "Cardiomegaly" consists of two concepts: "Cardiomegaly" and "Cardiac Hypertrophy". Each concept has a preferred term, which is also said to be the name of the Concept. For example, the concept "Cardiomegaly" has three terms "Cardiomegaly" (preferred term), "Enlarged Heart" and "Heart Enlargement". As in the example above, the concept "Cardiac Hypertrophy" is narrower than the preferred concept "Cardiomegaly".

B. Pretreatment

The first step is to split text into set of sentences. We use the Tokeniser module of GATE in order to split the document into tokens, such as numbers, punctuation and words. Then, the TreeTagger stems these tokens to assign a grammatical category (noun, verb...) and lemma to each token. Finally, system prunes the stop words for each medical article of the corpus.

This process of pretreatment is also carried out on the MeSH thesaurus.

Fig. 5 outlines the basic steps of the pretreatment phase.

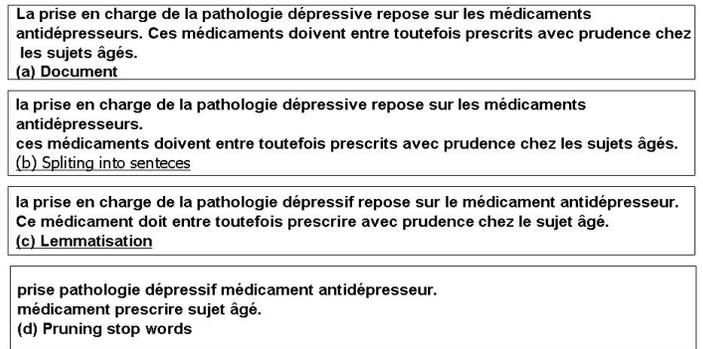


Fig. 5. Pretreatment step.

C. Extracting MeSH Terms

As mentioned above, a term can be either simple or composed. To extract the simple term, we project the Mesh thesaurus on the document by applying a simple matching. More precisely, each lemmatized term in the document is matched with the canonical form or lemma of MeSH terms. To recognize the composed terms, we have chosen to use YateA [27]. YateA (Yet Another Term ExtrAktor) [29] is a hybrid term extractor developed in the project ALVIS. After text processing, YateA generates a file composed of two columns: the inflected form of the term and its frequency. For instance, as shown in Fig. 6 which describes the result of the term extraction process by using YateA, the term "exercice physique" occurs 6 times.

#	Inflected form	Frequency
	activité physique	16
	activité sportive	9
	exercice musculaire	8
	exercice physique	6
	effets bénéfiques	6
	g de glucides	5
	contrôle glycémique	5
	insuffisance coronaire	5
	index glycémique	4
	risque cardiovasculaire	4
	adaptation des doses	4
	glycémie capillaire	4
	sensibilité à l'insuline	4
	fréquence cardiaque	3
	hydrates de carbone	3
	acides gras libres	3
	autosurveillance glycémique	3
	patient dnid	3
	acides gras	3
	dernier repas	3
	profil lipidique	3
	activité physique régulière	3
	insuline rapide	3

Fig. 6. An excerpt of the result of YaTeA.

D. Term Weighing

Given a set of extracted terms issued from the step of "Extracting MeSH Terms", we calculate the terms weight by using two measures: the Content Structure Weight (CSW) and the Semantic Weight (SW) [4].

¹<http://www.opengalen.org>

²<http://www.nlm.nih.gov/research/umls/>

³<http://www.nlm.nih.gov/mesh/>

1) *Content Structure Weight*: We can notice that the frequency is not a main criterion to calculate the CSW of the term. Indeed, the CSW takes into account the term frequency in each part of the document rather than the whole document. For example, a term of the Title receives a higher importance (*10) than to a term that appears in the Paragraphs (*2). Table 1 shows the various coefficients used to weight the term locations. These coefficients were determined in an experimental way in [3].

TABLE I. WEIGHING COEFFICIENTS

term location	Weight of the location
Title (T)	10
Keywords (K)	9
Abstract (A)	8
Paragraphs (P)	2

The CSW of the term t_i in a document d is given as follows:

$$CSW(t_i, d) = \frac{\sum_{A \in T, K, A, P} f(t_i, d, A) \times W_A}{\sum_{A \in T, K, A, P} f(t_i, d, A)} \quad (1)$$

Where,

- W_A is the weight of the location A (see Table I),
- $f(t_i, d, A)$ is the occurrence frequency of the term t_i in the document d at location A .

For example, the term *cancer* exists in the document d_{1683} : 1 time in the title, 2 times in the abstract and 9 times in the Paragraphs,

$$CSW(cancer, d_{1683}) = \frac{1 * 10 + 2 * 8 + 9 * 2}{1 + 2 + 9}$$

2) *Semantic Weight (SW)*: The Semantic Weight of term t_i in the document d depends on its synonyms existing in the set of Candidate Terms ($CT(d)$) generated by the term extraction step. To do so, we use the *Synof* function that associates for a given term t_i , its synonyms among the $CT(d)$. Formally the measure SW is defined as follows:

$$SW(t_i, d) = \frac{\sum_{g \in Synof(t_i, CT(d))} f(g, d)}{|Synof(t_i, CT(d))|} \quad (2)$$

For a given term t_i , we have on the one hand its Content Structure Weight ($CSW(t_i, d)$) and on the other its Semantic Weight ($SW(t_i, d)$), its Local Weight ($LW(t_i, d)$) is determined as follows:

$$LW(t_i, d) = \frac{CSW(t_i, d) + SW(t_i, d)}{2} \quad (3)$$

By examining the equation 3, we can notice that the terms (simple or composed) are weighted by the same way. Despite the several works dealing with the weighing of composed terms, there is so far no weighing technique shared by the community [17]. In our approach, we applied the weighing method proposed by [17]. For a term t composed of n words,

its frequency in a document depends on the frequency of the term itself, and the frequency of each sub-term. For this purpose, it proposes the measure cf is defined as follows:

$$cf(t, d) = f(t, d) + \sum_{st \in subterms(t)} \frac{length(st)}{length(t)} \cdot f(st, d) \quad (4)$$

where,

- $f(t, d)$: the occurrences number of t in the document d .
- $Length(t)$ represents the number of words in the term t .
- $subterms(t)$ is the set of all possible terms MeSH which can be derived from t .

For example, if we consider a term “cancer of blood”, knowing that “cancer” is itself also a MeSH term, its frequency is computed as:

$$cf(cancer\ of\ blood) = f(cancer\ of\ blood) + \frac{1}{2} \cdot f(cancer)$$

Consequently, in an attempt to take into account the case of composed terms, we calculate the csw measure as follows:

$$CSW(t_i, d) = \frac{\sum_{A \in T, K, A, P} f(t_i, d, A) \times W_A}{\sum_{A \in T, K, A, P} f(t_i, d, A)} + \sum_{st \in subterms(t_i)} \frac{length(st)}{length(t_i)} \cdot f(st, d) \quad (5)$$

where, $f(st, d)$ is the occurrences number of st in the document d .

It's important to note that in the case of simple terms, $subterms(t_i) = \emptyset$. Consequently the formulas presented by (5) and (1) are equivalent.

Finally, the weight of a term t_i in a document d_j ($Weight(t_i, d_j)$) is calculated as follows:

$$Weight(t_i, d_j) = LW(t_i, d_j) \cdot \ln(N/df) \quad (6)$$

where,

- N : the total number of documents,
- df (document frequency): the number of documents which term t_i occurs in.

E. Generation of Best Descriptors MeSH

A term MeSH may be located in different hierarchies at various levels of specificity, which reflects its ambiguity. As an illustration, Fig. 7 depicts the term “Pain”, which belongs to four branches of three different hierarchies (descriptors) whose the most generic descriptors are: Nervous System Disease (C10); Pathological Conditions, Signs and Symptoms (C23); Psychological Phenomena and Processes (F02); Musculoskeletal and Neural Physiological Phenomena (G11).

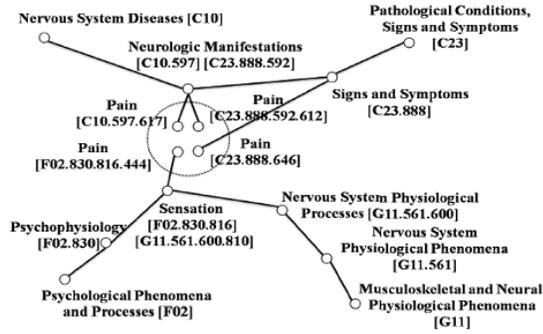


Fig. 7. Term Pain in MeSH.

In the last years, due to the amount of ambiguous terms and their various senses used in biomedical texts, term ambiguity resolution becomes a challenge for several researchers [1] [18]. For an ambiguous term, the task of WSD consists in answering the question: among its several senses, which is the best descriptor that can represent this term. The task of the WSD system is then to estimate, for each candidate descriptor MeSH, which is most likely to be the ideal concept. Differently from the proposed works in the literature, our method assign the appropriate descriptor related to a given term by using the language model approach.

In proposed approach, to determine for an ambiguous term, its best descriptor, we have adapted the language model of [20] by substituting the query by the Mesh descriptor. Thus, we infer a language model for each document and rank Mesh descriptors according to their probability of producing each one given this model. We would like to estimate $P(des|d)$, the probability of generation a Mesh descriptor des given the language model of document d .

For a collection (D), a document (d) and a MeSH descriptor (des) composed of n concepts, the probability $P(des|d)$ is done by:

$$P(des_k|d) =$$

$$\prod_{c_j \in RelatedtoDes(des_k, d)} ((1 - \lambda) \cdot P(c_j|d) + \lambda \cdot P(c_j|D)) \quad (7)$$

RelatedtoDes (respectively *RelatedtoCon*) is the function that associates for a given descriptor des (respectively concept con) and a document d , the concepts (respectively terms) MeSH which are related to des (respectively con) in d .

In this equation, we need to estimate two probabilities:

- $P(c_j|D)$ the probability of observing the concept c_j in the collection D .

$$P(c_j|D) = \frac{\sum_{t_i \in RelatedtoCon(c_j, d)} df(t_i, D)}{\sum_{c' \in D} f(c', D)}$$

$df(t, D)$: df (document frequency) is the number of documents which term t occurs in D .

- $P(c_i|d)$ the probability of observing the concept c_i in document d .

$$P(c|d) = \frac{f(c, d)}{|concepts(d)|}$$

$$f(c_j, d) = \sum_{t_i \in RelatedtoCon(c_j, d)} LW(t_i, d)$$

$LW(t, d)$ is determined by using (3).

Based on this approach, to assign the appropriate sense (Best Descriptor (BD)) related to an ambiguous term (t_i) in the context of document (d_j), we must go through these steps:

- 1) *Compute the descriptor relevance score*
Let

$$senses_{d_j}^i = \{des_{d_j}^{i1}, des_{d_j}^{i2}, \dots, des_{d_j}^{in}\} :$$

the set of descriptors MeSH that can represent the term t_i in the document d_j .

For each descriptor des_k existing in this set, we need to measure its ability to represent the term (t_i) in the document (d_j). To do so, we calculate $P(des_k|d_j)$.

- 2) *Selection of the best descriptor* The best descriptor (BD) to retain is the one which maximizes $P(des_k|d_j)$:

$$BD(t_i, d_j) = \max_{des_k \in senses_{d_j}^i} P(des_k|d_j)$$

Finally, in document's description of document d , we retain its Semantic Index (SI).

$$SI(d_j) = \bigcup_{t_i \in CT(d_j)} BD(t_i, d_j)$$

F. Extracting Temporal Expressions

This step extracts all temporal expressions in document, including the explicit temporal expressions and the implicit temporal expressions.

- **Explicit temporal expressions:**
These temporal expressions directly describe entries in some timeline, such as an exact date or year. For example, the token sequences "December 2017" or "September 12, 2011" in a document are explicit temporal expressions and can be mapped directly to chronons in a timeline.
- **Implicit temporal expressions:**
These temporal expressions represent temporal entities that can only be anchored in a timeline in reference to another explicit or implicit, already anchored temporal expression. For example, the expression "last Friday" or "next week" alone cannot be anchored in any timeline.

To extract explicit temporal expressions, we employ the GUTime tool [22]. The implicit temporal expressions are extracted by using the method presented in [27].

TABLE II. PRECISION (P) AND MAP(MEAN AVERAGE PRECISION) GENERATED BY THE INDEXING SYSTEMS USING TITLE AS INPUT

System	(P@10)	MAP
MTI	0.18	0.16
MetaMap	0.17	0.14
EAGL	0.18	0.17
KNN	0.43	0.47
TempIndex	0.40	0.45

TABLE III. PRECISION AND MAP (MEAN AVERAGE PRECISION) GENERATED BY THE INDEXING SYSTEMS USING TITLE AND ABSTRACT AS INPUT

System	(P@10)	MAP
MTI	0.32	0.25
MetaMap	0.19	0.16
EAGL	0.21	0.19
KNN	0.45	0.50
TempIndex	0.33	0.28

G. Generation of Best Temporal Expressions

The score of a temporal expression as a combination of an explicit score and an implicit score. The explicit score is related to the term frequency of a temporal expression, and accordingly the implicit score is related to the contribution made by all its children expressions [27]. The score of T_i , denoted as $Score(T_i)$, is the sum of its explicit score, denoted as $ES(T_i)$, and its implicit score, denoted as $IS(T_i)$.

$$ES(T_i) = TF_{ETE}(T_i) + d * TF_{ITE}(T_i)$$

$TF_{ETE}(T_i)$ refers to the term frequency of the explicit temporal expressions which are recognized as T_i .

$TF_{ITE}(T_i)$ refers to the term frequency of the implicit temporal expressions which are calculated as T_i . d is the weighting factor, if d is set to 1, it means that the explicit and implicit temporal expression have the same credible level; if d is set to 0, it means that we take no account of implicit temporal expressions. Finally, the N^4 top-ranked temporal expressions with the highest score are selected in document's description.

VI. COMPARISON OF PROPOSED SYSTEM WITH OTHERS INDEXING SYSTEMS

To prove the effectiveness of our indexing method, we compared system (TempIndex) to other medical indexing systems. We evaluate the performance of five indexing systems (MetaMap, EAGL, KNN, MTI and TempIndex) in terms of generating the manual MeSH annotations. For this evaluation, we used the same corpus⁵ used by [28] composed of 1000 random MEDLINE citations.

Table II shows the results generated by indexing systems using the title of a 1000 random MEDLINE citations.

Table III shows the results generated by indexing systems with the title and abstract of a 1000 random MEDLINE citations.

Fig. 8 illustrates the obtained results by the five indexing systems on the 1000 random MEDLINE citations.

⁴The N value is calculated experimentally

⁵The corpus can be downloaded in ([http](http://www.ebi.ac.uk/triesch/meshup/testset_v1.xml) : //www.ebi.ac.uk/triesch/meshup/testset_v1.xml)

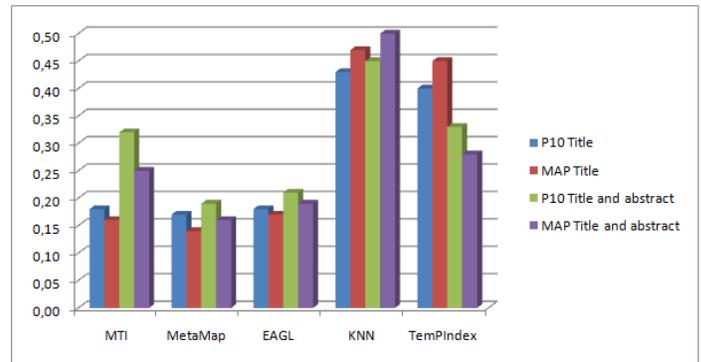


Fig. 8. Experimental results generated by the five indexing systems.

The system TempIndex serves as the baseline against which the other systems are compared. Both indexing systems MetaMap and EAGL perform worse than BIOINSY in all metrics. Indeed, MetaMap performs similarly to or slightly worse than EAGL when presented with the title only or both title and abstract of the citation to index. MTI performs worse than TempIndex when the title was available for indexing. For example, when title used as input, the value of P10 generated by MTI is equal to 0,18. Concerning TempIndex, it generates 0,40 as value of P10. When using title and abstract, MTI performs similarly to or slightly better than TempIndex in terms of MAP and P10. By using title as input, KNN and TempIndex echoed very similar performance. Given the title and abstract of a citation, KNN shows the best results in all metrics. The obtained results confirm the well interest to use the temporal criteria in the indexing process.

VII. RESULTS AND DISCUSSION

In this section, we try to answer the following question: Can proposed temporal indexing approach (described and evaluated above) improve the information retrieval process. The overview of this section is as follows. In subsection 7.1 we will present the test collection. In subsection 7.2 we will describe the experimental setup. In subsection 7.3, the experimental results will be analyzed and discussed.

A. Test Collection

To evaluate the retrieval effectiveness based on our conceptual indexing method, we use the ImageCLEF med 2007 collection⁶. Started from 2004, the ImageCLEFmed (medical retrieval task) aims at evaluating the performance of medical information systems, which retrieve medical information from a mono or multilingual image collection. This corpus [25] is based on a dataset containing images from the Casimage, MIR, PEIR, PathoPIC, CORI, myPACS and Endoscopic collections. For each image of this corpus, a textual description called diagnosis is attributed. This corpus comprises 47680 cases, 66662 images and 55485 Annotations.

B. Experimental Setup

The ImageCLEF data contains the qrels file (TREC format) which specifies the set of relevant images to a given query. In

⁶CLEF (Cross Language Evaluation Forum)

TABLE IV. THE COMPARISON OF OUR SYSTEM WITH OFFICIAL RUNS PARTICIPATED IN IMAGECLEF MED 2007

Run	(P@5)	MAP
LIG-MRIM-LIGMU	0.44	0.32
OHSU	0.42	0.27
IPAL4	0.39	0.27
miracleTxtFRT	0.43	0.17
IRIT RunMed1	0.05	0.04
system TempIndex in experiment 1	0.38	0.24
system TempIndex in experiment 2	0.43	0.33

our indexing method we are interested by the textual document. Hence, to evaluate the proposed approach we assume that “If a query is relevant to an image then it is also relevant to its textual description (diagnosis)”.

This evaluation process is structured around the following steps:

- *Indexing of diagnosis and queries* The indexing process is carried out on the diagnosis and queries. Thus, documents and eventually queries are expanded with MeSH descriptors and temporal expressions identified by our indexing method.
- *Calculation of Retrieval Status Value (RSV (q, d))* The relevance score of the document d_j with respect to the query q is given by

$$RSV(q, d_j) =$$

$$\sum_{des, temp \in q} TF_j(des, temp) * IDF(des, temp)$$

Where,

- TF_j : the normalized term frequency of the current descriptor MeSH (des) or expression temporal (temp) in document d_j .
- IDF : the normalized inverse document frequency of the current descriptor MeSH (des) or expression temporal (temp) in the collection.

C. Results and Discussion

To evaluate the effectiveness of integrating temporal criteria in the indexing process, we carried out two sets of experiments:

Experiment 1: Indexing without temporal criteria: indexing process consists of four main steps: (a) Pretreatment (b) term extraction (c) term weighing and (d) generation of best descriptors MeSH.

Experiment 2: Indexing with temporal criteria: indexing process consists of five main steps: (a) Pretreatment (b) Extracting MeSH concepts (c) Extracting temporal expressions (d) Generation of Best descriptors MeSH and (e) Generation of best temporal expressions.

We had compared the results of the indexing system TempIndex to official runs in medical retrieval task 2007 discussed as follows:

Table IV summarizes the results obtained by the participants in medical retrieval task 2007 and system TempIndex in experiment 1 and experiment 2.

In order to make clear these experimental results, we propose Fig. 9 which presents the precision and MAP value generated by each system. By examining this figure, we can note that the results generated by our system (even without using temporal criteria) close to those of the best run (LIG-MRIM-LIGMU).

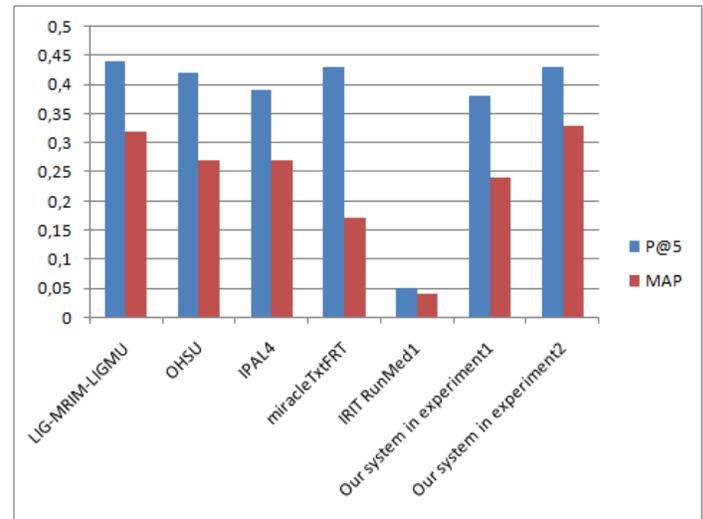


Fig. 9. Experimental results generated by the participants in medical retrieval task 2007 and system TempIndex.

As shown in Fig. 9, our temporal indexing approach (experiment 2) is really significant compared to the classical indexing approach (experiment 1). The obtained results confirm the well interest to integrate the temporal criteria in the indexing process. For instance, our system displayed 0.38 as precision in the case of experiment 1 and 0.43 in the experiment 2. Thus, we conclude that our temporal indexing approach proposed in this paper would significantly improve the IR performance.

VIII. CONCLUSION

The work developed in this paper outlined a temporal indexing approach using the Mesh thesaurus for representing the semantic content of medical articles. Our proposed approach consists of three main steps. At the first step (Term extraction), being given an article, MeSH thesaurus and the NLP tools, the system TempIndex extracts the article’s lemma. After that, these sets are used in order to extract the Mesh terms existing in the document. At step 3, these extracted terms are weighed by using the measures CSW and SW that intuitively interprets MeSH conceptual information to calculate the term importance. The step 4 aims to recognize the MeSH descriptors that represent the document by using the language model. At step 5, the system TempIndex extracts the list of temporal expressions. This list is used in step 6 to determine the best temporal expressions.

In order to assess its feasibility, our indexing approach was experimented on through training data sets containing 1000 random MEDLINE citations. An experimental evaluation and comparison of our system with others indexing tools confirms the well interest to use the temporal criteria in the indexing process.

Our future work aims at incorporating a kind of temporal smoothing into the language modeling approach.

ACKNOWLEDGMENT

We thank the deanship of Scientific Research of Majmaah University to support the research project number 27/40.

REFERENCES

- [1] B. Andreopoulos D. Alexopoulou and M. Schroeder. *Word Sense Disambiguation in biomedical ontologies with term co-occurrence analysis and document clustering*, IJDMB, 2008.
- [2] R. Campos G. Dias AM. Jorge and A.Jatowt. *Survey of temporal information retrieval and related applications*, ACM Computing Surveys 2014; 47(2): 15:115:41.
- [3] J. Gamet. *Indexation de pages web*, Report of DEA, universit de Nantes, 1998.
- [4] J. Majdoubi H.Loukil M.Tmar and F.Gargouri. *Using the Mesh Thesaurus to Index a Medical Article:Combination of Content, Structure and Semantics*, In KES Journal 16-4 pp. 278285, 2009.
- [5] J. Majdoubi H.Loukil M.Tmar and F.Gargouri. *An approach based on langage modeling for improving biomedical information retrieval*, In KES Journal,16-4 pp. 235-246, 2012.
- [6] J. Majdoubi H.Loukil M.Tmar and F.Gargouri. *Medical Case-based Retrieval by Using a Language Model: MIRACL at ImageCLEF 2012*, Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012.
- [7] J. Majdoubi M.Tmar and F.Gargouri. *Thesaurus based Semantic Representation in Language Modeling for Medical Article Indexing*, In 2th International Conference on Enterprise Information Systems, Volume 2, AIDSS, Funchal, Madeira, Portugal, pp. 65-74, June 8 - 12, 2010
- [8] W. Kim A. Aronson and W. Wilbur. *Automatic MeSH term assignment and quality assessment*, in: AMIA, 2001.
- [9] D. Dinh and L. Tamine. *Biomedical concept extraction based on combining the content-based and word order similarities*, in: SAC, 2011, pp. 11591163.
- [10] W. Dakka L. Gravano and P. Ipeirotis. *Answering general time-sensitive queries*. In CIKM08, 2008.
- [11] X. Li and W. B. Croft. *Time-based language models*. In CIKM03, 2003.
- [12] K. Massoudi E. Tsagkias M. de Rijke and W. Weerkamp. *Incorporating query expansion and quality indicators in searching microblog posts*. In ECIR11, 2011.
- [13] D. Metzler R. Jones F. Peng and R. Zhang. *Improving search relevance for implicitly temporal queries*, 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR 09. New York, NY, USA: ACM, 2009, pp. 700701.
- [14] A. neveol. *Automatisation des taches documentaires dans un catalogue de santé en ligne*, Institut National des Sciences Appliquées de Rouen, 2005.
- [15] A.Aronson J. Mork C.Gay S.Humphrey and W.Rogers. *The NLM Indexing Initiatives Medical Text Indexer*, in: Medinfo, 2004.
- [16] A. Aronson. *Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program*, in: AMIA, 2001, pp. 17-21.
- [17] M. Baziz. *Indexation conceptuelle guide par ontologie pour la recherche dinformation*. PhD thesis, Univ. of Paul sabatier (2006).
- [18] B. Dinh and L. Tamine. *Sense-based biomedical indexing and retrieval*. In: NLDB. (2011) pp- 2435.
- [19] M. Efron and G. Golovchinsky. *Estimation Methods for Ranking Recent Information*. In SIGIR11, 2011.
- [20] Hiemstra, D. *Using Language Models for Information Retrieval*. PhD thesis, University of Twente (2001).
- [21] Yu. PS X. Li and B.Liu. *On the temporal dimension of search*. In: Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers & Posters. WWW Alt. 04; New York, NY: ACM, 2004, pp.448449.
- [22] I. Mani and G.Wilson. *Robust Temporal Processing of News*, In Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, 2000, pp. 69–76
- [23] K. Mathews and S. Deepa Kanmani. *A Survey on Temporal Information Retrieval Systems*, International Journal of Computer Applications, November 2012, pp 24-28.
- [24] D. Metzler C. Cai and E. Hovy. *Structured Event Retrieval over Microblog Archives*, NAACL-HLT 12, 2012.
- [25] H. Miller T. Deselaers T. Deserno and P.Clough, E.Kim and W. Hersh. *Overview of the imageclefmed 2006 medical retrieval and annotation tasks*. In: In: CLEF 2006 Proceedings. Lecture Notes in Computer Science (2006) pp- 595608.
- [26] M-H. Peetz E. Meij M. de Rijke and W. Weerkamp. *Adaptive Temporal Query Modeling*, In ECIR12, 2012.
- [27] Sheng Lin Peiquan Jin Xujian Zhao and Lihua Yue. *Exploiting temporal information in Web search*, *Expert Systems with Applications*, 2014, pp. 331 - 341.
- [28] D.Trieschnigg P.Pezik V.Lee W.Kraaij F. Jong and D.Rebholz-Schuhmann. *MeSH Up: Effective MeSH Text Classification and Improved Document Retrieval*. *Bioinformatics* 25(11), (2009), pp- 14121418.
- [29] S. Aubin and T. Hamon. *Improving Term Extraction with Terminological Resources*, *Advances in Natural Language Processing*, 2006.

A Comparison of Predictive Parameter Estimation using Kalman Filter and Analysis of Variance

Asim ur Rehman Khan, Haider Mehdi, Syed Muhammad Atif Saleem, Muhammad Junaid Rabbani
Multimedia Labs,
National University of Computer and Emerging Sciences (NUCES),
Pakistan

Abstract—The design of a controller significantly improves if internal states of a dynamic control system are predicted. This paper compares the prediction of system states using Kalman filter and a novel approach analysis of variance (ANOVA). Kalman filter has been successfully applied in several applications. A significant advantage of Kalman filter is its ability to use system output to predict the future states. It has been observed that Kalman filter based predictive controller design outperforms many other approaches. An important drawback of such controllers is however that their performances deteriorate in situations where the system states have no correlation with the output. This paper takes a hypothetical model of a helicopter and builds system model using the state-space diagram. The design is implemented using SIMULINK. It has been observed that in situations where the states are dependent on system output, the ANOVA based state prediction gives comparable results with that of Kalman filter based parameter estimation. The ANOVA based parameter prediction, however outperforms Kalman filter based parameter prediction in situations where the output does not directly contribute in a particular state. The research was based on empirical results. Rigorous testing was performed on four internal states to prove that ANOVA based predictive parameter estimation technique outperforms Kalman based parameter estimation in situations where the system internal states is not directly linked with the output.

Keywords—Analysis of variance (ANOVA); Kalman controllers; predictive controller

I. INTRODUCTION

A linear time-invariant (LTI) continuous system can be inherently stable if all of its poles are on the left-hand side of s-plane. If, however some of the poles are on the right-hand side then it needs a controller to ensure that the poles in the right-hand side are cancelled by the zeros of the controller, making an inherently unstable system to become stable. Alternately a system may be inherently stable, but at higher gains its poles may move towards the right-hand side of s-plane. In case of a discrete system, the condition of stability requires the presence of poles within a unit circle resulting in a stable system. This research compares the estimates of Kalman filter based predictive parameter estimates with that of the ANOVA based predictive parameter estimates. The controllers are generally categorized as feedback controllers, adaptive controllers, and predictive controllers. Among these controllers, the predictive controller influences the activity of the system to adjust various parameters to achieve the targeted value. The system tracks output such that the difference

between the desired and actual output remains within limits as per given matrix. The predictive controller uses current output/states to adjust the parameters of system to change the future output/states. The controller is based on proactive approach. The time-series analysis predictive controller performs reasonably well in case of where the variations are relatively free from noise. One of known time series predictor is an auto-regressive (AR) controller. The prediction based on analysis of variance (ANOVA) is relatively new in predictive controller design. A significant advantage of using ANOVA is that the noise carried by the parametric variation is also accounted for in the model. A combination of auto-regressive model and ANOVA are successfully used to predict the computer utilization in an internet service provider (ISP) [1].

An adaptive controller design based on Kalman filter has provided an optimum control design during the last many decades. The implementation of Kalman filter observer for multivariable ship control system is discussed in [2]. In addition, the application of extended Kalman filter observer to estimate the state of time varying disturbance for robotic manipulator and industrial heating system is presented in [3]-[4]. An adaptive Kalman filter for state-of-charge (SOC) lithium-ion battery is discussed in [5]. Several other Techniques of predicting states using the Kalman filter have been discussed in [6]-[8].

This system uses more sophisticated state-space model to monitor the system states. If fast moving applications, the predictive controller performs better than the adaptive controller simply due to the fact that they predict future parameters based on past records. The proposed algorithm uses analysis of variance techniques (ANOVA) to predict the future states. The results further show that in certain situations the ANOVA based parameter prediction outperforms that of Kalman filter based parameter estimation. Predictive controllers based on ANOVA can be used in real time control applications. To the best of our knowledge, no work has been done in this area. Future work may involve using other statistical techniques, like regression analysis, etc.

After a brief introduction in this section, the state-space model in continuous and discrete system of a hypothetical helicopter is discussed in Section 2. The Kalman filter based observer and parameter estimation is given in Section 3. The ANOVA based parameter estimation is given in Section 4. Section 5 gives several results that compare the estimated parameters using Kalman filter and the ANOVA based approaches. Section 6 concludes this paper.

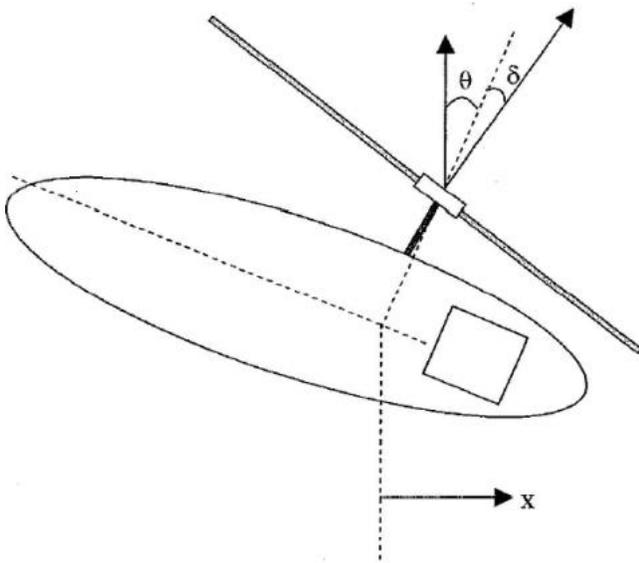


Fig. 1. A conceptual model of helicopter.

II. MATHEMATICAL MODELS

The model essentially consists of state-space model of a helicopter as in Fig. 1. The helicopter is expected to move only in the horizontal direction. The change in horizontal direction $x(t)$ is controlled by the input $\delta(t)$. The tilt in the horizontal direction is given by the angle $\theta(t)$. The model is described by using two second order equations.

$$\begin{aligned} \ddot{\theta} &= -\tau_1 \dot{\theta} - \alpha_1 \dot{x} + \vartheta_1 \delta \\ \ddot{x} &= g\theta - \alpha_2 \dot{\theta} - \tau_2 \dot{x} + g\delta \end{aligned} \quad (1)$$

Where the first equation gives horizontal tilt, and the second equation gives helicopter position in the horizontal direction. All other parameters are coefficients with known and fixed values as given by

$$\begin{aligned} \tau_1 &= 0.415 & \alpha_1 &= 0.0111 & \vartheta_1 &= 6.27 \\ \tau_2 &= 0.019 & \alpha_2 &= 1.43 & g &= 9.81 \end{aligned} \quad (2)$$

A. Continuous Time State-Space Model

The state-space model of a general system is represented by

$$\begin{aligned} \dot{x} &= Ax + Bu \\ y &= Cx + Du \end{aligned} \quad (3)$$

Where, ' x ' is the set of system states of a helicopter. The ' u ' is given input signal, and ' y ' is the output signal. The **A**, **B**, **C**, and **D** are parameter of state-space model giving the system's characteristics. The bold letters represent matrices. The angle of tilt θ is represented by system state x_1 . The derivative of tilt angle ($\dot{\theta}$) is equal to x_2 . The horizontal position x is represented by x_3 , and the derivative of horizontal position \dot{x} is represented by x_4 .

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} \theta \\ \dot{\theta} \\ x \\ \dot{x} \end{bmatrix} \quad (4)$$

The input u and the output state vector is defined as

$$u = \delta, \text{ and } \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \theta \\ x \end{bmatrix} \quad (5)$$

Where, δ is the amount of input signal, and the expected output y_1 , and y_2 are the amount of horizontal motion and the amount of tilt. The state equations of a helicopter are given by the following set of equations:

$$\begin{aligned} \dot{x}_1 &= \dot{\theta} = x_2 \\ \dot{x}_2 &= \ddot{\theta} = -0.415x_2 - 0.0111x_4 + 6.27\delta \\ \dot{x}_3 &= \dot{x} = x_4 \\ \dot{x}_4 &= \ddot{x} = 9.81x_1 - 1.43x_2 - 0.0198x_4 + 9.81\delta \end{aligned} \quad (6)$$

These are represented in more compact form using matrices:

$$\begin{aligned} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} &= \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & -0.415 & 0 & -0.0111 \\ 0 & 0 & 0 & 1 \\ 9.81 & -1.43 & 0 & -0.0198 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \\ &+ \begin{bmatrix} 0 \\ 6.27 \\ 0 \\ 9.81 \end{bmatrix} \delta + v(k) \end{aligned} \quad (7)$$

The $v(k)$ represents undesired effects and this is approximately by a zero mean, normally distributed signal with constant variance. The output is given by

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + w(k) \quad (8)$$

The **D** matrix in the standard state-space model is considered to be zero. The $w(k)$ is due to an undesired effect at the input, again approximated by a normally distributed signal having zero mean, and constant variance.

B. State Space Model in Discrete Time

The continuous state-state model is transformed into discrete state-space mode by using MATLAB routine. The sampling rate is $h = 0.1$. A relatively smaller sampling rate results in large number of samples, and thus gives more accurate replica of the continuous state-space model. The general form of discrete state-space model is given by,

$$\begin{aligned} x(k+1) &= Fx(k) + Gu(k) + v(k) \\ y(k) &= Cx(k) + w(k) \end{aligned} \quad (9)$$

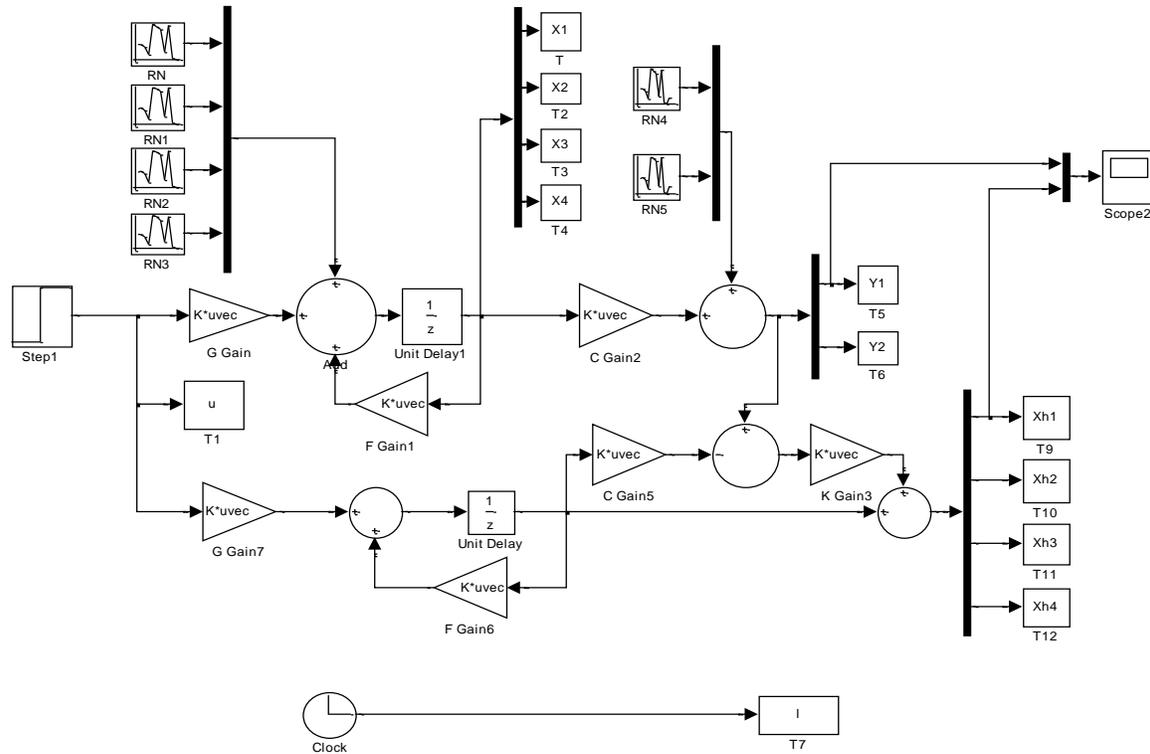


Fig. 2. SIMULINK model with Kalman filter observer.

The D matrix is again considered as zero. The $v(k)$ and $w(k)$ are corresponding process and measurement noises. The respective matrices are given by

$$F = \begin{bmatrix} 1.0000 & 0.0980 & 0 & -0.0001 \\ -0.0005 & 0.9594 & 0 & -0.0011 \\ 0.0490 & -0.0054 & 1.0000 & 0.0999 \\ 0.9801 & -0.0916 & 0 & 0.9981 \end{bmatrix} \quad (10)$$

$$G = \begin{bmatrix} 0.0309 \\ 0.6136 \\ 0.0478 \\ 0.9460 \end{bmatrix} \quad C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

This model is given in Fig. 2 using SIMULINK.

III. KALMAN FILTER OBSERVER

The Kalman filter observer uses output signal to estimate various system states. A predictor based on these estimated state values can be designed to control the future states to minimize the amount of error. The estimated state has two components, a predicted state and a correction term based on the present output.

$$\hat{x}(k) = \bar{x}(k) + K\{y(k) - C \bar{x}(k)\} \quad (11)$$

Where, $\hat{x}(k)$ is the estimated value at sample k, $\bar{x}(k)$ is the predicted state, and $K\{y(k) - C \bar{x}(k)\}$ is the correction term. K is the gain matrix. The predicted state $\bar{x}(k)$ is

$$\bar{x}(k) = F\hat{x}(k-1) + G u(k-1) \quad (12)$$

The residual error is

$$e(k+1) = x(k+1) - \hat{x}(k+1) \quad (13)$$

The objective is to reduce the residual error. A system is controllable only if this is observable. The system is tested for observability and controllability using MATLAB routines. The positions of poles are also checked. It is found that some of the poles are outside the unit circle. The system is inherently unstable however with an appropriate controller design the system will become stable as all poles outside the unit circle are cancelled by zeros of controller. The Kalman gain matrix K is found using the MATLAB routine DLQE (discrete-linear-quadratic-estimator) command.

$$[K, P] = DLQE (F, H, C, R_v, R_w) \quad (14)$$

Where, K is the Kalman gain. The H is 3x3 identity matrix. F and C are previously defined. R_v is the covariance matrix of system disturbance $\vartheta(k)$, and R_w is the covariance matrix of output disturbance $w(k)$. These matrices are taken as,

$$R_v = \begin{bmatrix} 0.01 & 0 & 0 & 0 \\ 0 & 0.01 & 0 & 0 \\ 0 & 0 & 0.01 & 0 \\ 0 & 0 & 0 & 0.01 \end{bmatrix} \quad R_w = \begin{bmatrix} 0.01 & 0 \\ 0 & 0.01 \end{bmatrix} \quad (15)$$

TABLE I. PERCENTAGE CHANGES IN THE FOUR PARAMETERS USING KALMAN FILTERING

S.No	Samples	x1			x2			x3			x4		
		Actual	x1_k	Error	Actual	x2_k	Error	Actual	x3_k	Error	Actual	x4_k	Error
1	25	11.8	9.4	2.4	5.6	-9.5	15.0	149.4	104.7	44.7	166.2	148.8	17.3
2	50	30.7	27.9	2.7	6.2	-83.5	89.7	1114.8	887.8	227.1	672.6	980.4	-307.8
3	75	25.3	31.4	-6.1	-6.8	-186.2	179.4	3672.7	3236.5	436.2	1362.7	2335.4	-972.6
4	100	-5.6	0.7	-6.3	-19.7	-217.6	197.9	7656.3	7188.3	468.0	1697.2	3044.5	-1347.4
5	125	-61.1	-58.3	-2.7	-19.1	-123.4	104.2	11180.4	10950.9	229.5	893.1	1927.6	-1034.6
6	150	-89.0	-92.4	3.4	5.1	65.1	-60.0	11157.9	11317.4	-159.5	-1032.8	-1013.0	-19.8
7	175	-13.9	-28.1	14.2	47.7	178.0	-130.3	6437.1	6732.2	-295.1	-2457.4	-3442.1	984.7
8	200	138.8	125.0	13.8	62.4	-43.5	105.9	1249.0	933.2	315.8	-1087.2	-1491.5	404.4
9	225	226.2	240.1	-13.8	0.7	-615.9	616.7	3886.0	2333.4	1552.5	3573.7	6154.7	-2581.0
10	250	94.6	140.4	-45.8	-103.2	-1019.1	915.9	18898.0	16724.5	2173.5	7947.6	13948.9	-6001.2

TABLE II. PERCENTAGE CHANGES IN THE FOUR PARAMETERS USING ANOVA

S.No	Samples	x1			x2			x3			x4		
		Actual	x1_a	Error	Actual	x2_a	Error	Actual	x3_a	Error	Actual	x4_a	Error
1	25	11.8	9.8	2.0	5.6	3.5	2.0	149.4	134.2	15.2	166.2	159.0	7.1
2	50	30.7	28.8	1.9	6.2	4.5	1.7	1114.8	1065.9	48.9	672.6	654.8	17.8
3	75	25.3	25.5	-0.2	-6.8	-6.8	0.0	3672.7	3590.6	82.1	1362.7	1352.6	10.1
4	100	-5.6	-5.9	0.3	-19.7	-21.2	1.5	7656.3	7568.6	87.7	1697.2	1707.8	-10.6
5	125	-61.1	-61.3	0.3	-19.1	-20.6	1.4	11180.4	11160.9	19.5	893.1	931.3	-38.3
6	150	-89.0	-89.7	0.7	5.1	5.0	0.1	11157.9	11255.3	-97.4	-1032.8	-991.4	-41.4
7	175	-13.9	-16.8	2.8	47.7	47.8	-0.1	6437.1	6578.0	-140.8	-2457.4	-2471.7	14.3
8	200	138.8	135.0	3.7	62.4	62.9	-0.5	1249.0	1248.7	0.4	-1087.2	-1186.4	99.2
8	225	226.2	227.9	-1.7	0.7	3.0	-2.3	3886.0	3595.5	290.4	3573.7	3458.0	115.7
10	250	94.6	104.1	-9.5	-103.2	-99.4	-3.7	18898.0	18432.6	465.4	7947.6	7945.2	2.4

The Kalman gain K is found to be equal to

$$K = \begin{bmatrix} 0.6405 & 0.0006 \\ 0.3695 & -0.1370 \\ 0.0006 & 0.6683 \\ 0.1669 & 0.8703 \end{bmatrix} \quad (16)$$

The state estimators using Kalman gain is found by

$$\begin{bmatrix} \hat{x}_1(k+1) \\ \hat{x}_2(k+1) \\ \hat{x}_3(k+1) \\ \hat{x}_4(k+1) \end{bmatrix} = \begin{bmatrix} \bar{x}_1(k) \\ \bar{x}_2(k) \\ \bar{x}_3(k) \\ \bar{x}_4(k) \end{bmatrix} + G(k) \quad (17)$$

$$G(k) = \begin{bmatrix} 0.6405 & 0.0006 \\ 0.3695 & -0.1370 \\ 0.0006 & 0.6683 \\ 0.1669 & 0.8703 \end{bmatrix} \begin{Bmatrix} y_1 \\ y_2 \end{Bmatrix} - \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{Bmatrix} \bar{x}_1(k) \\ \bar{x}_2(k) \\ \bar{x}_3(k) \\ \bar{x}_4(k) \end{Bmatrix}$$

The system design including the Kalman filter observer is given in Fig 2.

IV. ANOVA PREDICTOR DESIGN

The regression based predictive modeling has been used in several fields. The human behavior is predicted based on the

known facts. The statistical based prediction has been used in marketing, financial services like banking & insurance. It has also been used in the telecommunications industry. A software package, Statistical Analysis System (SAS) has been developed by SAS Institute that helps in the advance analysis like multivariate analyses, data management, business intelligence, and predictive analysis. The package also uses ANOVA to provide the necessary analysis. An ANOVA based prediction of moisture buildup in electronic enclosure is proposed in [9]. The prediction of dataset by software engineers using ANOVA is demonstrated in [10]. An Intrusion Detection System by feature elimination has been demonstrated in [11]. ANOVA based parameter prediction is also demonstrated in [12].

This paper is an enhanced version of a recently published conference paper [13]. The analysis of variance (ANOVA) based proposed model is given by,

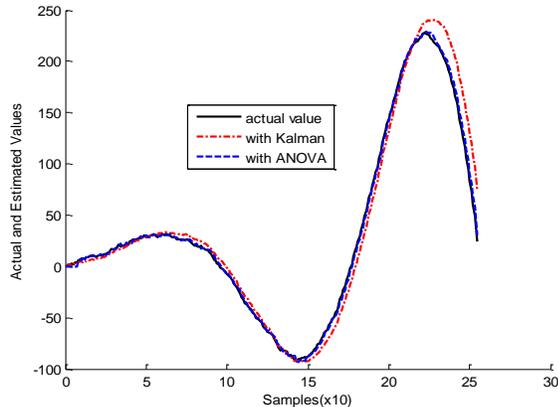
$$\hat{x}(k) = \hat{\mu}_k + \hat{\alpha}_k + \hat{\beta}_k + \epsilon(k) \quad (18)$$

where $\hat{x}(k)$ is the estimated value of k^{th} sample. This is approximated by an estimated general mean $\hat{\mu}_k$, sample parameters $\hat{\alpha}_k$, and $\hat{\beta}_k$. The unknown effects are given in the last term $\epsilon(k)$. The sample mean and the other parameters are found using

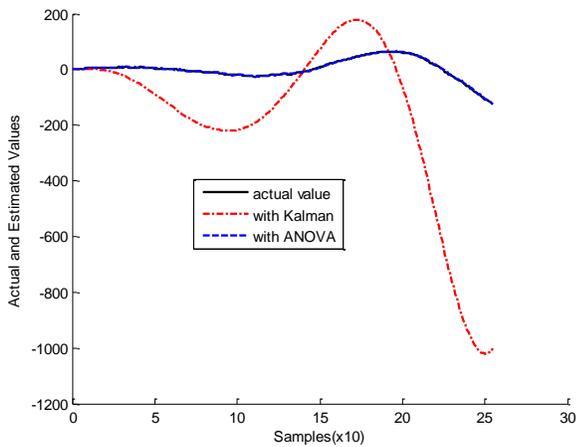
$$\hat{\mu}_k = \frac{1}{8} \sum_{i=(k-8)}^{k-1} x(i)$$

$$\hat{\alpha}_k = \frac{1}{7} \sum_{i=(k-7)}^{k-1} [x(i) - x(i-1)]$$

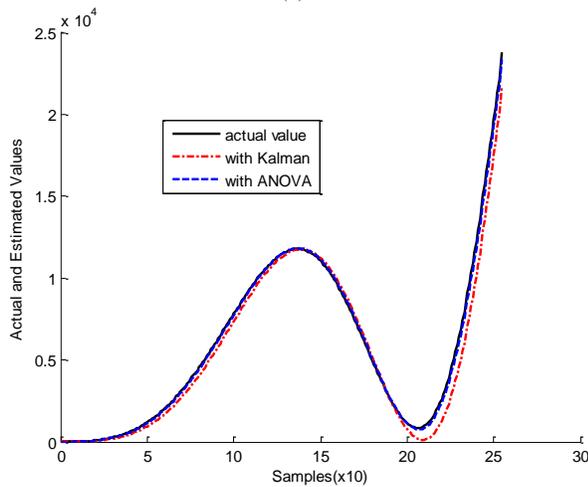
$$\hat{\beta}_k = \frac{1}{2} \{ [x(k-1) - x(k-4)] + [x(k-5) - x(k-8)] \}$$
(19)



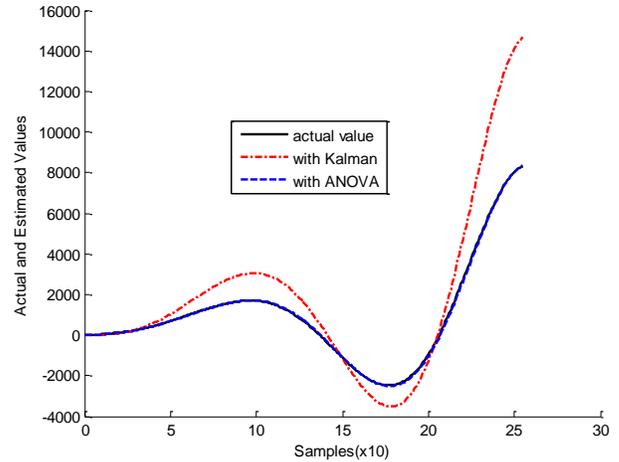
(a)



(b)



(c)



(d)

Fig. 3. (a)-(d) Actual and Estimated values of parameters, x_1 , x_2 , x_3 , and x_4 .

The above equation extracts three parameters based on sample mean of past eight samples $\hat{\mu}_k$, the mean difference of past eight samples by $\hat{\alpha}_k$, and the sample mean of past four samples with its preceding four samples.

V. SIMULATION RESULTS

The main objective of this work is the prediction of four states that represent the motion of a helicopter. This is possible only due to state-space analysis. The states are initially in continuous domain which is converted into discrete model using the sampling rate of 0.1. The sampling rate is critical in the sense that a smaller sample time results in more accurate analog-digital conversion, but results in much larger number of samples. Alternately a larger sampling rate results in coarse analog-digital conversion but a fewer samples. The most appropriate sampling rate depends on the application. It has been observed that a sampling rate of 0.1 results in most appropriate combination of refinement, and amount of data for a helicopter model. The output signal comprises of two parameters: an angle of tilt (θ) and the horizontal motion (x) based on four internal states. These states are given as x_1, x_2, x_3 , and x_4 as given in (4). The first step is the conversion of continuous state-space model into a discrete state-state model (9). The MATLAB routines are used to find matrices F, G, and C in discrete state-space model (10). These matrices, respectively, correspond to A, B, and C in continuous state-space model. Next it is desired to find if the system needs a controller, and if it is controllable. A discrete system is inherently stable if all of its poles are inside the unit circle. Such a system may only need a very simple controller that monitors system gain and ensures that the poles remain within the unit circle. If however, one or more poles are outside the unit circle then the system is inherently unstable, and sophisticated controller is required to cancel all poles that are outside the unit circle. It was found that the proposed model is not inherently stable, as it has one pole outside the unit circle. A second challenge is to learn if the system is indeed controllable. A system is controllable only if the internal states are observable. The validation of the observability is performed with a series of tests (14)-(17). The

model is implemented by using SIMULINK. The original equations are based on a second order system. A significant advantage of using state-space analysis is that the model is implemented by using a single delay element, thus reducing the system order to one-dimension model. Other advantages of using discrete signal are the requirement of much smaller storage space, and the flexibility of using advanced signal processing techniques.

The original model and the Kalman filter estimate of states are given in Fig. 2. The actual values of selected samples, the estimated values using Kalman filter, and the error is given in Table 1. The graphically plots are given in Fig. 3(a)-(d). It is clear that the two states x_1 , and x_3 are estimated with good accuracy. However, the other two states x_2 , and x_4 are not estimated with good accuracy. The main reason is the fact that Kalman filter observer uses output signals y_1 , and y_2 for the prediction of x_1 , and x_3 . It however, does not use output y_1 and y_2 for the prediction of other two states x_2 , and x_4 . This inherent shortcoming in Kalman based parameter estimator results in poor performance in x_2 , and x_4 .

The performance of ANOVA based estimator design is given in Fig. 3, where the estimators of states x_1, x_2, x_3 , and x_4 are super imposed on the previous drawn estimates using Kalman filter. Each estimator is approximate by a time-series analysis. The analysis is based on a general mean, mean value of a first order gradient, and the mean value of gradient of a group of values as given in (18). The estimated values specific parameters are found using (19). The number of samples in this particular example is restricted to only 8 samples. It is expected that sampling time of 0.1 seconds in this example is small enough to highlight the minute change with sufficient number of samples. It is clear that any change in sampling time would change the matrices F, and G which would change the Kalman gain matrix K, resulting in completely new model. The actual values of states, the estimated values using ANOVA, and the corresponding error of selected samples are given in Table 2. The graphical plots of these samples are given in Fig. 3(a)-(d). A comparison of Tables 1 and 2, and Fig. 3 clearly shows that both Kalman filter and ANOVA based parameter prediction performs reasonable well for states x_1 , and x_3 , however, the performance of ANOVA based parameter prediction is much superior to that of Kalman filter based parameter predictive estimation.

The list of parameters in an ANOVA design can be changed to suit a particular analysis. As an example, in this particular analysis the objective of this model was to have an estimator based on the mean value, and the gradients at two levels. The model was relatively simple as the estimator was based on the aggregation of three independent terms. This model can be further simplified by considering only the mean value, or made complicated by considering terms related to multiple effects.

A significant advantage of ANOVA based estimator is that these estimators are robust and are able to correctly estimation even if the states are corrupted with high degree of normally distributed noise.

VI. CONCLUSIONS

This paper presents a model to estimate the internal states of a helicopter using analysis of variance (ANOVA). The results are compared with parameter estimation using Kalman filter. It has been observed that both the approaches yield comparable results when states have some form of dependencies on the system output. The ANOVA based approach however outperforms Kalman filter approach in situations where the internal states do not depend on the system output. The model is implemented on SIMULINK.

REFERENCES

- [1] Wei Xu, Xiaoyun Zhu, Sharad Singhal, Zhikui Wang "Predictive control for dynamic resource allocation in entrepreneur data centers", IEEE/IFIP Network Operations and Management Symposium NOMS, pp. 115-126, (2006)
- [2] M. Tomera, "Nonlinear observers design for multivariable ship motion control", Polish Maritime Research. 19, Issue Special, pp. 50-56 (2012)
- [3] C. A. Lightcap, S. A. Banks, "An extended Kalman filter for real-time estimation and control of a rigid-link flexible-joint manipulation", IEEE Transactions on Control Systems Technology, no. 1, vol. 8, pp. 91-103 (2010)
- [4] B. Sohlberg, "Gray box modeling for model predictive control of a jeating process", Journal of Process Control, no. 3, vol. 13, pp. 225-238 (2003)
- [5] H. He, R. Xiong, X. Zhang, F. Sun, & J. Fan, "State-of-charge estimation of the lithium-ion battery using an adaptive extended Kalman filter based on an improved thevenin model", IEEE Transaction on Vehicular Technology, no. 4, vol. 60, pp. 1461-1469 (2011)
- [6] Wang, Weiwen, and Gao Zhiqiang. "A comparison study of advanced state observer design techniques", Proceedings of the IEEE American Control Conference, vol. 6 (2003)
- [7] D. Bak, M. Michalik, J. Szafran, "Application of Kalman filter technique to stationary and non-stationary static observer design." IEEE Power Tech Conference Proceedings IEEE Bologna IEEE, vol. 3 (2003)
- [8] G. Heredia, A. Ollero, Mechatronics, 2009, "Sensor fault detection in small autonomous helicopter using observer/Kalman filter identification." IEEE International Conference on ICM (2009)
- [9] P.S. Nasirabadi, et. al. "Semi-empirical prediction of moisture build-up in an electronic enclosure using analysis of variance (ANOVA)", 18th Electronic Packaging Technology Conference, IEEE (2016)
- [10] M. Azzeh, Y. Elseikh, M. Alseid, "An optimized analogy-based project effort estimation", International Journal of Advanced Computer Science and Applications (IJACSA), vol. 5, no. 4, (2014)
- [11] H. Nkiama, S.Z.M. Said, M. Saidu, "A subset feature elimination mechanism for intrusion detection system", International Journal of Advanced Computer Science and Applications (IJACSA), vol. 7, no. 4, (2016)
- [12] J.N. Rouder et. al., "Model comparison in ANOVA" Psychon Bull Rev, Springer, 23:1779-1786, (2016)
- [13] A. R. Khan, M. J. Rabbani, "An ANOVA based predictive parameter estimation", International Conference on Innovations in Electrical Engineering and Computational Technologies (ICIEECT), IEEE, 2017.

Fine-grained Accelerometer-based Smartphone Carrying States Recognition during Walking

Kaori Fujinami, Tsubasa Saeki, Yinghuan Li, Tsuyoshi Ishikawa, Takuya Jimbo, Daigo Nagase, and Koji Sato
Department of Computer and Information Sciences, Tokyo University of Agriculture and Technology
2-24-16 Naka-cho, Koganei, Tokyo 184-8588

Abstract—Due to the dependency of our daily lives on smartphones, the states of the device have impact on the quality of services offered through a smartphone. In this article, we focus on the carrying states of the device while the user is walking, in which 17 states, e.g., in the front-left trouser pocket, calling phone in the right hand, in a backpack are subjects to recognition based on supervised learning with accelerometer-derived features. A large-scale data collection from 70 persons with three walking speeds allows reliable evaluation regarding suitable features and classifiers model, the feature selection method, robustness of localization against unknown person, and effect of walking speed in training a classifier. Person-independent evaluation shows that average F-measures of 17 class classification and merged 9 class classification were 0.823 and 0.913, respectively.

Keywords—Smartphone; on-body localization; accelerometer; machine learning; feature selection; wearable computing

I. INTRODUCTION

Our daily lives heavily depend on smartphones that provides not only phone calling functionality, but also ubiquitous access to the Internet and replacement of objects for specific purposes, e.g., camera, pedometer, etc. as software. Various sensors are embedded into the device, which allows the a system to extract a user's and/or a device's context such as engaging activity [19], [23], [26] and a person/device location [16], [24], identity of pedestrian [28], environmental conditions around a user [8], [10], [15], [27], and so on, which contributes to provide appropriate information/services to a user based on the context.

According to a phone carrying survey, 17% of people determine the position of storing a mobile phone based on *contextual restrictions*, e.g. no pocket in the T-shirt, too large phone size for a pants pocket, comfort for an ongoing activity [4]. These factors are variable throughout the day, and thus users change their positions in a day. This suggests that a context, *on-body device position*, has great potentials in improving the usability of a smartphone and the quality of sensor-dependent services, facilitating human-human communication, the reduction of unnecessary energy consumption, etc. [7]. Note that the position is not an exact 3D coordinate, but the names of the parts of the body, clothes and items to carry the device during walking such as “inside a chest pocket”, “inside a bag”, and “calling (attaching to the ear)”.

In this article, we propose a machine learning-based classifier and classification features to identify 17 storing positions of a smartphone on the body against a segment of data, i.e., window, obtained while a person is walking. The contribution of this article is summarized as follows:

- Classification features suitable for classifying 17 classes are specified, in which we show a subset-based feature evaluation is superior to a collection of individual “good” features.
- We show raw acceleration signal shows better classification performance than linear and vertical component of acceleration signals.
- A large scale user independent classification performance evaluation is presented, in which 70 persons provided acceleration signals of smartphone carrying during walking.
- The effect of heterogeneity of walking speed in training a classifier is evaluated, in which a training dataset with various speed can build more robust classifier than training with single, i.e., normal, speed data only.

The remaining part of this article is organized as follows. In Section II examines related work regarding on-body position sensing. Section III describes about dataset used in this study. The localization method is presented in Section IV, in which the notion of *series* is introduced, and classification features are presented. Section V shows experiments from various aspects including suitable features and classification model, the feature selection method, robustness of localization against unknown person, and effect of walking speed in training a classifier. Finally, in Section VI, we conclude the article.

II. RELATED WORK

On-body position sensing is getting attention to researchers in machine learning and ubiquitous computing communities [7], [25], [29]. A research direction is on the type of a device which is actually realized or intended to be utilized in the future as wearable devices [18], [21], [29] or a smartphone [1], [5]–[7], [12], [14], [22], [25], [30]. The type of a device relates to the selection of target positions. In the wearable device approach, the target positions range from the head to the ankle including fine-grained discrimination such as upper arm vs. forearm and shin vs. thigh [29]. A device is usually attached firmly using a belt or a special mounting fixture. This indicates that the direction of the device might not change so irregularly within a specific activity in a frequent manner, given that small displacement might occur during activities [17]. By contrast, a smartphone terminal is usually stored into containers such as the pockets of jacket, chest and trousers pockets and a wide variety of bags, as well as in a user's hand, hanging from the neck and on a table as surveyed in [4], [30]. In this case, the degree of freedom of irregular movement in a large container, e.g., jacket pocket, handbag,

would increase. Another aspect is the modality of sensing, in which an accelerometer is dominant due to its low power operation and the availability in most commercial smartphones and wearable devices. Incel [14] shows an extensive study on acceleration-based phone localization, in which recognition features are proposed that represent the movement, rotation and orientation of devices during diverse activities of a person such as walking, sitting and biking. Fujinami proposed 63 classifier-independent features for 9 on-body phone positions including bags during walking, which selected based on as what are more predictive of classes and less correlated with each other [7]. Shi et al. [25], Alanezi et al. [1], and Incel [14] utilized a gyroscope in combination with an accelerometer. They reported that the combined approach slightly improved the accuracy [1], [14]; however, considering the power-hungry nature of a gyroscope [32], the improvement would not be the major reason for utilizing a gyroscope.

Regarding the evaluation method, n -fold cross validation is often utilized [1], [12], [18], [22], [25], [29], which utilizes $(n-1)/n$ of dataset for training a classifier and $1/n$ for testing the classifier; it tends to result in good recognition performance because the training dataset may contain $(n-1)/n$ of data from each person in theory, and hence the classifier “knows” about the subjects in advance. By contrast, Leave-One-Subject-Out (LOSO) cross validation is carried out by testing a dataset from a particular person with a classifier that is trained without a dataset from the person. So, LOSO-CV is regarded as a fairer and practical test method, and recently getting attention [1], [7], [14], [30]. To validate the generalization of a classification model, the number of subjects is important, i.e., small number of subjects fail in capturing the characteristic of the population. Incel [14] carried out a performance evaluation using LOSO-CV against an integrated dataset from 35 persons in total; however, the number of persons varies between positions (35 persons for trouser pocket, 25 for backpack, 15 for hand and 10 for messenger bag, jacket, belt and wrist), and the average number is 15.6. Fujinami utilized LOSO-CV using dataset from 20 persons, in which data from 9 positions were equally collected [7]. By contrast, we equally collect data from 17 positions on the body of 70 persons including three states in hands, i.e., swinging, talking on the phone, watching the screen from both left and right hands, as well as carrying items, i.e., bags, which is a unique aspect of our work. In existing work, the type of a bag is not clearly defined [30] or limited to a messenger bag [14], [30]. We consider that the scale of experiment in this article, i.e., 17 positions of 70 persons, is the largest one in the literature.

III. DATASET

A. Target Positions

We targeted 11 popular positions as shown in Fig. 1, among which both the left and right sides of three types of “hand”, trousers front/back pockets, and jacket pockets were collected. Three type of “hand” correspond to calling, watching the screen in the portrait direction and swinging during walking. In total, 17 classes are defined and analyzed.

B. Sensor Modality

The three-axis accelerometer employed in this study is a primary sensor embedded into almost all of today’s smart-

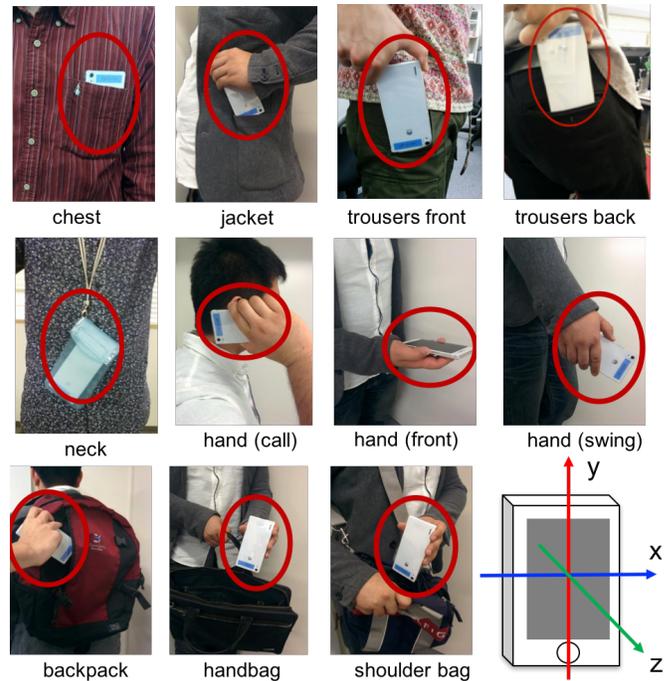


Fig. 1. Target storing positions.

phones. The signals can be used to characterize the movement patterns generated while a person is walking. Although the combination of an accelerometer and a gyroscope slightly improved the classification accuracy [1], [14], because a gyroscope is power-hungry sensor [32]. Typical waveforms of the target classes are presented in Fig. 2.

C. Data Collection

In data collection, 70 subjects (53 male and 17 female, undergraduate or graduate students at the age of 20’s) were recruited with a 2,000 yen equivalent worth of remuneration. The subjects carried Huawei Ascend P7 smartphone terminals running Android 4.4 in 2 to 5 positions simultaneously, and were asked to walk about 5 min per position in the campus of our university including straight ways and corners at three walking speeds, i.e., slow, normal and fast. The speeds could be chosen by the subjects themselves; however, the order of the trial in the walking speed was kept constant such that fast, normal and slow. The subjects may get tired as the experiments proceeds. So, we consider that it is preferable to start walking with fast speed. We collected raw acceleration signals from Android API at the speed of `SENSOR_DELAY_FASTEST`. The sampled data from Android sensor system are added to an internal queue of our data collection system and polled at 50 Hz. Note that the data on a phones being carried in trousers pockets covered four orientations: downward and upward and with the display surface facing towards and away from the body.

IV. ON-BODY LOCALIZATION SYSTEM

A. Overview

The localization is carried out to recognize a class of a position from the 17 candidate positions based on the similarity

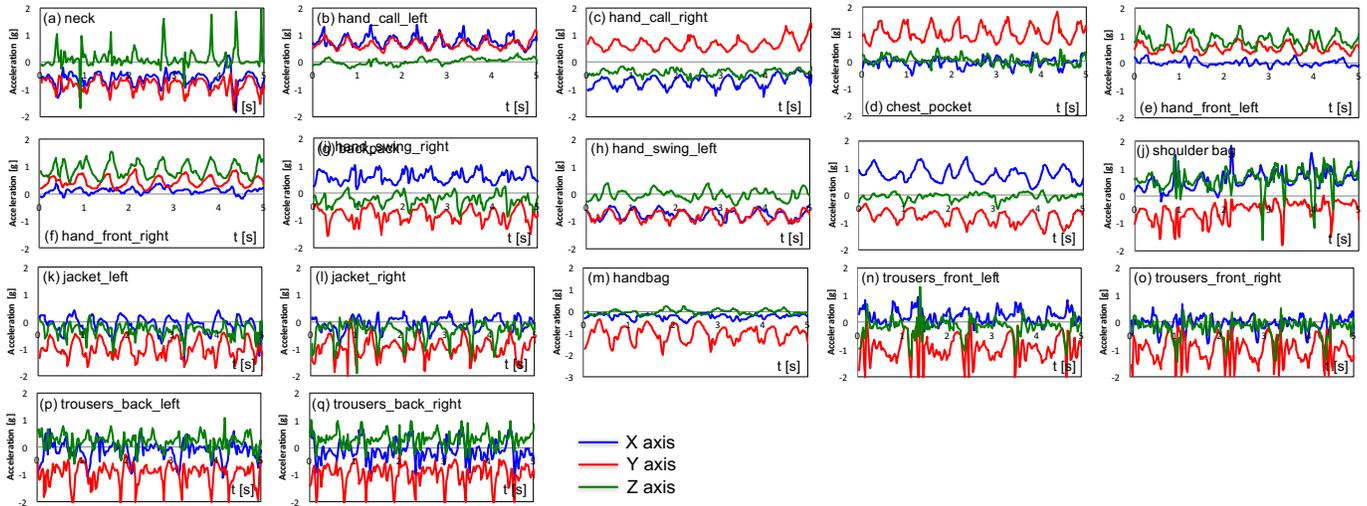


Fig. 2. Signal variations in acceleration during walking.

of patterns of acceleration signals. The recognition process is carried out window-by-window, in which a window consists of a certain number of sampled acceleration signals. In line with the principles of Vahdatpour, et al. [29], Fujinami [7], and Mannini, et al. [21], primarily recognizes the storing position of a smartphone while a person is walking. In this article, we assume that a segment representing a person is walking is already identified.

B. Signal Series and Axis

The term “series” indicates the type of basic time series data, which includes *raw* acceleration signal, *linear* acceleration component, and *vertical* acceleration component. As described in Section III-C, the raw acceleration signal is what is just obtained from accelerometer. In this section, we present the other two series. Here, the notation $a_{s,a,i}$ represents the i -th sample in a window of a -axis in the s -signal series. The coordinates in the definition of the three series is illustrated in Fig. 3.

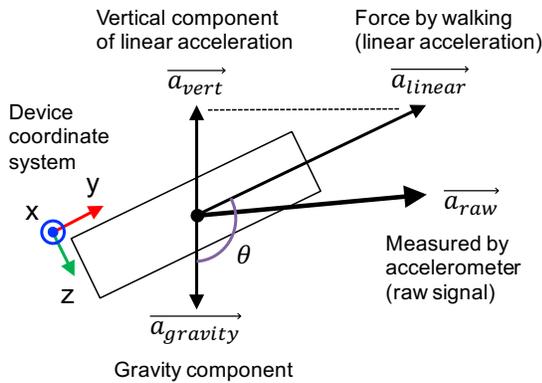


Fig. 3. The definition of three series.

1) *Linear Component*: Linear acceleration is obtained by removing gravity components from the measured signals. Sophisticated linear acceleration signal estimation methods have

been proposed by combining gyroscope and magnetometer [13]; however, we utilize only accelerometer for the same reason as the choice of an accelerometer as a modality of storing position recognition. We adopted the method proposed by Cho et al. [3], in which the gravity components are approximated by the mean of raw acceleration signals (1) in a window, and the linear components are obtained by subtracting the gravity component from the raw acceleration signals (2). Here, \bar{a}_{raw} is a vector of the mean raw acceleration signals of x , y and z axes in a window. Also, \vec{a}_{linear} and \vec{a}_{raw} indicate a vector of a sample of linear acceleration signal and raw acceleration signal in a window, respectively.

$$\vec{a}_{gravity} = \bar{a}_{raw} \quad (1)$$

$$\vec{a}_{linear} = \vec{a}_{raw} - \vec{a}_{gravity} \quad (2)$$

2) *Vertical Component*: The vertical component is obtained by decomposing the linear component based on the component of gravity in each axis (3) [13].

$$\begin{aligned} \vec{a}_{vertical} &= |\vec{a}_{vertical}| \frac{\vec{a}_{gravity}}{|\vec{a}_{gravity}|} \\ &= (|\vec{a}_{linear}| \cos\theta) \frac{\vec{a}_{gravity}}{|\vec{a}_{gravity}|} \end{aligned} \quad (3)$$

Here, $\cos\theta$ is obtained based on the definition of *inner product* (\cdot) as represented by (4).

$$\cos\theta = \frac{\vec{a}_{linear} \cdot \vec{a}_{gravity}}{|\vec{a}_{linear}| |\vec{a}_{gravity}|} \quad (4)$$

Then, (3) is represented with the gravity and the linear components by assigning (4) as represented by (5).

$$\begin{aligned} \vec{a}_{vertical} &= (|\vec{a}_{linear}| \frac{\vec{a}_{linear} \cdot \vec{a}_{gravity}}{|\vec{a}_{linear}| |\vec{a}_{gravity}|}) \frac{\vec{a}_{gravity}}{|\vec{a}_{gravity}|} \\ &= (\frac{\vec{a}_{linear} \cdot \vec{a}_{gravity}}{\vec{a}_{gravity} \cdot \vec{a}_{gravity}}) \vec{a}_{gravity} \end{aligned} \quad (5)$$

In addition to the three axes, i.e., x , y and z , we introduce the magnitude of the three-axes signals (m) as the fourth dimension for series s as shown in (6).

$$a_{s,m} = |\vec{a}_s| \quad (6)$$

C. Recognition Features

Recognition features play very important role on determining the performance of a recognition system. In this section, we describe the definition of the candidates of features. The localization is carried out to recognize a class of a position from the 17 candidate positions based on the similarity of patterns of acceleration signals. The recognition process is carried out window-by-window, in which a window consists of a certain number of sampled acceleration signals. A feature vector is obtained per window, in which features are calculated against the three series of acceleration signals.

We take an approach of listing up candidates of features from literature [1], [7], [14], [30] and observation of waveforms (Fig. 2), and selecting relevant and non-redundant features based on a machine learning technique. We systematically calculate the candidates of features from a window of four-dimensional vector of raw acceleration signals by the combination of feature types and the axes. Totally, 326 features are obtained (72 types \times 4 axes for individual axes, 5 (one for time domain and four for frequency domain) types \times 6 ($=_4C_2$) pairs for correlation-based features and two types \times 4 ($=_4C_3$) triples for features obtained from combination of three axes). Tables I, II, and III show the features calculated from the four axes individually, the features regarding the correlation of two axes, i.e., *correlation coefficient*, and the features representing the relationship among three axes, respectively. The feature selection is described in Section V-B.

Regarding the subscript L , M and H , the frequency spectrum is equally divided into three “frequency ranges”, which correspond to 0.001-5.000 Hz, 5.001-10.000 Hz and 10.001-25.000 Hz, respectively. In addition, the subscript *all* indicates the entire frequency range of 0.20-25.00 Hz. Note that a feature maxSdev_F is obtained in a way similar to “sliding window average”; a subwindow with 2.9 Hz range is created in an entire frequency spectrum to calculate standard deviation (*sdev*); the subwindow is slid by 0.1 Hz throughout the frequency spectrum; and the maximum *sdev* is found. maxSdev_F is the central frequency of a particular subwindow that gives maxSdev_F . The size and sliding-width (0.1 Hz) of subwindow were heuristically determined. A feature calculated as the sum of squared values of frequency components is sumPower_F (a.k.a “FFT energy” in [9]) [2]. The FFT entropy (entropy_F) is then calculated as the normalized information entropy of FFT component values of acceleration signals, which represents the distribution of frequency components in the frequency domain [2].

V. EXPERIMENT

In this section, we describe experiments from various aspects.

A. Condition

The window size is set to 256 samples, i.e., 5.12 seconds, with the sliding of every 128 samples (overlapping 50 %).

Throughout the experiment, we utilized a machine learning toolkit Weka 3.7.13 [20] running on Apple Mac Pro (3.5 GHz 6-Core Intel Xeon E5, 32 GB RAM, OS X El Capitan). Table IV summarizes average number of recognition instances, i.e., feature vectors, and standard deviation per person.

TABLE IV. AVERAGE NUMBER AND STANDARD DEVIATION (S.D.) OF RECOGNITION INSTANCES PER PERSON

Carrying state	Average (S.D.)	Carrying state	Average (S.D.)
bag_backpack	369.7 (32.6)	jacket_left	366.6 (27.7)
bag_handbag	362.5 (24.8)	jacket_right	368.5 (25.8)
bag_shoulderbag	373.6 (31.4)	neck	372.5 (30.9)
chest_pocket	367.5 (29.9)	trousers_back_left	366.7 (26.0)
hand_call_left	365.6 (29.3)	trousers_back_right	367.6 (26.2)
hand_call_right	370.2 (33.5)	trousers_front_left	365.9 (25.2)
hand_front_left	367.0 (28.6)	trousers_front_right	368.8 (26.6)
hand_front_right	366.3 (28.7)		
hand_swing_left	369.8 (30.7)	total	6253.9 (427.6)
hand_swing_right	364.9 (27.6)		

B. Feature Selection

1) *Methodology*: Feature selection consists of three phases: feature subset evaluation, feature subset search, and series selection. As feature subset evaluation, we utilized a correlation-based feature selection (CFS) [11], which is called *CfsSubsetEval* in Weka. CFS has a heuristic evaluation function *merit*, which can specify subset of features that are highly correlated with classes, i.e., more predictive of classes, but uncorrelated with each other, i.e., more concise. As described in Section IV-C, a large number of features were listed up, which may contain redundant features. So, we consider that the capability of CFS is suitable for this problem. Formula (7) defines the heuristic merit M_S of a feature subset S that contains k features, in which $\overline{r_{cf}}$ is the mean feature-class correlation and $\overline{r_{ff}}$ is the mean feature-feature inter-correlation. For more detail, please refer to [11]. M_S acts as a ranking on feature subsets in the search space of all possible feature subsets. Note that CFS is a classifier-independent method of feature selection.

$$M_S = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \quad (7)$$

To find the subset of features based on the CFS evaluation, we initially attempted to utilize the forward greedy stepwise search against entire feature set (*GreedyStepwise* in Weka). The method searches the best feature subsets, which begins with no features and greedily adds features one by one. However, the computation ended up with out of memory error. So, we needed to take another approach, which finds a subset with much smaller number of features than entire dataset, i.e., 326 features, with lightweight computation at first and applies the greedy stepwise search on the subset. As a lightweight computation of searching the space of feature subsets, we utilized, *BestFirst* in Weka, a greedy hill-climbing method augmented with a backtracking facility. Setting the number of consecutive non-improving samples allowed controlling the level of backtracking done. In this experiment, the number was set to five.

The dataset of selected feature subset for each “series” is evaluated by 10 fold cross-validation (10 fold CV) to specify

TABLE I. CLASSIFICATION FEATURES (x , y AND z AXES AND THE MAGNITUDE (m) OF THE THREE AXES)

Name	Description or Formula
$mean_T$	Average of time-series data
var_T	Variance of time-series data
max_T	Maximum value of time-series data
min_T	Minimum value of time-series data
$range_T$	Difference between max and min, i.e., $max_T - min_T$
$skew_T$	Skewness of time-series data, i.e., $\frac{\frac{1}{N} \sum_{i=1}^N (a_i - mean_T)^3}{var_T^{\frac{3}{2}}}$
$kurto_T$	Kurtosis of time-series data, i.e., $\frac{\frac{1}{N} \sum_{i=1}^N (a_i - mean_T)^4}{var_T^2}$
RMS_T	Root mean square of time-series data, i.e., $\sqrt{\frac{1}{N} \sum_{i=1}^N a_i^2}$
$absMean_T$	Absolute value of $mean_T$, i.e., $ mean_T $
$absMax_T$	Absolute value of max_T , i.e., $ max_T $
$absMin_T$	Absolute value of min_T , i.e., $ min_T $
$meanAbsDT$	Averaged absolute value of successive value's difference, i.e., $\frac{1}{N-1} \sum_{i=1}^{N-1} a_{i+1} - a_i $
$meanXing_T$	The number of crossing the mean value
$1^{st}Q_T$	1^{st} quartile (1/4 smallest value) of time-series data
$3^{rd}Q_T$	3^{rd} quartile (3/4 smallest value) of time-series data
IQR_T	Inter-quartile range of time-series data, i.e., $3^{rd}Q_T - 1^{st}Q_T$
$energy_{F,\{all L M H\}}$	Sum of energy spectrum, i.e., $\sum_{i=1}^{N/2} f_i^2$
$entropy_{F,\{all L M H\}}$	Frequency entropy, i.e., $-\sum_{i=1}^{N/2} p_i \times \log_2 p_i$, where $p_i = f_i^2 / \sum_{i=1}^{N/2} f_i^2$
$max_{F,\{all L M H\}}$	Maximum value in an entire frequency spectrum
$maxF_{F,\{all L M H\}}$	Frequency that gives max_F
$meanF_{F,\{all L M H\}}$	Mean frequency, i.e., $\frac{\Delta f}{N/2} \sum_{i=1}^{N/2} (f_i \times i)$
$1^{st}Q_{F,\{all L M H\}}$	1^{st} quartile (1/4 smallest) frequency spectrum
$3^{rd}Q_{F,\{all L M H\}}$	3^{rd} quartile (3/4 smallest) frequency spectrum
$IQR_{F,\{all L M H\}}$	Inter-quartile range of frequency spectrum, i.e., $3^{rd}Q_F - 1^{st}Q_F$
$1^{st}Q_{F_{F,\{all L M H\}}}$	Frequency that gives $1^{st}Q_F$
$3^{rd}Q_{F_{F,\{all L M H\}}}$	Frequency that gives $3^{rd}Q_F$
$var_{F,\{all L M H\}}$	Variance in the low-frequency range
$maxSdev_{F,\{all L M H\}}$	Maximum standard deviation in subwindows in frequency spectrum
$maxSdevF_{F,\{all L M H\}}$	Central frequency of subwindow that gives $maxSdev_F$
$cepDens_{F,\{all L M H\}}$	Cepstrum density, i.e., $\frac{1}{N/2} \sum_{i=1}^{N/2} Cep_i ^2$, where Cep_i is the i -th element of cepstrum coefficient

TABLE II. CLASSIFICATION FEATURES BASED ON CORRELATION COEFFICIENTS BETWEEN TWO AXES

Name	Description
$corr_T$	Pearson's correlation coefficient of signals from two axes in time-series data
$corr_{F,\{all L M H\}}$	Correlation coefficient in an entire frequency spectrum

TABLE III. CLASSIFICATION FEATURES OBTAINED FROM THREE AXES ($i, j, k \in \{x, y, z, m\}, i \neq j \neq k$)

Name	Description
$max3axes_T$	Max of the max of 3 out of 4 axes, i.e., $max(max_{i,T}, max_{j,T}, max_{k,T})$
$min3axes_T$	Min of the min of 3 out of 4 axes, i.e., $min(min_{i,T}, min_{j,T}, min_{k,T})$

the best feature subset for later analysis. The RandomForest classifier with 10 trees is utilized as a base classifier for the cross-validation. The classification result is evaluated by F-measure. F-measure is a harmonic mean between recall and precision. F-measure for class i is defined by (8), which is averaged over 17 classes. The recall and precision for class i are represented by (9) and (10), where $N_{correct_i}$, N_{tested_i} , and N_{judged_i} represent the number of cases correctly classified into $class_i$, the number of test cases in $class_i$, and the number of cases classified into $class_i$, respectively, while i corresponds to either one of 17 classes.

$$F\text{-measure}_i = \frac{2}{1/recall_i + 1/precision_i} \quad (8)$$

$$recall_i = \frac{N_{correct_i}}{N_{tested_i}} \quad (9)$$

$$precision_i = \frac{N_{correct_i}}{N_{judged_i}} \quad (10)$$

2) *Result*: The BestFirst search filtered out about 70 features for each series as initial “meaningful” features. We then applied GreedyStepwise search against these features to understand the best combination in particular number of features. Fig. 4 shows the relationship between the number of features and the merit score M_s . Here, the number of features increases in the order of adding to the feature subset. As shown in the figure, the increase of M_s becomes saturated at a particular number of features. This indicates that the redundancy of features increased and/or the predictiveness of an added feature decreases after a particular number of features. The merit scores of “raw” series are larger than the other two series in almost all cases of the number of features. This suggests that “raw” series contains more predictive and less redundant than the features from the other two series and may performs best.

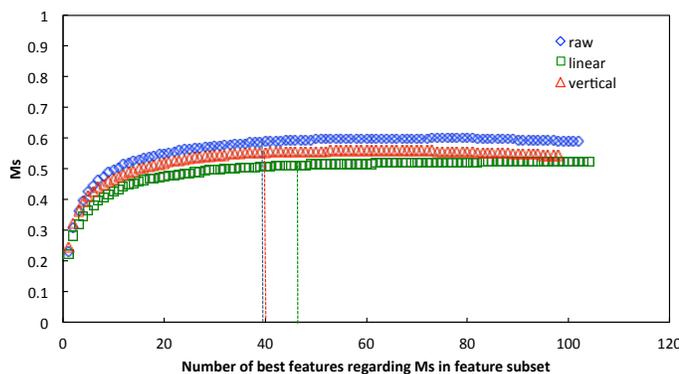


Fig. 4. Relationship between the size of feature subset and merit score of the subset (partially).

We utilized 40, 45 and 40 features for “raw”, “linear” and “vertical” data series near the saturation points, respectively, in series selection. Table V summarizes average F-measures of the three series. As shown in the table, “raw” series performed the best in the three series using selected feature subsets. Table VI summarizes the selected features for “raw” data series.

TABLE V. F-MEASURE FOR EACH DATA SERIES

series	raw	linear	vertical
F-measure	0.980	0.940	0.948

TABLE VI. SELECTED FEATURES FOR “RAW” DATA SERIES. “#” REPRESENTS THE ORDER OF ADDING TO THE FEATURE SUBSET, WHILE M_s INDICATES THE MERIT SCORE OF THE SUBSET

#	Name	M_s	#	Name	M_s
1	$mean_{T,y}$	0.231	21	$var_{F,all,z}$	0.554
2	$energy_{F,M,y}$	0.309	22	$3^{rd}Q_{T,z}$	0.557
3	$1^{st}Q_{T,x}$	0.363	23	$min3axes_{T,xy,m}$	0.560
4	$mean_{T,z}$	0.398	24	$3^{rd}Q_{F,M,x}$	0.562
5	$3^{rd}Q_{F,M,z}$	0.426	25	$corr_{F,L,ym}$	0.564
6	$3^{rd}Q_{T,x}$	0.445	26	$corr_{T,xy}$	0.566
7	$entropy_{F,H,z}$	0.462	27	$meanXing_{T,m}$	0.568
8	$corr_{T,ym}$	0.476	28	$max_{T,x}$	0.570
9	$meanAbsD_{T,x}$	0.489	29	$kurt_{OT,z}$	0.572
10	$maxSdev_{F,L,z}$	0.498	30	$skew_{T,y}$	0.574
11	$meanXing_{T,x}$	0.506	31	$corr_{T,zm}$	0.576
12	$meanF_{F,L,y}$	0.514	32	$mean_{T,x}$	0.577
13	$1^{st}Q_{T,z}$	0.519	33	$3^{rd}Q_{F,H,y}$	0.579
14	$meanXing_{T,y}$	0.525	34	$corr_{F,M,ym}$	0.581
15	$cepDens_{F,M,x}$	0.531	35	$cepDens_{F,M,m}$	0.582
16	$maxF_{F,all,x}$	0.535	36	$meanXing_{T,z}$	0.584
17	$skew_{T,z}$	0.539	37	$maxF_{F,all,z}$	0.585
18	$mean_{T,m}$	0.543	38	$cepDens_{F,M,z}$	0.586
19	$corr_{T,xm}$	0.547	39	$entropy_{F,H,x}$	0.587
20	$corr_{F,L,zm}$	0.550	40	$1^{st}Q_{F,H,z}$	0.588

C. Classifier Selection

To find the best classifier, we compare popular classifiers by taking into account not only correctness of classification, but also the computational load for running on a smartphone.

1) *Methodology*: We carried out 10 fold CVs against Naïve Bayes, Multi-Layer Perceptron (MLP), J48 Tree and RandomForest classifiers using 40 features from raw dataset. We also measured the elapsed time to complete one fold of evaluation (test) that contains approximately 44,000 instances. Note that different numbers of trees in RandomForest were tested, i.e., 10, 50 and 100. The Support Vector Machines (SVM) has not been tested because it is parameter sensitive.

2) *Result and Analysis*: Fig. 5 shows the average F-measure of the classifiers and elapsed time for testing dataset per one fold. As shown in the figure, three types of RandomForest performed better than the others. Paired t-tests showed significant difference between RandomForest with 10 trees and Naïve Bayes, MLP, and J48 with $p < .05$ ($t(9)=224.80$, $t(9)=94.83$, and $t(9)=45.99$, respectively). Also, the performance gets better as the number of threes in creased from 0.980 to 0.987. To determine the number of trees, it is important to consider the trade-off between the classification performance and processing workload. As shown in Fig. 5, the number of trees in RandomForest influences the processing speed because of the nature of the algorithm. Although paired t-tests showed that RandomForest with 10 trees was significantly lower in F-measure than RandomForest with 50 and 100 trees with $p < .05$ ($t(9)=-44.03$ and $t(9)=-47.19$, respectively), we took the number of trees 10 for the following experiments by taking the processing speed in processing on the smartphone. Hereinafter, RandomForest with 10 trees is utilized as a classifier in this article.

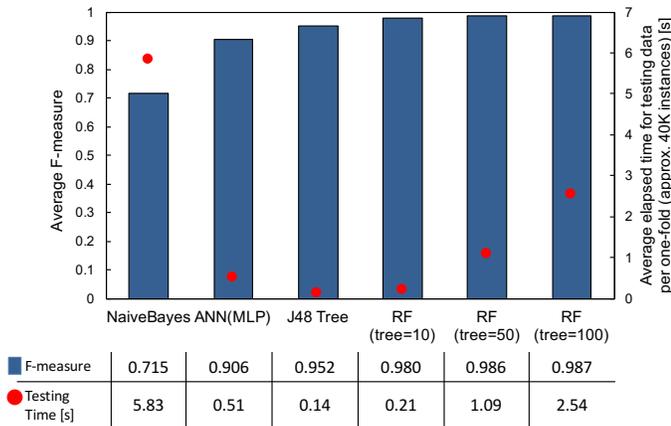


Fig. 5. The difference of F-measure in various classifiers. Note on the acronyms of classifiers: ANN = Artificial Neural Networks, MLP = Multi-Layer Perceptron, RF=RandomForest.

D. Feature Subset Evaluation vs. Individual Feature Collection

In Section V-B, 40 features were selected using CFS, which allows us to find subset of features that are more predictive of classes yet less correlated with each other. In this section, we evaluate the subset evaluation approach by comparing a collection of individual “good” features.

1) *Methodology*: The contribution of each feature is evaluated based on information gain (IG). IG is commonly used in feature selection, where the gain of information provided by a particular feature is calculated by subtracting a conditional entropy with that feature from the entropy under random guess [31]. So, the more informative feature has the higher IG.

After specifying the same number of features as those obtained by CFS method, i.e., 40 features, 10 fold CVs are performed against these two feature subsets, and F-measures are compared.

2) *Result and Analysis*: Table VII summarizes top-40 informative features based on IG. The features derived from x -axis show their effectiveness by appearing 7 in top-10 features. The table also shows the order of adding to the subset of 40 features obtained by CFS, which shows that individually informative feature, i.e., high IG, is not always selected in early stage (or not at all) of adding to CFS-based feature subset, i.e., low CFS value. This is natural because CFS is designed to take into account the redundancy among features and find the best combination of features, while IG is used to represent the informativeness of individual features.

Regarding the classification performance, the F-measures obtained from classifiers trained by IG-based features and CFS-based features are 0.967 and 0.980, respectively, and CFS-based feature subset is significantly contributive in classification compared to IG-based one ($t(9)=83.06, p<.05$). Therefore, suppose that the same number of features is utilized, we consider that the approach of feature subset evaluation was effective in building better classifier than collecting individual features with good evaluation results.

TABLE VII. TOP-40 INFORMATIVE FEATURES BASED ON INFORMATION GAIN FEATURE EVALUATION. THE COLUMN “CFS” INDICATES THE ORDER OF ADDING TO THE FEATURE SUBSET AS SHOWN IN TABLE VI, IN WHICH “-” REPRESENTS THAT THE FEATURE IS NOT INCLUDED IN THE SUBSET OF 40 FEATURES OBTAINED BY CFS

Rank	Name	IG [bit]	CFS	Rank	Name	IG [bit]	CFS
1	$mean_{T,y}$	1.21	1	20	$3^{rd}Q_{F,H,y}$	0.97	33
2	$mean_{T,x}$	1.13	32	20	$mean_{F,F,M,y}$	0.97	-
3	$1^{st}Q_{T,x}$	1.10	3	20	$energy_{F,M,z}$	0.97	-
4	$3^{rd}Q_{T,x}$	1.09	6	24	$3^{rd}Q_{F,all,y}$	0.96	-
5	$RMS_{T,x}$	1.08	-	24	$3^{rd}Q_{F,M,z}$	0.96	5
6	$cepDens_{F,all,x}$	1.07	-	24	$var_{F,M,z}$	0.96	-
7	$cepDens_{F,H,x}$	1.05	-	24	$mean_{F,F,M,z}$	0.96	-
8	$corr_{T,ym}$	1.03	8	24	$meanAbsD_{T,y}$	0.96	-
9	$energy_{F,M,y}$	1.03	2	29	$var_{F,L,y}$	0.95	-
10	$cepDens_{F,L,x}$	1.01	-	30	$meanAbsD_{T,x}$	0.94	9
11	$cepDens_{F,M,x}$	1.00	15	30	$maxSdev_{F,L,x}$	0.94	-
11	$max_{F,all,x}$	1.00	-	30	$maxSdev_{F,all,x}$	0.94	-
11	$max_{F,H,x}$	1.00	-	30	$maxSdev_{F,H,x}$	0.94	-
11	$max_{F,L,x}$	1.00	-	34	$IQR_{F,M,y}$	0.93	-
15	$3^{rd}Q_{F,L,y}$	0.99	-	34	$1^{st}Q_{T,z}$	0.93	13
15	$mean_{T,z}$	0.99	4	36	$mean_{F,F,all,y}$	0.94	-
17	$3^{rd}Q_{T,y}$	0.98	-	36	$3^{rd}Q_{T,z}$	0.94	22
17	$meanAbsD_{T,z}$	0.98	-	36	$1^{st}Q_{F,all,z}$	0.94	-
19	$var_{F,M,y}$	0.97	-	36	$entropy_{F,H,z}$	0.94	7
20	$mean_{F,F,L,y}$	0.97	12	40	$1^{st}Q_{T,y}$	0.93	-

E. Recognition against Unknown Person

As described in Section II, LOSO-CV is regarded as a fairer and more practical test method under a condition in which individual difference exists. In this section, we apply LOSO-CV to 70 subjects, which we consider the largest case in on-body smartphone localization.

1) *Methodology*: The dataset from one subject is treated as a test set, while the dataset from remaining 69 subjects are utilized for training a classifier. The train-and-test process is iterated 70 times.

2) *Result and Analysis*: Table VIII shows the detail of the classification result in the form of confusion matrix. The average F-measure in the classification of 17 classes against 70 subjects is 0.823. Although the value decreased by 0.157 from the one by 10 fold-CV, we consider that the performance is rather good given that there are 17 classes. Especially, by taking into account that no data from person for testing are included in the training data, it is surprising that left and right sides in “hand_call” and “hand_swing” were separated with very high F-measure (>0.93), in which clear differences in x -axis are observed as shown in Fig. 2. “Neck” also has high F-measure (0.932). A smartphone hanging from the neck is hit by the user’s body as he/she walks forward, which causes strong impact on z -axis (Fig. 2(a)). However, as shown in Table VIII, the discriminations of left and right sides of “hand_front”, “trousers_back”, and “trousers_front” are often confused with each other. As shown in Fig. 2, less differences are observed in the left and right sides of these classes than the successful cases. Also, the confusion within “bags” is slightly observed.

So, we merged the left and the right sides into one class, e.g., “hand_call_left” and “hand_call_right” \rightarrow “hand_call”, against “hand_call”, “hand_front” and “hand_swing”. Also, three types of bags are merged into one “bag” class. The result of the merging is summarized in Table IX, in which the mean F-measure is 0.913 (increased by 0.090 from original 17 classes). Furthermore, three subclasses of hand and two subclasses of trousers pockets, i.e., front and back, are merged

into single class “hand” and “trousers pocket”, respectively, resulting in six class classification, which is shown in Table X. As shown in these tables, merging of multiple classes into a single class increases the performance metrics. Application designers should consider the required resolution, i.e., the level of detail of position recognition, for their target applications.

F. Effect of Various Walking Speed in Classifier Training

As described in Section III-C, we collected data with three walking speeds based on the decision of the subjects. The above experiments were carried out with dataset that contains all walking speeds. Training a classifier with single, i.e., “normal”, speed is easy for the participants in data collection; however, it may sacrifice the robustness against different speed. Data collection process can be simplified if no difference exists in the robustness between classifiers modeled with heterogeneous speed and single speed. In this section, we explore the effect of walking speed in classifier training.

1) *Methodology*: The experiment follows LOSO-CV principle with a slight difference in walking speed between training and test datasets. More specifically, two classifiers for 17 class classification are trained using 1) dataset that contains all speeds and 2) dataset with only “normal” speed, in which training a classifier with “normal” speed is a traditional approach. Here, a dataset obtained from a test subject is excluded from the training dataset. Meanwhile, the dataset for test is either “slow”, “normal”, and “fast” speed. For example, a combination of “normal” speed for training with “fast” speed for testing represents a case where a person is walking faster than what the classifier knows.

In training classifiers with three walking speeds, we reduced the size of dataset to 1/3 so that it can become similar size to that of “normal” speed to avoid the bias of the number of training instances. Actually, three sets of 1/3 sampled dataset are applied, and F-measures are averaged. Regarding the classification features for training with “normal” speed, we selected dedicated ones in the same way as with all speeds (Section V-B) because we consider that suitable set of features can be different from each other due to the variation of walking speed in “all speed” case.

2) *Result and Analysis*: Table XI summarizes average F-measures in different combinations of walking speed in training and test datasets. Paired t-tests regarding the heterogeneity in training datasets showed that using three walking speeds performed better classification than using single, i.e., “normal”, speed ($p < .05$) in all cases of walking speeds in test datasets ($t(69) = -2.34$, $t(69) = 2.64$, and $t(69) = 5.30$ for “slow”, “normal”, and “fast”, respectively). The result shows that, in building a classifier, heterogeneity of walking speed is important for robust classifier.

TABLE XI. AVERAGE F-MEASURES IN DIFFERENT COMBINATIONS OF TRAINING AND TEST DATASETS

Trained\Tested	Slow	Normal	Fast
All	0.799	0.814	0.787
Normal	0.786	0.801	0.758

VI. CONCLUSION

In this article, we proposed a machine learning-based classifier and classification features to identify 17 states of a smartphone while the user is walking. A large-scale data collection from 70 persons were carried out with three different walking speeds to evaluate the effect of heterogeneity of walking speed in training a classifier. The following results were obtained:

- In feature calculation, we introduced three series of acceleration signals, *raw*, *linear*, and *vertical* components, in which the *raw* acceleration series showed the highest classification performance in the three series.
- 40 features in the *raw* series were selected from 326 candidates features based on correlation-based feature subset evaluation. The comparison with a subset by collecting individually informative features based on information gain showed that the subset evaluation method was superior to the collection-based method with the same number of features.
- Person-independent evaluation (LOSO-CV) showed that an average F-measure of 17 class classification was 0.823, while 9 class classification by merging left and right sides into one class showed an average F-measure of 0.913.
- Comparison of the heterogeneity of walking speeds in training dataset showed that the classifier built from various walking speed allowed us to realize more robust classifier than using a classifier with a single walking speed (normal speed).

We consider that the F-measure of 0.824 for 17 class classification has still room for improvement by using suitable classifier to address “classifier compatibility” issue as suggested in [7]. In addition, the classification in the experiment was carried out against a window, which means that decisions of successive windows may differ due to mis-classification. For practical recognition, we will investigate temporal smoothing techniques to smoothen such “discontinuity” of recognition results. We have already developed a mechanism to identify a segment of walking, to which these future investigation will be integrated.

ACKNOWLEDGMENT

This work is partially supported by the grant in aid from The Telecommunication Advancement Foundation.

REFERENCES

- [1] K. Alanezi and S. Mishra, “Design, implementation and evaluation of a smartphone position discovery service for accurate context sensing,” *Computers & Electrical Engineering*, vol. 44, pp. 307–323, 2015. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S004579061500021X>
- [2] L. Bao and S. S. Intille, “Activity recognition from user-annotated acceleration data,” in *Proceedings of the 2nd International Conference on Pervasive Computing (Pervasive 2004)*, vol. LNCS 3001. Springer-Verlag, 2004, pp. 1–17.
- [3] S.-J. Cho, E. Choi, W.-C. Bang, J. Yang, J. Sohn, D. Y. Kim, Y.-B. Lee, and S. Kim, “Two-stage Recognition of Raw Acceleration Signals for 3-D Gesture-Understanding Cell Phones,” in *Proceedings of the Tenth International Workshop on Frontiers in Handwriting Recognition*, 2006.

TABLE VIII. CONFUSION MATRIX OF LOSO-CV AGAINST 70 SUBJECTS, IN WHICH THE ROW AND THE COLUMN INDICATE THE LABELED AND THE PREDICTED CLASSES, RESPECTIVELY. THE NUMBER OF INSTANCES IN EACH CLASS IS NORMALIZED SO THAT THE SUM OF EACH ROW BECOMES 100

Labeled\Predicted	a.	b.	c.	d.	e.	f.	g.	h.	i.	j.	k.	l.	m.	n.	o.	p.	q.
a. bag_backpack	78	0	6	4	0	2	0	0	0	0	4	3	0	0	1	0	0
b. bag_handbag	0	94	0	1	1	1	0	1	0	0	1	1	0	0	0	0	0
c. bag_shoulderbag	5	0	87	1	0	0	1	0	0	0	2	2	2	0	0	0	0
d. chest_pocket	2	2	0	89	1	1	0	0	0	0	1	2	0	0	0	0	0
e. hand_call_left	0	0	0	1	97	0	0	0	0	0	0	0	0	0	0	0	0
f. hand_call_right	0	0	0	1	0	96	0	2	0	0	0	0	0	0	0	0	0
g. hand_front_left	0	0	1	2	0	0	86	10	1	0	0	0	0	0	0	0	0
h. hand_front_right	0	0	0	1	0	3	13	77	0	6	0	0	0	0	0	0	0
i. hand_swing_left	0	0	0	0	2	0	0	0	96	0	0	0	0	0	0	0	0
j. hand_swing_right	0	0	0	0	0	1	0	3	1	93	0	0	0	0	0	0	0
k. jacket_left	3	1	1	3	1	0	0	0	1	0	72	11	2	0	0	2	2
l. jacket_right	3	1	1	4	0	0	0	0	0	1	8	71	3	1	1	2	2
m. neck	0	0	1	0	0	0	0	0	0	0	1	1	96	0	0	0	0
n. trousers_back_left	1	0	0	2	0	0	0	0	0	0	1	2	1	66	23	2	2
o. trousers_back_right	1	0	1	1	0	0	0	0	0	0	1	2	0	25	67	1	1
p. trousers_front_left	0	0	0	1	0	0	0	0	0	0	2	2	1	2	0	75	17
q. trousers_front_right	0	0	0	1	0	0	0	0	0	0	1	4	0	2	2	27	62
Recall (average: 0.825)	0.784	0.937	0.869	0.887	0.975	0.963	0.856	0.767	0.958	0.931	0.718	0.715	0.959	0.660	0.672	0.749	0.622
Precision (average: 0.824)	0.815	0.954	0.875	0.801	0.937	0.909	0.842	0.832	0.959	0.927	0.759	0.697	0.907	0.680	0.710	0.682	0.715
F-measure (average: 0.823)	0.799	0.945	0.872	0.842	0.956	0.935	0.849	0.798	0.959	0.929	0.738	0.706	0.932	0.670	0.690	0.714	0.666

TABLE IX. PERFORMANCE OF LOSO-CV FOR MERGED 9 CLASSES AGAINST 70 SUBJECTS

Labeled\Predicted	bag	chest_pocket	hand_call	hand_front	hand_swing	jacket_pocket	neck	trousers_back	trousers_front	average
Recall	0.900	0.887	0.969	0.926	0.949	0.815	0.959	0.903	0.908	0.913
Precision	0.862	0.897	0.920	0.956	0.949	0.821	0.953	0.940	0.921	0.913
F-measure	0.881	0.892	0.944	0.941	0.949	0.818	0.956	0.921	0.914	0.913

TABLE X. PERFORMANCE OF LOSO-CV FOR MERGED 6 CLASSES AGAINST 70 SUBJECTS

Labeled\Predicted	bag	chest_pocket	hand	jacket_pocket	neck	trousers_pocket	average
Recall	0.900	0.887	0.982	0.815	0.959	0.935	0.913
Precision	0.884	0.922	0.928	0.857	0.959	0.925	0.912
F-measure	0.892	0.904	0.954	0.835	0.959	0.930	0.912

[4] Y. Cui, J. Chipchase, and F. Ichikawa, "A Cross Culture Study on Phone Carrying and Physical Personalization," in *Proceedings of HCI International 2007*, 2007, pp. 483-492.

[5] I. Diaconita, A. Reinhardt, D. Christin, and C. Rensing, "Inferring Smartphone Positions Based on Collecting the Environment's Response to Vibration Motor Actuation," in *Proceedings of the 11th ACM Symposium on QoS and Security for Wireless and Mobile Networks*, ser. Q2SWinet '15. New York, NY, USA: ACM, 2015, pp. 99-106. [Online]. Available: <http://doi.acm.org/10.1145/2815317.2815342>

[6] I. Diaconita, A. Reinhardt, F. Englert, D. Christin, and R. Steinmetz, "Do you hear what {I} hear? Using acoustic probing to detect smartphone locations," in *2014 {IEEE} International Conference on Pervasive Computing and Communication Workshops, PerCom 2014 Workshops, Budapest, Hungary, March 24-28, 2014*, 2014, pp. 1-9. [Online]. Available: <http://dx.doi.org/10.1109/PerComW.2014.6815157>

[7] K. Fujinami, "On-Body Smartphone Localization with an Accelerometer," *Information*, vol. 7, no. 2, p. 21, 2016. [Online]. Available: <http://www.mdpi.com/2078-2489/7/2/21>

[8] K. Fujinami, "Smartphone-based Environmental Sensing Using Device Location as Metadata," *International Journal on Smart Sensing and Intelligent Systems*, vol. 9, no. 4, 2016. [Online]. Available: <http://s2is.org/Issues/v9/n4/papers/abstract31.pdf>

[9] K. Fujinami and S. Kouchi, "Recognizing a Mobile Phone's Storing Position as a Context of a Device and a User," in *Mobile and Ubiquitous Systems: Computing, Networking, and Services*, ser. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, K. Zheng, M. Li, and H. Jiang, Eds. Springer Berlin Heidelberg, 2013, vol. 120, pp. 76-88. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-40238-8_{_}7

[10] J. Goldman, K. Shilton, J. Burke, D. Estrin, M. Hansen, N. Ramanathan, S. Reddy, V. Samanta, M. Srivastava, and R. West, "Participatory Sensing: A citizen-powered approach to illuminating the patterns that shape our world," *Foresight and Governance Project, White Paper*, 2009.

[11] M. A. Hall, "Correlation-based Feature Selection for Machine Learning," Ph.D. dissertation, The University of Waikato, 1999.

[12] C. Harrison and S. E. Hudson, "Lightweight material detection for placement-aware mobile computing," in *UIST '08: Proceedings of the 21st annual ACM symposium on User interface software and technology*. New York, NY, USA: ACM, 2008, pp. 279-282.

[13] S. Hemminki, P. Nurmi, and S. Tarkoma, "Gravity and Linear Acceleration Estimation on Mobile Devices," in *Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, ser. MOBIQUITOUS '14. ICST, Brussels, Belgium, Belgium: ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2014, pp. 50-59. [Online]. Available: <http://dx.doi.org/10.4108/icst.mobiquitous.2014.258034>

[14] O. D. Incel, "Analysis of Movement, Orientation and Rotation-Based Sensing for Phone Placement Recognition," *Sensors*, vol. 15, no. 10, p. 25474, 2015. [Online]. Available: <http://www.mdpi.com/1424-8220/15/10/25474>

[15] T. Ishikawa and K. Fujinami, "Smartphone-Based Pedestrian's Avoidance Behavior Recognition towards Opportunistic Road Anomaly Detection," *ISPRS International Journal of Geo-Information*, vol. 5, no. 10, p. 182, oct 2016. [Online]. Available: <http://www.mdpi.com/2220-9964/5/10/182>

[16] T. Jimbo and K. Fujinami, "Detecting mischoice of public transportation route based on smartphone and GIS," in *UbiComp and ISWC 2015 - Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the Proceedings of the 2015*

- ACM International Symposium on Wearable Computers, 2015, pp. 165–168.
- [17] K. Kunze and P. Lukowicz, “Dealing with Sensor Displacement in Motion-based Onbody Activity Recognition Systems,” in *Proceedings of the 10th International Conference on Ubiquitous Computing*, ser. UbiComp '08. New York, NY, USA: ACM, 2008, pp. 20–29. [Online]. Available: <http://doi.acm.org/10.1145/1409635.1409639>
- [18] K. Kunze, P. Lukowicz, H. Junker, and G. Tröster, “Where am I: Recognizing On-body Positions of Wearable Sensors,” in *Proceedings of International Workshop on Location- and Context-Awareness (LoCA 2005)*, LNCS 3479, 2005, pp. 264–275.
- [19] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, “Activity recognition using cell phone accelerometers,” *SIGKDD Explor. Newsl.*, vol. 12, no. 2, pp. 74–82, 2011. [Online]. Available: <http://doi.acm.org/10.1145/1964897.1964918>
- [20] Machine Learning Group at University of Waikato, “Weka 3 - Data Mining with Open Source Machine Learning Software in Java,” <http://www.cs.waikato.ac.nz/ml/weka/>. [Online]. Available: <http://www.cs.waikato.ac.nz/ml/weka/>
- [21] A. Mannini, A. M. Sabatini, and S. S. Intille, “Accelerometry-based recognition of the placement sites of a wearable sensor,” *Pervasive and Mobile Computing*, vol. 21, pp. 62–74, 2015. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1574119215001108>
- [22] E. Miluzzo, M. Papandrea, N. D. Lane, H. Lu, and A. T. Campbell, “Pocket, Bag, Hand, etc.-Automatically Detecting Phone Context through Discovery,” in *Proc. of the First International Workshop on Sensing for App Phones (PhoneSense'10)*, 2010.
- [23] S. Pirttikangas, K. Fujinami, and T. Nakajima, “Feature Selection and Activity Recognition from Wearable Sensors,” in *2006 International Symposium on Ubiquitous Computing Systems (UCS2006)*, 2006, pp. 516–527.
- [24] A. Rai, K. K. Chintalapudi, V. N. Padmanabhan, and R. Sen, “Zee: Zero-effort Crowdsourcing for Indoor Localization,” in *Proceedings of the 18th Annual International Conference on Mobile Computing and Networking*, ser. Mobicom '12. New York, NY, USA: ACM, 2012, pp. 293–304. [Online]. Available: <http://doi.acm.org/10.1145/2348543.2348580>
- [25] Y. Shi, Y. Shi, and J. Liu, “A rotation based method for detecting on-body positions of mobile devices,” in *Proceedings of the 13th international conference on Ubiquitous computing*, ser. UbiComp '11. ACM, 2011, pp. 559–560.
- [26] M. Shoaib, S. Bosch, O. D. Incel, H. Scholten, and P. J. M. Havinga, “Fusion of Smartphone Motion Sensors for Physical Activity Recognition,” *Sensors*, vol. 14, no. 6, p. 10146, 2014. [Online]. Available: <http://www.mdpi.com/1424-8220/14/6/10146>
- [27] M. Stevens and E. D'Hondt, “Crowdsourcing of Pollution Data using Smartphones,” in *1st Ubiquitous Crowdsourcing Workshop at UbiComp*, 2010.
- [28] D. Sugimori, T. Iwamoto, and M. Matsumoto, “A Study about Identification of Pedestrian by Using 3-Axis Accelerometer,” in *Embedded and Real-Time Computing Systems and Applications (RTCSA), 2011 IEEE 17th International Conference on*, 2011, pp. 134–137.
- [29] A. Vahdatpour, N. Amini, and M. Sarrafzadeh, “On-body device localization for health and medical monitoring applications,” in *Proceedings of the 2011 IEEE International Conference on Pervasive Computing and Communications*, ser. PERCOM '11. IEEE Computer Society, 2011, pp. 37–44.
- [30] J. Wiese, T. S. Saponas, and A. J. B. Brush, “Phoneprioception: Enabling Mobile Phones to Infer Where They Are Kept,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '13. New York, NY, USA: ACM, 2013, pp. 2157–2166. [Online]. Available: <http://doi.acm.org/10.1145/2470654.2481296>
- [31] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, third edit ed. San Francisco, CA, USA: Morgan Kaufmann Publishers, 2011.
- [32] L. Zhang, P. H. Pathak, M. Wu, Y. Zhao, and P. Mohapatra, “AccelWord: Energy Efficient Hotword Detection Through Accelerometer,” in *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '15. New York, NY, USA: ACM, 2015, pp. 301–315. [Online]. Available: <http://doi.acm.org/10.1145/2742647.2742658>

Automated Player Selection for a Sports Team using Competitive Neural Networks

Rabah Al-Shboul
Al Al-Bayt University,
Mafrag, Jordan

Tahir Syed
National University
of Computer and &
Emerging Sciences, Pakistan

Jamshed Memon
Barrett Hodgson University,
Pakistan

Furqan Khan
French Institute for Research
in Computer Science
and Automation (INRIA), France

Abstract—The use of data analytics to constitute a winning team for the least cost has become the standard *modus operandi* in club leagues, beginning from Sabermetrics for the game of basketball. Our motivation is to implement this phenomenon in other sports as well, and for the purpose of this work we present a model for football, for which to the best of our knowledge, previous work does not exist.

The main objective is to pick the best possible squad from an available pool of players. This will help decide which team of 11 football players is best to play against a particular opponent, perform prediction of future matches and helps team management in preparing the team for the future. We argue in favour of a semi-supervised learning approach in order to quantify and predict player performance from team data with mutual influence among players, and report win accuracies of around 60%.

Keywords—Team selection; match outcome prediction; neural networks

I. INTRODUCTION

Predicting game outcomes in sports is both challenging and interesting for their potential value for betting houses, team managements, sports fans, sports media, etc. More specifically, management of football teams are interested in selecting the best team to play to maximize their chances of winning a game, hence optimizing their return on investments. Betting houses on the other hands, would like to have this ability so that they could adjust the betting odds to maximize their profits. Sports media can optimize their contract values for teams and players based on their likelihood of winning. We would not be a miss if we consider sports prediction market to be multi-billion and still not well tapped in.

Sports including baseball and basketball have used analytics based on past records and statistics to analyze and produce results for future use. It is a realistic expectation nowadays that automated systems should be used to predict results. Sabermetrics [8] (the baseball analytic model) is one such practical implementation. Basketball coaches and managers have used analytics regularly to maximize their results.

Creating an optimal lineup of players that is capable of winning over another lineup is a major challenge [2]. The difficulty comes from different player positions requiring different skills. This means that taken collectively, players performing optimally on as individuals does not necessarily translate to an optimal combination, because of new team dynamics.

This work aims to aid team managers and selectors in identifying the best possible team to play that has the best odds at winning. We focus on team games and on player statistics and analytics computed from past data. We also perform match-by-match analytics of matches played between specific opponents. Specifically, our objective is two-fold:

- 1) Selecting the best possible team combination for a specific match given the knowledge of a particular opponent, and
- 2) Predicting the likelihood of victory in a match given the knowledge of the two lineups.

This paper is concerned with creating a tool that allows football team management to do analysis on their pool of players and thereby generate a ranking based on that analysis. This give rise to two main questions, the first of which being, “What is being analyzed and whether it could help construct a ranking?” and the second being, “How is this analysis being performed?”

We focus on football (soccer) for it is arguably the most popular and one of the most unpredictable games in the world where the occurrence of an upset is arguably more likely than any other sport. The unpredictability factor in this sport will make this project challenging for us to complete.

Van Haaren et al. [1] have discussed the limitation of data available for the purpose of analysis in football. Most techniques for predicting football match outcomes have been derived from methods regularly used by statisticians. Most of these include estimation techniques that used parameterized models. The values of these parameters are learned (or estimated, from a statistical standpoint) from scores of a history of football matches. The authors give two difficulties with this approach:

- 1) Match statistics for club football are usually not publicly available, in contrast to American sports like basketball and baseball, where they are available in detail online. That has led to a profusion of interest in those sports than in football.
- 2) It is not obvious how to derive meaningful measures and statistics from football matches [1].

Our interpretation of the second problem lies in the fact that these sports are fundamentally different. In baseball it is relatively simple because it is a series of individual matchups between a hitter and a pitcher. In basketball, it is a more

challenging than in baseball because there is a lot of scoring and rolling substitutions. Still, this leaves us with a wealth of data of different groups of players on the pitch. The problem is substantially more complex in football due to:

- the dearth of scoring,
- the dynamic nature of the play, and
- data sparsity.

For example, no two teams play 4-4-2 the same way. Some play narrow, some wide - some use a diamond in the midfield, while some use a double-pivot. Things get even more complex with the variations of 4-5-1, which is the most used formation across the top 5 leagues in Europe in recent years. So if we account for all the variations, we will obtain with very sparse datasets.

In recent past, association football has undergone a metamorphosis in terms of professionalization and the incorporation of technical advances [3]. The previous generation of predictive models for football almost exclusively worked on the basis of the number of goals scored in a match. As an example, Maher's [4] model predicts outcomes of football matches given two lineups. The model proposed uses individual Poisson variables to calculate the score for both teams separately. Dixon and Coles [5] adopt Maher's model while proposing some changes. They show that there exists a strong dependency between individual scores in low-scoring football matches.

Current football-related prediction techniques are typically applications of statistical methods that fail to exploit the full range of information in the available data and are limited to learning from football match histories [6]. In spite of the vast popularity of football, research has been lacking in terms of more sophisticated models that take the numerous events that influence a match result (e.g., a red card) as well as time-dependent and positional information (e.g., dribbles and passes) into account.

The biggest challenge is that we do not know which data are trustworthy - "clean" in machine learning parlance, and which contain duplicates, invalid records and inaccurate data entries. According to Gartner, a shocking proportion of all business data are inaccurate [7].

II. METHODOLOGY

The work presented here is mainly in two major directions:

- Player rating.
- Team predictions.

The main goal is what we call *team selection*. Team selection includes mainly both tasks but player rating has to be changed from independent player rating to relational player rating that contribute in team winning and Team predictions are less important and instead we have to find a combination that maximizes the team's winning chances. This approach is converse of team prediction. A major challenge is to compete with the human mind. Coach will think of a player in a different way rather than statistics but we want to provide a statistical comparison to use the power of stats in determining the importance of a player. We take advantage of neural

networks to learn relative ranking criteria from individual players' statistics.

In the following we first described our prediction work and then introduced our learning methodology for the given predictive model and then finally we selected an optimal squad for a given team.

A. Predictive Model

The predictive model is used to generate player contribution for an individual player relative to others in his/her team. However a naïve approach of using the winning ratio might be misleading for it might be too similar for multiple players. What is the legitimation of winning ratio arguably being the correct true label? To address this concern, we used a semi-supervised approach to find player importance. For this purpose we trained a neural network. Neural network analyzes the input features of all the players, separately for selected and opponent team and uses the final outcome to generate two individuals team scores. Both outcome are combined to get the final win/loss. The learning process assigns weights to the input links of each player. So in this manner we would be able to get an evaluation measure of each player with respect to his position. These weights are kept non-negative by saturating them at 0 so the impact of the player is in favor of his team rather than with the opponent.

B. Team Selection

In the team selection using the weights learned by the neural network team we generate player rating according to its playing position in the team. The quantified attribute of players are first multiplied by the weights generated by the first visible and first hidden layer of neural network. This will generate scores for individual player for his performance in a particular team.

$$P_{si} = \sum_{j=1}^n (D_j \times 1j)_i \quad (1)$$

$$\mathbf{P}_{si} = \mathbf{D}_i \cdot \Theta_1 i \quad (2)$$

where, P_s = Player score, j = Attribute of one player, θ_1 = Weights from first hidden layer of the neural network.

The individual player scores are then multiplied by the weights generated by the first and the second hidden layer of the neural network.

$$T_s = \sum_{i=1}^n (P_{si} \times \theta_2 i) \quad (3)$$

$$\mathbf{T}_s = \mathbf{P}_i \Theta_2 \quad (4)$$

where, T_s =Team Score and θ_2 = Weights from second hidden layer of the neural network.

After summing up the calculated scores we will get magnitude of how good a particular team is. This procedure is initially applied for the opponent team to get its best team score, secondly we apply the same method for subject team which will give us all the possible combination rated higher

than the opponent team. By this we achieve our target in two ways, firstly selecting a better subject team then its opponent and secondly giving multiple team combinations for subject team.

III. EXPERIMENTS

The dataset comprises four categories of players with the features listed. Note that the three categories sharing the complete feature set are mentioned together.

- *Keepers*: The feature set of keepers include age, game starts, substitute ins, saves, goals conceded, fouls committed, fouls suffered, yellow cards, red cards and wins. The last one is binary, while the rest are real-valued.
- *Defenders, Midfielders, Strikers*: The feature set of keepers include age, goals scored, goals conceded, shots taken, assists, fouls committed, fouls suffered, yellow cards, red cards and wins.

The data for Keepers include 285 samples, Defenders have 957 samples, and Midfielders have 1026 samples and Forwards have 549 samples in all. This dataset comprises 10 years of data from the English Premier League (EPL).

A. Network Architecture for Player Rating

We use neural networks for semi-supervised learning using data of matches played during the past 10 years. Random weights are initialized to the neural networks and data inputs are the features of players that played the game. The input to neural network is a set of matches played. The target variable is the match outcome for the subject team. The architecture consists of 11 input nodes one for each player. The first hidden layer has the same number of neurons as input layer with one to one connection and it receives the transformed sigmoid values of the initial attributes. Lastly, the output layer is the combination of both of these. The accuracy is calculated thus: (true positives + true negatives) / total test data.

The results show 54% accuracy with the given set of data. With more data, this accuracy is likely to improve. Fig. 1 shows the architecture. Our process is justified due to the fact that we do not need to rate individual players on their own individual abilities rather we have to maximize the team winning probability. The train and test accuracies on 5-fold cross validation and their averages are shown below:

TABLE I: Accuracies for the player rating neural network.

ALPHA	1	0.1	0.3
a1	52.34	51.44	56.431
a2	54.32	54.22	55.034
a3	55.65	55.63	55.697
a4	53.23	52.43	49.877
a5	54.54	51.522	53.454
AVG	54.016	53.0484	54.0986

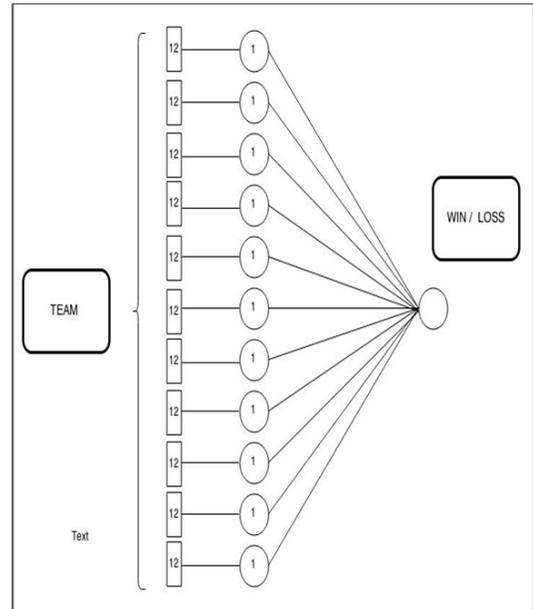


Fig. 1: The player selection neural network architecture.

B. Network Architecture for Team Predictions

We then revised our architecture by considering the players of the opponent team as well. New architecture consists of 22 input nodes, 11 for each team’s players and an additional neuron for home/away value. First hidden layer also has the same number of neurons and it receives the transformed sigmoid values of the initial attributes. The second hidden layer then combines all players of subject team into one neuron and all players of opponent team into another neuron. Lastly, the output layer is the combination of both of these. Fig. 2 shows the neural network architecture. Fig. 2 illustrates the network. In general we can introduce arbitrary number of hidden layers before L1 and L2 of Fig. 2. Hence our model is extensible and can learn non-linear player features.

TABLE II: Accuracies for the Team Prediction Neural Network

ALPHA	1	0.1	0.3
a1	61.6774	61.2903	58.8710
a2	61.6774	55.6452	59.6774
a3	61.6774	54.8387	59.6774
a4	61.8710	55.6452	59.6774
a5	60.8065	59.6774	59.6774
AVG	60.741	57.4194	59.5161

IV. RESULTS

In the first experiment, we used the neural network with 11 players in the input layer only considering the subject team. It gave a training accuracy of 54% with 5-fold cross validation.

In the second experiment, we use the neural network with 22 players, 11 for each side (subject and opponent teams) with an additional hidden layer. It gave training accuracy of 5-fold cross validation of 60%, improving by approximately 6%.

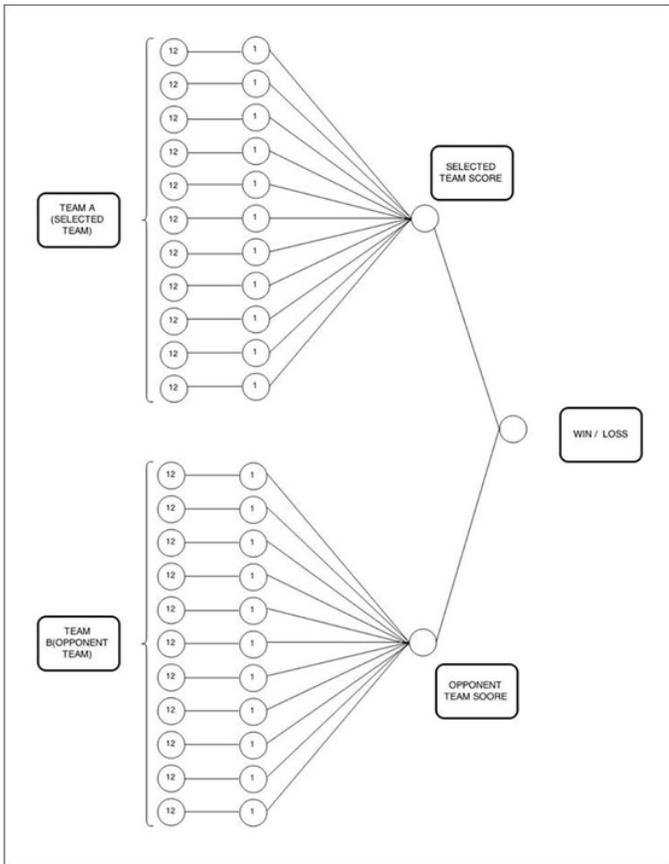


Fig. 2: The team selection neural network architecture.

V. CONCLUSIONS

We present a design of an ANN that is tailored to assist team managers in the selecting a team that will provide the best performance against a given opposition. The architecture is currently designed for a (4, 4, 2) team combination which means 4 defenders, 4 mid-fielders and 2 forwards. Secondly the system will be able to help team managers select players to buy and drop from their own and other teams in order to form team combination that can provide best possible performance.

A future direction of research could be to make this problem generic to work with all possible team formations. Another could be working on a methodology for new players that do not have previous records, to help predict their future performance.

Individual player ratings can be useful in the proper context. If a team is trying to replace a player with someone who is very similar, ratings like these can be used to short-list transfer targets. Better still, if and when we get to a point where youth and academy players have the same level of detailed data as professionals, the ratings can be used to gauge the progress and map the career trajectory of academy players.

ACKNOWLEDGEMENTS

The authors appreciate support from M. Hani, M. Jhone, M. Raza.

REFERENCES

- [1] Jan Van Haaren, Albrecht Zimmermann, Joris Renkens, Guy Van den Broeck, Tim Op De Beck, Wannes Meert, and Jesse Davis, "Machine Learning and Data Mining for Sports Analytics," Department of Computer Science, KU Leuven.
- [2] Ohlmann, Michael J. Fry and Jeffrey W., "Introduction to the Special Issue on Analytics in Sports Part I: General Sports Applications," Department of Operations, Business Analytics and Information Systems, University of Cincinnati.
- [3] Jan Van Haaren and Guy Van den Broeck, "Relational Learning for Football-Related Predictions," Katholieke University Leuven, no. 2010.
- [4] Maher, "Modelling Association Football Scores," *Statistica Neerlandica* 36(3), 1982.
- [5] Dixon, M, Coles, "Modelling Association Football Scores and Inefficiencies in the...," *Journal of the Royal Statistical Society: Series C (Applied)*, 1997.
- [6] Thomas H. Davenport and Jeanne G. Harris, "Competing on Analytics," *The New Science of Winning*. Harvard Business School Press, March 2007.
- [7] J. V. Haaren, "Analyzing Football Matches Using Relational Performance Data," Department of Computer Science, Katholieke Universiteit Leuven, Celestijnenlaan 200A, 3001 Heverlee, Belgium.
- [8] James Albert, Jay M. Bennett, "Curve Ball: Baseball, Statistics, and the Role of Chance in the Game". Springer. pp. 170171, 2001