

Context Aware SmartHealth Cloud Platform for Medical Diagnostics

Using Standardized Data Model for Healthcare Analytics

Sarah Shafqat, Almas Abbasi

Department of basic & Applied Sciences,
International Islamic University (IIU),
Islamabad, Pakistan

Muhammad Ahsan Qureshi, Tehmina Amjad

Department of basic & Applied Sciences,
International Islamic University (IIU),
Islamabad, Pakistan

Muhammad Naeem Ahmad Khan

Department of Computer Science,
Shaheed Zulfiqar Ali Bhutto Institute of Science and
Technology (SZABIST),
Islamabad, Pakistan

Hafiz Farooq Ahmad

Department of Computer Science, College of Computer
Sciences and Information Technology (CCSIT)
King Faisal University (KFU)
Alahsa 31982, Kingdom of Saudi Arabia

Abstract—Healthcare has seen a great evolution in current era in terms of new computer technologies. Intensive medical data is generated that opens up research in healthcare analytics. Coping with this intensive data along with making it meaningful to deliver knowledge and be able to make decisions are the most important tasks. To deduce the authenticity of the data on basis of precision, correction, associations and true meaning is important to validate the understanding of correct semantics. In case of medical diagnosis to form accurate understanding of associations while removing ambiguity and forming a correct picture of the case is of utmost importance. To come up with the right metrics for the diagnostic solution we have explored the known criteria to validate healthcare analytics techniques involved in formation of diagnosis that results in betterment and safety of patients under observations and heading towards possible treatments. In this work, we have proposed a thematic taxonomy for the comparison of existing healthcare analytics techniques with emphasis on diabetes and its underlying diseases. This analysis lead us to propose a data model for hybrid distributed simulation model for future Context Aware SmartHealth cloud platform for diagnostics. This platform is designed to inherit smartness of unsupervised learning which in turn would keep updating itself under supervised learning by qualified experts. Finally, the accuracy would be determined using HUM approach with biomarkers or a better accuracy model than AUC. The recommended action plan is also presented.

Keywords—Healthcare analytics; medical diagnostics; HL7; cloud platform; SmartHealth; big data

I. BACKGROUND

Going through some early histories of medical informatics [1] we get to know of the major research that was carried out in 1977 in Bosnia and Herzegovina. In it the periodic data analysis of healthcare services and performance available in Bosnia and Herzegovina was done and later in 1982, the first Local Health Information System (LHIS) with health databases with supervision of 6000 citizens was tested. Izet Masic and

Arif Agovic, an electronics engineer in Energoinvest Ltd. in Sarajevo, were known as the creator and pioneer of Medical Informatics in Bosnia and Herzegovina. A Healthcare Informatics Society BiH was established in 1987. Realization of 'Development of Information System of healthcare B&H in circumstances of electronic data manipulation' project started in 1986 after approval by the Executive Board of Association of healthcare communities B&H. It was planned to be in three phases: (i) first phase (1985-1990), (ii) second phase (1991-1995), and (iii) third phase (1996-2000). Several trials and projects failed during their inception or due to war.

Only, biggest progress was seen in pharmaceutical industry where 43 pharmacies in Sarajevo centrally connected for receipt collection and analysis. It took three years of testing and Izet Masic finally defended it in his Master's thesis [1] but the later planned activities got interrupted during and after war (1992-1995) and due to lack of funds. A need was there to introduce health informatics in medical profession to train medical specialists and physicians to be able to give quality services in healthcare using technologies that have quantitative and qualitative growth in diagnostics and therapy. In 1985, Professor Izet Masic launched a separate course "Informatics and Economics in Health" of 30 hours' duration and some of these postgraduate students became MSc and PhDs in this subject who were then able to offer their services in B&H universities and abroad. After the war (1992-1995), the B&H went through a very tough time with lack of electricity, gas, water supply and food. And, during these circumstances Society for Medical Informatics carried out eight scientific and professional events where 500 papers were presented and published in the proceedings. That was a miracle.

From then onwards several proceedings and congress participations were done. Finally in 1993, SMI BiH launched the professional and scientific journal Acta Informatica Medica (AIM) and it is being published continually since then [1]. In Bosnia and Herzegovina, there is also a library system BiH-

SBMNI been established during 1984 to 1990, by a group of medical librarians and informatics professionals in BiH. Currently, the interest is in establishing higher education in the field of Medical Informatics and the question revolves around Cathedra for Medical Informatics, University of Sarajevo. Several surveys have been carried out to analyze the level of higher education in medical informatics and future development required.

From then onwards, the healthcare informatics is rapidly adopting current trends of cloud computing to store big data coming in from miscellaneous platforms; hospitals, social networks, IoT and other wearable devices, etc. The study in [2] led researchers community to form consensus for adopting cloud computing in hospitals in a strategical manner as part of smart city project led by ICT [3] in Bandung city of Indonesia [4].

Currently, the use of ICT in hospitals [3] for playing strategic role is a viewpoint that is being worked out for rightfully implementing healthcare analytics in the field. Exploration through exploitation is carried out extensively to enhance hospital performance in administration as well as services sector; prediction, diagnosis, treatment, etc.

The need is realized to conquer these problems through computation [5]. The use of computation specially in medical diagnostics is a complex task. Till now to expect a complete diagnostic system is understood as unrealistic. But, no matter how much complex it is advances are being made using artificial intelligence (AI) techniques. Computers have advantage over humans as they do not get fatigued or bored. Computers update themselves within seconds and are rather economical. If automated diagnostic system is made such that it would take care of routine clinical tasks in which patients are not too sick, then doctors would be free to focus on serious patients and complex cases. For analysing complex medical data, the use of artificial intelligence is known to be more capable [6]. The way artificial intelligence exploit and relate the complex datasets giving it a meaning to predict, diagnose and treat the particular disease in a clinical setup. Several artificial intelligence techniques with significant clinical applications are reviewed. Its usefulness is explored in almost all fields of medicine. It was found that artificial neural networks were the most used technique while other analytical tools were also used and those included fuzzy expert systems, hybrid intelligent systems, and evolutionary computation to support healthcare workers in their duties, and assisting in tasks of manipulation with data and knowledge. It was concluded that artificial intelligent techniques have solution to almost all fields of medicine but much more trials are required in a carefully designed clinical setup before these techniques are utilized in real world healthcare scenarios. Artificial intelligence (AI) is termed to be part of science and engineering known to exhibit computational understanding to be said as behaving intelligently through the creation of such artifacts that form the stimulus in it. Alan Turing (1950) explained intelligent behaviour [6] of a computer that could act as any human for cognitive tasks applying logics (right thinking) and this theory was later termed as 'Turing Test'. The application of AI techniques in the field of surgery was first experimented in 1976 by Gunn, exploring the diagnosing of

abdominal pain with computer analysis. The challenge lies in collecting, analysing and applying all the medical knowledge for solving complex medical problems. Medical AI relates to developing AI programs such that they help in forming the diagnosis, making therapeutic decisions and predicting the outcome. Artificial Neural Networks (ANN) on the basis of their recognition for classifying and pattern identification have been known to be widely used for solving several clinical problems.

This paper is to highlight important factors and criteria required for establishing successful and trusted means of automated medical diagnostics through healthcare analytics. The challenge lies in development of automated medical diagnostics to reap its long term benefits to human kind is enough of motivation to dig deep into this domain.

II. LITERATURE REVIEW

Previous research confirms [3], [7], the successful application of analytics is found to be effective after significant degree of digitization accomplished. It was also recognized that the potential of analytics in clinical domain is stronger than in the administrative domain. Specific to automate medical diagnosis, the artificial neural networks (ANN) approach is greatly studied in combination with fuzzy approach [5]. There are some most common risks and precautions associated when forming a medical diagnosis [5]:

- To reach a well-established diagnosis a physician is required to be well versed with some very experienced cases and that experience does not come through completing academics but after lots of experience in a specialized field or disease.
- In case of new or rare disease even the experienced physicians feel to be at the same level as an entry level doctor.
- Humans are good at observing patterns or objects but fail when need to find the associations or probabilities for their observations. Here computer statistics helps.
- Quality in diagnosis relates to physician's talent and years of experience.
- Doctor's performance gets affected with the emotions or fatigue.
- To train the doctors in specialties is expensive as well as a long procedure.
- Medical field is always evolving with new treatments and new diseases are coming up with time. It is not easy for a physician to keep abreast with so much change and new trends in medicine.

Currently, 14 hospitals in Italy were observed for their activities and users' involvement in the use, adoption, and improvement of ICT infrastructure and services based solutions [3]. The prime data source was 107 semi-structured interviews of level C and other hospital informants in a time period of 3 years. There are three paths taken for exploration and exploitation paradox management for enhancing hospital performance: (i) digitization of the assets utilized within

hospitals, (ii) ICT-based integration among healthcare stakeholders, and (iii) the disruption of clinical and administrative decision making through the use of analytics, coping the conflicting demands attached to medical sector. Analysing various case studies, the possible wastage of energies is saved in ICT-based innovation and prioritization of possible paths is achieved when moving towards the management of exploration and exploitation paradox. This research [3] divides its analysis activities in three domains: (i) ICT introduction process, (ii) hospital, and (iii) healthcare system within a hospital. The analysis through different case studies produces evidences to demonstrate the theory in Fig. 1 [3]. As shown in Fig. 1 [3], the three complimentary ICT-based paths maintaining a balanced exploratory and exploitative stimuli act in a given approach:

- i. Digitization of assets utilized within hospitals ((1)—(2))
- ii. ICT-based integration among healthcare stakeholders ((2)—(3)—(5))
- iii. The use of analytics for disruption of clinical and administrative decision making ((2)—(4)—(5))

The three paths are successful in managing the exploration-exploitation paradox for short as well as long term. It is found that their overall effectiveness is felt stronger in the clinical domain rather the administration domain. The other limitation observed is that it is not easy to distinguish and separate these interlinked forces. The whole operationalization has to keep into consideration all the pros and cons and there are many factors attached to its success.

Disease outbreaks are happening all over and at all times. Computational prediction, identification, confirmation and responsiveness to these diseases is important as well. Therefore, Predictive Analytical Decision Support System (PADSS) [8] integrates in a cloud based healthcare platform that is Message Oriented Middleware (MOM). It connects healthcare organizations to share patients' data using a customized Health Level Seven (HL7) platform having Fast Healthcare Interoperability Resources (FHIR) specification.

Considering chronic kidney disease (CKD) [9], accurate prediction with time is important for lowering cost and mortality rate. Adaptive Neurofuzzy Inference System (ANFIS) is proposed that uses real clinical data of 10 years for newly diagnosed patients with CKD to predict renal failure time frame. It is deemed as highly successful measure to predict GFR variations in long future periods within uncertain body condition and dynamic nature of CKD progression. The limitation of this experiment was only that urine protein was missing as variable input for predicting GFR in 6 to 18 months' evaluation time.

For detection in medical imaging [10] is also required in healthcare practice and Automated Computer-Aided Detection (CADe) tool is there. There is a high rate of false positives (FP) as well that would add to cost for achieving success (high true positive rate) in highly sensitive cases per patient using state-of-the-art methodologies. But for high false positive or false negative rate that would add to the low sensitivity the application of CADe is not seen in the clinical practice. An updated multi-tiered hierarchical CADe system improved in

performance and was named as deep convolutional neural networks (ConvNets). ConvNets is open for advancements as it has been commercially tried on natural images as well as biomedical applications for detection of mitosis in digital pathology. It learns from supervised training to detect features from images through two cascaded layers for filtration using convolutional filters. It can be applied to detect from two and two and a half dimensional observations.

Usage of ConvNets through multiple 2D, 2.5D and 3D representations on CADe problems proves its success by avoiding overfitting analogous data. ConvNets in applications build better accurate classifiers for CADe systems pruning FPs maintaining high sensitivity recalls. Even with the threshold value of 95% sensitivity with 1FP per patient it is not found optimal for colonic polyp CADe System for polyp sizes between 6mm to 9mm. Better results are observed with the polyp sizes larger than 10mm.

Diagnostic accuracy is determined by applying statistical measures in medicine [11] in the context of multi-category classification.

To determine the success of data fuzzification an experimental setup was created to analyse the medical diagnosis done by physicians and automate it with machine implementable format. Eight different diseases were selected for extracting symptoms from few hundreds of cases and MLP Neural Networks was applied [5]. Later results were discussed to conclude that effective symptoms selection for data fuzzification using neural networks could lead to automated medical diagnosis system. A diagnostic procedure is work of an art of specialized doctors and physicians. It starts from patient's complaints and discussion with doctor that leads to perform some tests and examinations and on basis of results the patient's status is judged and diagnosis is formed. Then the possible treatment is prescribed. Patient remains under observation for some time where the whole diagnostic procedure is repeated and refined or even rejected if needed. We all are aware of the complexity of forming a medical diagnosis as even the profession requires twice the study than other professions. There are diversified symptoms history that is caused for diversified reasons. All these causes have to be included in the patient history.

To propose [5] a medical diagnostic system several interviews were conducted of expert doctors getting some diagnostic flow diagrams of various diseases and associated list of symptoms. The dataset was created such that hundreds of patients were tested against 11 symptoms (features) and 9 diseases (classes). First eight classes were associated with specific illnesses and the ninth one was defined for normal/healthy person. Multilayer perception (MLP) neural network was used with application of back propagation GDR training algorithms. The simulation was developed on MATLAB with NETLAB toolbox. A three-layer feed forward perceptron was kept to keep the structure simple and focus on the hidden nodes with the training iterations as variable parameters. The performance of classifier was tested while changing the parameters. Also, feature fuzzification rule [5] on the accuracy of diagnosis was investigated. With k-folding scheme, the lack of dataset was overcome to give better

accuracy for validation. So, the training procedure was repeated $k=5$ times with 80% as training dataset and 20% as testing. The mean was taken for all the outcomes of 5 tests. 88.5% best accuracy was achieved with 30 nodes at the hidden layer. Then, membership-based fuzzification scheme [5] was applied to the dataset converting it to fuzzified set of symptoms. A linear membership function was selected for each symptom with experts' consultation. From three to five linguistic variables were linked to each symptom and classification tests were repeated. Maximum performance was achieved with diagnostic accuracy of 97.5%.

ANNs an exploration of Baxt is also found a successful technique in clinical domain [6] for diagnostics. He came up with a neural network model to diagnose acute myocardial infarction accurately validating his work later with similar accuracy. ProstAsure Index is a classification algorithm extracted from ANN that classifies prostates as benign or malignant. This model gave the diagnostic accuracy of 90%, sensitivity of 81% and 92% specificity. Other relevant surgical diagnostics with application of ANNs are appendicitis and abdominal pain, glaucoma, retained common bile stones and back pain. PAPNET [6] is another computerized screening application based on ANN that assists cytologist in cervical screening and this application was commercially available. Thyroid, breast, oral epithelial, gastric, urothelial cells, peritoneal effusion cytology and pleural enjoyed varied level of diagnostic accuracy with application of ANNs. In the field of radiology, the inputs to ANNs are both human observations and digitized images. ANNs are good at analysing plain radiographs, CT, ultrasound, radioisotope and MRI scans. Various wave forms like ECGs, EEG, EMG, and Doppler ultrasound as well as hemodynamic patterns are well interpreted using ANNs pattern recognition ability. Correct prognosis is also very important to carry out appropriate treatment strategy and follow up. ANNs exploiting non-linear relations between variables are well suited to analyze complex cancer data and it can predict survival of breast and colorectal cancer patients better than colorectal surgeons. ANNs have also been applied to predict outcomes of lung and prostate cancer.

Then there is another AI domain 'fuzzy logic' [6]. It is the science of thinking, reasoning, and inference been applied in real world phenomenon of varying degrees. It sees beyond black and white into the varying shades of grey. Medicine is known to be a continuous field and the data is imprecise and thus fuzzy logic applies to it. Fuzzy expert systems reach to the conclusion through its 'if-then' structure modelling. Fuzzy logic was found to be better than multiple logistic regression analysis for diagnosing the patients with lung cancer having tumour. Then, fuzzy logic has been applied to acute leukaemia, and breast and pancreatic cancer diagnosis. It has also been applied to capture ultrasound images of breast and ultrasound, and CT scan images of liver lesions and MRI images of brain tumours. It is also being used to predict survival of patients with breast cancer. Fuzzy controllers are there for administration of vasodilators that controls blood pressure, and anesthetics in the operating room.

Evolutionary computation is yet another AI domain that mimic the natural selection and survival of the fittest

mechanism in natural world to solve the problems and its most widely used form in medical field is 'genetic algorithms' [6]. Genetic Algorithms are applied to reach diagnosis, prognosis, processing signals and analyse medical images, and plan and schedule. They are used to predict outcomes in critically ill patients, melanoma, lung cancer and response to warfarin. They are also used to analyse mammographic micro calcification, MRI segmentation of brain tumours for measuring efficacy of treatment strategies, 2-D images to diagnose malignant melanomas.

Then finally there is hybrid intelligent systems [6] combining the strengths of ANNs, fuzzy logic, and evolutionary computation.

In spite of many different AI techniques [6] that can be readily used for solving several clinical problems there is hindrance in its acceptance by clinicians who would mostly reject the biochemical results produced by autoanalyzer or images resulted from magnetic resonance imaging. So the obligation lies at the researcher's end to authenticate and validate its successful application in real clinical setup. There is no doubt that this future technology resulting in 'medical intelligence' would add to the ability of future clinician. It is concluded that for managing the exploration-exploitation paradox for short as well as long term the overall effectiveness is felt stronger in the clinical domain rather the administration domain. The other limitation observed is that it is not easy to distinguish and separate the three interlinked forces identified in Fig. 1 [3]. The whole operationalization has to keep into consideration all the pros and cons and there are many factors attached to its success. It is also observed that there is no universal solution till now for healthcare diagnostics meeting a standard protocol.

III. ANALYSIS OF PREVIOUS DIAGNOSTICS SYSTEMS

For managing the exploration-exploitation paradox for short as well as long term, it is found the overall effectiveness is felt stronger in the clinical domain rather the administration domain. The other limitation observed is that it is not easy to distinguish and separate the three interlinked forces identified in Fig. 1 [3]. The whole operationalization has to keep into consideration all the pros and cons and there are many factors attached to its success. The successful application of PADSS [8] is a strong motivation to take it to the next step in healthcare towards medical diagnostics in compliance with HL7 standards. It also demonstrates the successful utilization of Google Cloud Platform with Google BigQuery where there is need to make SQL queries more efficient by tuning. While evaluating of various healthcare analytics platforms it is visualized that mostly artificial neural networks or fuzzy logic algorithms form the basis of several proposed techniques and recent research has included deep learning algorithms as deep convolutional neural networks ConvNets on CAde problems proved its success.

Further it is observed that by selecting symptoms of eight different diseases a dataset containing few hundreds of cases was created and Multilayer Perception (MLP) Neural Networks was applied [5]. Later results were discussed to conclude that effective symptoms selection for data fuzzification using neural

networks could lead to automated medical diagnosis system. Simulation was developed on MATLAB with NETLAB toolbox. A three-layer feed forward perceptron was kept to keep the structure simple and focus on the hidden nodes with the training iterations as variable parameters.

Recently, evolutionary computation is another AI domain exploited that mimic the natural selection and survival of the fittest mechanism in natural world to solve the problems and its most widely used form in medical field is 'genetic algorithms' [6]. Genetic Algorithms are applied to reach diagnosis, prognosis, processing signals and analyze medical images, and plan and schedule. Then finally there is hybrid intelligent systems [6] combining the strengths of ANNs, fuzzy logic, and evolutionary computation.

These previously proposed healthcare analytics systems and techniques have been validated against the accuracy achieved. There are different metrics for evaluating accuracy level. It is seen that sensitivity of CADe [10] is improved from 57% to 70%, from 43% to 77%, and 58% to 75%. Statistical model based on HUM analysis [11] over biomarkers showed 65% accurate classification. ANFIS [9] is evaluated by accessing Normalized Mean Absolute Error that is lower than 5%. PAPNET [6] is said to have diagnostic accuracy of 90%, sensitivity of 81%, and 92% specificity. In MLP neural network [5] 88.5% best accuracy was achieved with k-modelling and maximum performance was achieved with diagnostic accuracy of 97.5% with membership-based fuzzification scheme. Further accuracy results are depicted in Table I.

TABLE I. TAXONOMY MATRIX FOR EVALUATION OF HEALTHCARE ANALYTICS

Authors	Diseases	Proposed Technique/System	Algorithms/Platforms	Tested	Accuracy Level
Gastaldi and Corso, 2015	X	Management of exploration and exploitation paradox in 14 Italy Hospitals	X	Case study using 107 semi-structured interviews	X
Ramesh et al., 2004	Acute myocardial infarction and Surgical Diagnostics including appendicitis and abdominal pain, glaucoma, retained common bile stones and back pain	Neural network model	Artificial Neural Networks (ANN), ProstAsure Index	Simulation	Diagnostic accuracy of 90%, sensitivity of 81%, and 92% specificity
Ramesh et al., 2004	Thyroid, breast, oral epithelial, gastric, urothelial cells, peritoneal effusion cytology and pleural	Neural network model	ANN	Simulation	Diagnostic accuracy of 90%, sensitivity of 81%, and 92% specificity
Ramesh et al., 2004	Cervical screening	PAPNET	ANN	Experimented	Diagnostic accuracy of 90%, sensitivity of 81%, and 92% specificity
Alan Turing, 1950	X	Turing Test	Cognitive behavior	Theory	X
Gunn, 1976	Diagnosing of abdominal pain	Computational analysis	AI techniques	Applied in field of surgery	X
Neto & Ferraz, 2016	X	PADSS integrated in cloud based healthcare platform	Google Cloud Data Store	Prototype	Compliant to HL7 standard
Boukenze et al., 2016	Chronic kidney diseases (CKD)	Predicts the chronic kidney diseases (CKD)	C4.5 learning algorithm using classification model with ORANGE or TANAGRA	Simulation	63%

Norouzi J. et al., 2016	Chronic kidney diseases (CKD)	Adaptive Neurofuzzy Inference System (ANFIS)	Neurofuzzy	Simulation	Normalized Mean Absolute Error lower than 5%
Roth et al., 2015	Lymph nodes, sclerotic metastases and colonic polyps	Computer-Aided Detection (CADe)	Deep convolutional neural networks (ConvNets)	Experimentation on three datasets	Improves the sensitivity of CADe from 57% to 70%, from 43% to 77%, and 58% to 75%
Li J., 2013	Synovitis and leukemia	Statistical Accuracy Model	(HUM) based analysis	1000 simulations using Monte Carlo	65% accurately classified
Moein et al., 2009	Eight diseases classified with ninth one a healthy person	Medical diagnostic system	Multilayer perception (MLP) neural network	Simulation was developed on MATLAB with NETLAB toolbox	88.5% best accuracy was achieved with k-modeling , maximum performance was achieved with diagnostic accuracy of 97.5% with membership-based fuzzification scheme
Izet Masic, 1977	X	First Local Health Information System (LHIS) in Bosnia and Herzegovina	X	It took 3 years to process and test health data by a specially designed software package ARCHIVE on a Personal Sinclair QL computer that was later redesigned in Clipper, DBASE III, and FoxBASE	Realization of Healthcare Informatics was formed
Sordo, 2002	Screening of Pap (cervical) smears	PAPNET	ANN	In cardiovascular domain, neural networks were trained using ECG measurements of 1120 patients suffering from heart disease and 10,452 normal people without any history of heart attack	15.5% more sensitive than the interpretation program and 10.5% more sensitive than the cardiologist for diagnosing abnormalities
Sordo, 2002	Predict breast cancer	Entropy Maximization Network (EMN)	ANN	Experimented	Most accurate in predicting five-year survival of 25 cases
Sordo, 2002	Cancer	Drug Development Solution	ANN	Simulation	Correctly classified 91.5% from anticancer agents (drugs)

A. Comparison of Analytical Tools and Techniques used in HealthCare

In Fig. 1, the taxonomy, clearly demonstrates that there is an exemplary work already been done in developing and using various healthcare analytics techniques for prediction, detection and diagnosis of various chronic diseases.

There are various artificial intelligence and machine learning algorithms been employed that contributed in emergence of different clinical systems. There are some specific tools and platforms used that are mostly believed to be state-of-the-art technologies including google cloud platform, MATLAB, Weka, and design methodology like DFDs. But in most cases these systems have been proposed and only limited number of systems have been commercialized yet. Major example of commercially used systems is Papnet that is for diagnostics of cervical cancer. This is due to the limitation posed by the level of accuracy achieved and in acceptance from clinical domain.

If we carefully elaborate the picture we see that there have been systems developed for focused areas like; prediction, detection and diagnosis with respect to the particular disease mostly chronic. Narrowing our focus, we see that PADSS is a prediction system proposed for large scale patient data that was distinguishable by the standard nomenclature understood by HL7 using FHIR specification. It became part of the Global Public Health Intelligence Network (GPHIN) project that was supported by WHO and Centers for Disease Control. This is the most unique and large scale project that underwent in medical field. The platform used was Google cloud platform and Google bigquery. Its basis was previous system PREVENT that used XML-based syntax and object-oriented approach for which it was criticized.

Then for detection there is hierarchical two-tiered CADE system been discussed here that uses ConvNets a deep convolutional neural networks algorithm to detect lymph nodes and colonic polyps. Analogous data from three datasets is represented in 2D, 2.5D and 3D representation.

For diagnosis there is system that is prominent is Papnet which is commercially available. For other diagnostic measures there are fuzzy expert systems and hybrid intelligent systems that use various state-of-the-art algorithms and tools.

In 2009, there was conducted an experiment for proposing a diagnostic system that included eight different diseases. The algorithm used was Multilayer Perception (MLP) neural networks within MATLAB integrated with NETLAB. The accuracy achieved was 97%.

Accuracy is determined in all these systems through measuring sensitivity or using statistical accuracy tools like HUM or biomarkers. The most authentic accuracy measure would be to compare the system with a known standard or the system should be in compliance with a standard as in PADSS.

B. Application of Healthcare Analytics on Data of Diabetes and Underlying Diseases using Various Learning Algorithms

Taking the comparison done in Table II of various healthcare analytics systems we aim to provide universal

diagnostic solution for diabetics and patients forming risk or suffering from underlying other diseases of liver cirrhosis, kidney disease, cardiovascular complications, etc. [12] already refers to a detailed systematic review on application of data mining and machine learning techniques to perform analytics on diabetes mellitus (DM) for prediction and diagnosis, finding complications associated to it and its linkage with the genetic background and environment to assist betterment in healthcare management.

It is known that the patient diagnosed with diabetes has to be very careful in keeping the blood sugar controlled otherwise there are chances that long-term diabetes may develop certain complications in form of some known chronic diseases (mayoclinic.org) mainly; cardiovascular disease, brain stroke, nerve damage (neuropathy), kidney damage (nephropathy), eye damage (retinopathy), foot damage, worse skin conditions, hearing impairment, alzheimer's disease, etc.

IV. AIMS AND OBJECTIVES

LHS [17] is in its development phase with a lot to compute and found extendable. Based on the thematic taxonomy in Tables I, II and III the optimal diagnostics system for diabetics patients those are threatened by underlying diseases is yet a problem to be achieved with many challenges.

1) Data and Simulation Model for a universal diagnostic system primarily for diabetes leading to other chronic diseases with the perspective of efficiency and reliability with maximum accuracy is required.

2) Providing a test bed with automated real time patient data to input into diagnostics system is an important concern for validation through simulation.

3) diagnostics system for diabetic and its underlying diseases is not present in current scenario and has our attention.

4) Finding an optimal big data healthcare analytics technique for diagnostics in compliance with HL7 standard still remains a challenge.

TABLE II. COMPARISON OF VARIOUS HEALTHCARE ANALYTICS SYSTEMS

	Platform	Model	Algorithm	Tool	Standard
PADSS	Google Cloud	OLAP	X	BigQuery	HL7
Predictive Analytics	X	Classification Model	C4.5	Weka	X
ANFIS	X	ANFIS prediction Model	Genfis3	MATLAB	X
Hierarchical two-tiered CADe	CADe	X	Deep ConvNets	X	X
Medical Diagnostic System	X	X	MLP neural networks	MATLAB with NETLAB	X
PAPNET	X	Neural network model	ANN	X	X

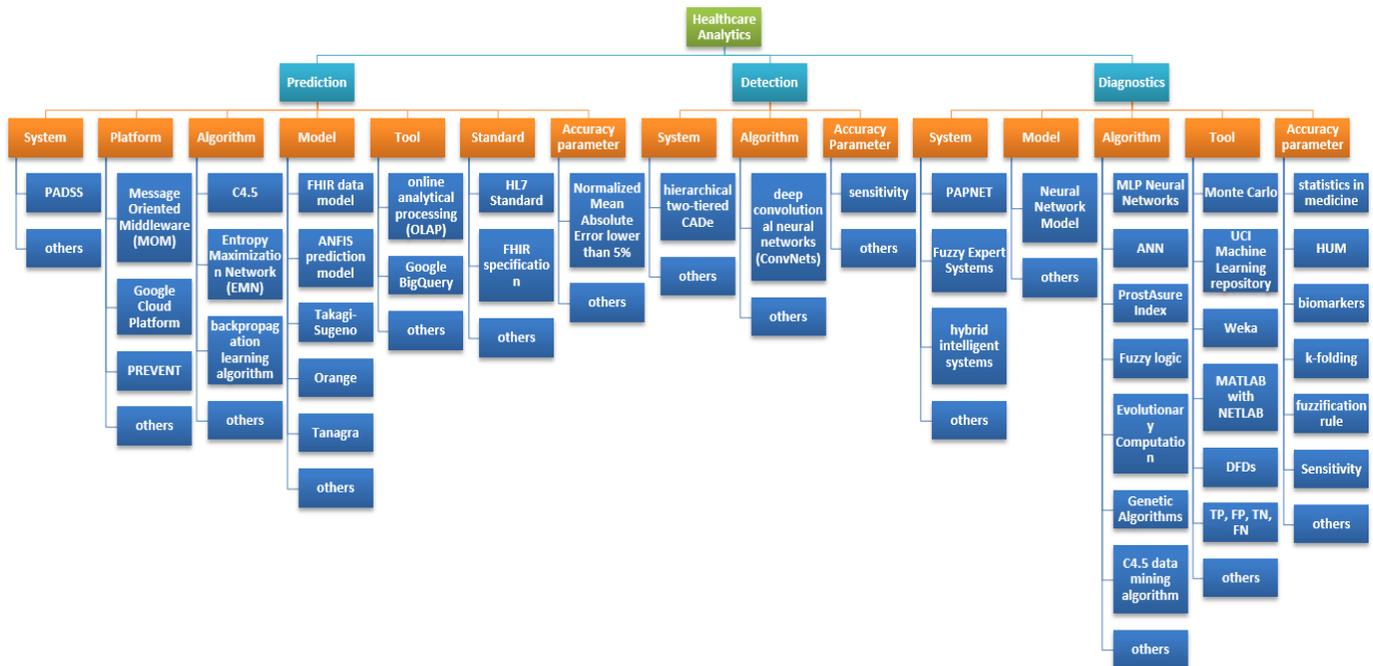


Fig. 1. Detailed taxonomy of healthcare analytics.

TABLE III. HEALTHCARE ANALYTICS APPLIED FOR DIAGNOSTICS OF DIABETES AND LINKED DISEASES

Citation	Disease	Application	Algorithm/System	Dataset	Tool/s	Accuracy
[13]	Diabetes	Diagnosis	Hybrid System (ANN & FNN)	Pima Indians Diabetes (taken from UCI machine learning repository)	Matlab 7.0.0	84.2%
[13]	Heart Disease	Diagnosis	Hybrid System (ANN & FNN)	Cleveland Heart Disease	Matlab 7.0.0	86.8% but IncNet outperformed it with 90% accuracy
[14]	Diabetes, Liver, Breast Cancer, Hepatitis, etc.	Diagnosis	Hybrid Intelligent System (optimizing SVM & MLP using GSA, PSO & FA evolutionary algorithms)	11 datasets of various diseases obtained from UCI (that included Pima Indian Diabetics & BUPA Liver Disease)	Weka	Optimized MLP technique with Firefly Algorithm (FMLP) in proposed Hybrid System was found better
[15]	Diabetes, heart disease, breast cancer, iris, skin cancer etc.	Compression of Training Set	Comparison of instance selection algorithms	Wisconsin breast cancer, Cleveland heart disease, appendicitis, Iris, wine, Pima Indians diabetes, Skin Cancer	Tested for classification algorithms kNN, SVM, SSV, NRBF, FSM & IncNet	Prototype selection algorithms Explore, RMHC, MC1, LVQ, DROP2-4 and DEL, are the most effective. RMHC and LVQ algorithms also finished tests with high level of accuracy on the unseen data. ENN algorithm came at the top, especially for KNN & NRBF and ENRBF. These algorithm may stabilize the learning process, except the SSV or SVM models.
[16]	Clinical Medicine (Psychological)	Approximation and classification (Diagnosis)	IncNet is integrated with Extended Kalman Filter learning algorithm and bi-central transfer functions while rotating constant values in multidimensional spaces	Minnesota Multiphasic Personality Inventory (MMPI) test	Not mentioned	IncNet outperforms C4.5 & FSM Rule Opt. with accuracy of 99.23%

V. PROPOSED SOLUTION

This paper has started off with proposing a solution to our first problem of defining a data model and validate a proposed simulation model in [18] based on it that would give us a testbed in future to propose detailed solution.

A. Universal Data Model

As mentioned previously researchers are interested to propose Healthcare Analytics Platform for Diagnostics of Diabetic and underlying other chronic diseases like; Liver Cirrhosis. Comprehensible Knowledge Model for Cancer Treatment (CKM-CT) proposed in [19] greatly contributes to visualize the limitations posed when defining a universal data model. In [19] researchers acknowledge the close interaction and support of clinicians with technologists. Experts in knowledge domain are highly recognized to assist in retrieving raw patient data in form of EHRs and transforming it into structured form using CRT algorithm integrated in CKM-CT model. CKM-CT is said to be generalized with some limitations of availability of latest technological facilities and infrastructure. Our aim is to use machine learning techniques that would assist to apply deep learning over the universal diagnostic data model that would become part of context aware SmartHealth cloud platform derived for our problem. Our confidence in proposed model is based on study of various healthcare analytics platforms earlier been used in cloud context as in PADSS mentioned in Table II.

B. Proposed Data Modelling Methodology

Our research would be an extended version to connect with LHS in future that would be in compliance with latest HL7 coding standards for coming up with medical diagnostics as in PADSS [8] that is for prediction of diseases integrated in cloud based healthcare platform. Further when setting up a benchmark for our system Adaptive Neurofuzzy Inference System (ANFIS) is seen as one of the successful approach for

diagnosis of chronic kidney diseases [9]. It is focused over a single disease where Multilayer perception (MLP) neural network [5] is tested over eight diseases with 97% accuracy level achieved. Our measure of accuracy would be based on HUM using biomarkers or a better approach than AUC. The major simulation platform for SmartHealth used would be AnyLogic as modelled in [18] and Google Cloud as in PADSS [8]. The system may further be standardized if the simulation platform is HIPAA enabled or the dataset is following or is convertible to HIPAA HL7 coding standard.

By standardized we mean the specification given by Fast Healthcare Interoperability Resources (FHIR) regulating HIPAA HL7 standard to be followed. But to converge the heterogeneous healthcare big data coming in from different platforms it has to be structured and modeled to fit FHIR data modeling guidelines. After being transformed in structured diagnostic data model if still it fails to map with FHIR specified data model fields then it may be modified to fit our built model (shown in Fig. 2). Later these modeled patients' diagnostic profiles would be clustered to train diagnostic learning model using analytics.

EHR is relatively structured form of health data. Therefore, it may be used to train the diagnostic learning model as in [20]. We would use it to cluster patient's data in three categories: (i) diabetic, (ii) prediabetic, and (iii) having family history (as shown in Fig. 3). Further nested clusters would be found for various other variables like; age, and diabetes type, etc.

Thus, learning model (Fig. 3) would be trained to find different sequences in symptoms and lab test results through finding correlations in one diagnosis or multiple diagnosis characteristics [21]-[23] of a patient in multiple visits. This diagnostics and history of diagnosis would also reach prognosis of co-occurrence [24], [25] of other chronic underlying diseases.

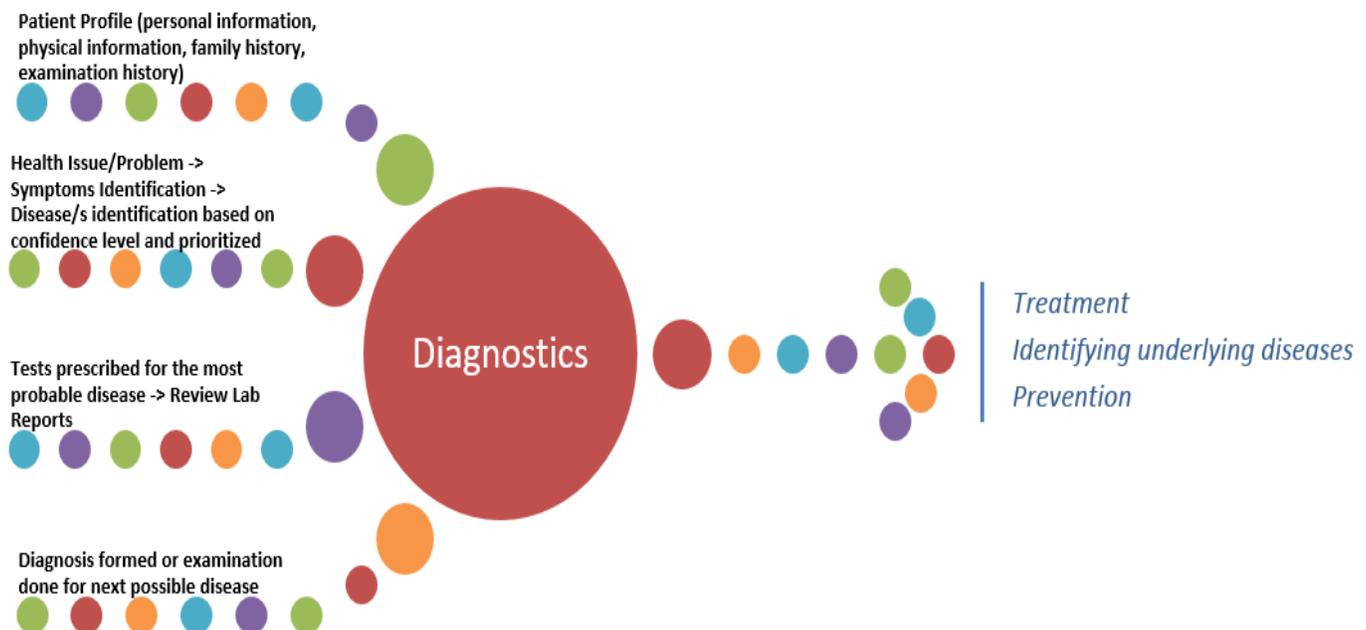


Fig. 2. Universal data model for diagnostics to apply on any disease.

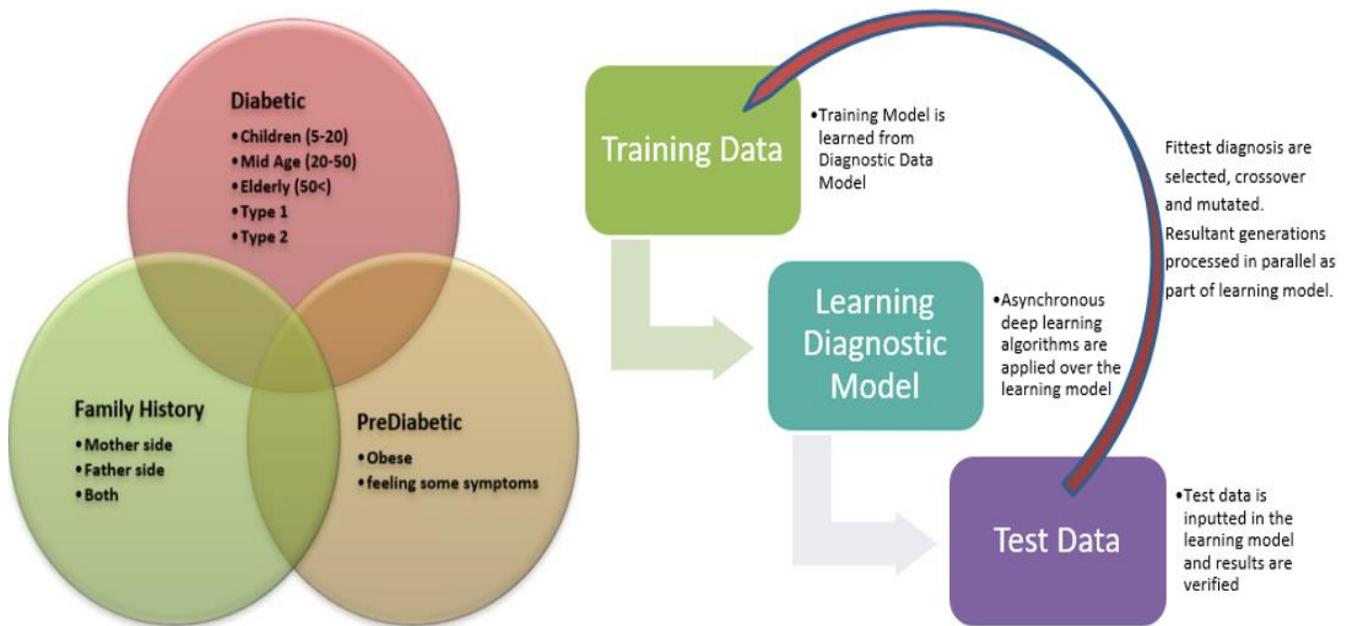


Fig. 3. Nested cluster model to develop accurate learning diagnostics model for diabetes.

VI. EVALUATING PROPOSED SOLUTION

Based on the extensive analysis of healthcare analytics platforms (Table I) detailed taxonomy (Fig. 1) is extracted for prediction, detection and diagnostics systems spanning various diseases. Moving forward we focused particularly on systems involving our domain of diabetes and linked diseases (Table III). This analysis lead us to compose our problem statement to construct a universal diagnostic system for diabetes and its underlying diseases that is context aware based on its ability to integrate with the latest technology of IoT embedded with cloud to give personalized service to patients wherever they go. To start with solvation of our problem we were able to devise a generic data model (Fig. 2) for any particular disease and in our case that would take patients data related to diabetes or other interlinked diseases. This data would be heterogeneous in nature coming from biosensors and other apps to become part of socially context aware SmartHealth cloud platform. The data would be structured with respect to its nature based on patient profile and examination history, current health issue or symptoms identified, the list of most probable diseases that are associated to a particular patient, any tests underwent and reports, etc. If the diagnosis is reached then patient gets the feedback for the next immediate action otherwise he/she is examined for next possible disease. The detailed scenario of how this data is evaluated at different stages using healthcare analytics techniques (as shown in Fig. 3) as part of hybrid distributed simulation environment within AnyLogic is demonstrated in Fig. 3 [18].

Detailed data model abstraction enabled us to build the simulation in detail in [18] and not compromising on abstract level to miss out important details. Our cloud platform considers to be integrated with societal role in future but at the moment is assumed to take input from biosensors and computing the in-formal information gathered by patients forming a patient profile graph as mentioned in [26] and shown in Fig. 1 of [8]. We have already selected our data formation

clustering method to be Random Dynamic Subsequent method proposed in [27]. Further when performing machine learning heuristics we consider the proposed deep patient representation using unsupervised deep feature learning method shown in Fig.2 of [28] to form diagnostics of diabetes and the underlying disease of liver cirrhosis. Still it is kept in supervision of human doctors not to let patients suffer any risk. The metrics to validate the accuracy of our model would be determined using HUM approach with biomarkers or a better accuracy model than AUC.

VII. CONCLUSION AND FUTURE DIRECTIONS

Based on our study of previous healthcare analytics techniques been utilized for prediction, diagnosis, treatments, and prognosis it is determined that researchers have focused over best approaches with respect to the maximum accuracy level achieved. These approaches properly integrated with the latest techniques would be used for coming up with future standardized medical diagnostics. We are working towards proposing a diagnostic system keeping in view for its commercial universal use by clinicians for various chronic diseases like; diabetes mellitus and its underline diseases particularly liver cirrhosis. The system should be such that it would be visualized for coming up with optimal healthcare analytics technique. In future the system would be enhanced exclusively for treatments or cover the diagnosis of diseases that have been left out due to limited time or limitation of proposed healthcare analytics technique.

The simulation in [18] embedded with proposed data model for SmartHealth cloud platform is found to be technically intelligent and sound in its process flow when compared in [18] to SelfServ platform and Societal Information System for Healthcare simulations [29], [30]. It acquires the smartness of unsupervised learning while at the same time it keeps it under supervision of qualified doctors that would assist in complicated diagnostic case and continuously update the

knowledgebase not compromising on the risk associated. This SmartHealth cloud platform is a very comprehensive and complete visualization of our future vision of LHS [17]. Currently, it lacks actual implementation due to limited resources and time. Still we have done considerable effort in paving the way with clear detail of the scenario for making a hybrid distributed cloud in future that is intelligent enough to assist in Complex Event Processing (CEP) in service oriented community (SOC) for health. It is clearly understandable as we explored the challenges of building DSM [31] integrated with big data that forms the bases of our hybrid distributed cloud that actual simulation and implementation of SmartHealth cloud is a complex task. It is estimated to take huge span of time and expert skills giving us a novel machine learning framework to support our SmartHealth cloud platform.

VIII. RECOMMENDED ACTION PLAN

Our detailed data model would be in compliance with latest HL7 coding standards for coming up with self-learning medical diagnostics as in PADSS [8] that is for prediction of diseases integrated in cloud based healthcare platform. Further when setting up a benchmark for our system Adaptive Neurofuzzy Inference System (ANFIS) is seen as one of the successful approach for diagnosis of chronic kidney diseases [9]. But, it is focused over a single disease where Multilayer perception (MLP) neural network [5] is tested over eight diseases with 97% accuracy level achieved whereas we would structure the data using sequential and hierarchical clustering to become part of learning diagnostic model in Fig. 3. Our measure of accuracy would be based on HUM using biomarkers or a better approach than AUC. The simulation platform mostly used is MATLAB with NETLAB tool box. Other good platform for real scenario comprising of large scale patients' data would be Google Cloud Platform [8] used for PADSS or AnyLogic for hybrid cloud simulation as evaluated in [18] against NetLogo and other distributed simulation tools discussed.

REFERENCES

- [1] Masic I., & Novo A., "History of Medical Informatics in Bosnia and Herzegovina", *MedArh* 2006, [Medical Faculty, University of Sarajevo]
- [2] Shafqat S., Awan M. D., & Javaid Q., "Evaluating Cloud Computing for Futuristic Development", (2013), *International Journal of Computer Applications*, Volume 61-No.6, [SZABIST, Pakistan]
- [3] Gastaldi L., & Corso M., "Managing ICT to Solve the Exploration-Exploitation Paradox in Healthcare", (2015), [Italy]
- [4] Kusnandar T., & Surendro K., "Adoption model of hospital information system based on cloud computing: Case study on hospitals in Bandung city", (2013), *International Conference on ICT for Smart Society*, [STMIK Mardira Bandung, Indonesia]
- [5] Moein S., Moallem P., & Monadjemi A., "A novel fuzzy-neural based medical diagnosis system", (2009), [University of Isfahan, Isfahan, Iran]
- [6] Ramesh A., Kambhampati C., Monson J., & Drew P., "Artificial Intelligence in Medicine", (2004), [The Royal College of Surgeons of England]
- [7] Shafqat S., Kishwer S., Rasool R. U., Qadir J., Amjad T., and Ahmad H. F., "Big data analytics enhanced healthcare systems: a review," *The Journal of Supercomputing*, pp. 1–46, 2018.
- [8] Neto S., & Ferraz F., "Disease Surveillance Big Data Platform for Large Scale Event Processing", (2016), ANCS'16, March 17–18, ACM, [Recife Center for Advanced Studies and Systems, Brazil]
- [9] Norouzi J. et al., "Predicting Renal Failure Progression in Chronic Kidney Disease Using Integrated Intelligent Fuzzy Expert System", (2016), *Computational and Mathematical Methods in Medicine Volume 2016*, Hindawi Publication, [University of Technology and University of Medical Sciences, Iran]
- [10] Roth H. R., Lu L., Liu J., Yao J., Seff A., Cherry K., Kim L., & Summers R. M., "Improving Computer-aided Detection using Convolutional Neural Networks and Random View Aggregation", (2015), [National Institutes of Health Clinical Center, USA]
- [11] Li Jialiang, "Multicategory reclassification statistics for assessing improvements in diagnostic accuracy", (2013), [National University of Singapore]
- [12] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I., "Machine Learning and Data Mining Methods in Diabetes Research", (2017), *Computational and Structural Biotechnology Journal*
- [13] Kahramanli H. & Allahverdi N., "Design of a Hybrid System for the Diabetes and Heart Diseases", *Expert Systems with Applications* 35 (2008) 82–89, [Selcuk University, Konya, Turkey]
- [14] Nalluri et al., "Hybrid Disease Diagnosis Using Multiobjective Optimization with Evolutionary Parameter Optimization", *Journal of Healthcare Engineering Volume 2017*, [SASTRA University, Thanjavur, Tamil Nadu, India]
- [15] Grochowski M. & Jankowski N., "Comparison of Instances Selection Algorithms", *International Conference on Artificial Intelligence and Soft Computing 2004*, LNAI 3070, pp. 580–585, 2004. [Nicholaus Copernicus University, Poland]
- [16] Jankowski N., "Approximation and Classification in Medicine with InNet Neural Networks", *Workshop on Machine Learning in Medical Applications*, [Norbert Jankowski Department]
- [17] Kaggal et al., "Toward a Learning Healthcare System – Knowledge Delivery at the Point of Care Empowered by Big Data and NLP", (2016), *Innovations in Clinical Informatics*, [Division of Information Management and Analytics, Mayo Clinic, Rochester, MN, USA]
- [18] Shafqat, S., Abbasi, A., Amjad, T., & Ahmad, H. F. (2018). SmartHealth simulation representing a hybrid architecture over cloud integrated with IoT: a modular approach. In *Future of Information and Communications Conference (FICC)*.
- [19] Afzal et al., "CKM-CT: Comprehensive knowledge model creation for cancer treatment decision making", (2017), *Computers in Biology and Medicine*, [Kyung Hee University, South Korea and King Faisal University, KSA, Saudi Arabia]
- [20] Syahputra M. F., Felicia V., Rahmat R. F., and Budiarto R., "Scheduling Diet for Diabetes Mellitus Patients using Genetic Algorithm," *International Conference on Computing and Applied Informatics 2016*. Conf. Series: Journal of Physics: Conf. Series 801 (2017). [IOP Publishing]
- [21] Cynthia M. Boyd, Bruce Leff, Jennifer L. Wolff, Qilu Yu, Jing Zhou, Cynthia Rand, Carlos O. Weiss. "Informing Clinical Practice Guideline Development and Implementation: Prevalence of Co-existing Conditions Among Adults with Coronary Heart Disease." *Journal of the American Geriatrics Society*, 2011.
- [22] Heather A. Close, Li-Ching Lee, Christopher N. Kaufmann, Andrew W. Zimmerman. "Co-occurring Conditions and Change in Diagnosis in Autism Spectrum Disorder." *American Academy of Pediatrics*, 2012.
- [23] Konstadina Griva, Nandakumar Mooppil, Eric Khoo, Vanessa Yin Woan Lee, Augustine Wee Cheng Kang, Stanton P Newman. "Improving outcomes in patients with coexisting multimorbidity conditions—the development and evaluation of the combined diabetes and renal control trial (C-DIRECT): study protocol." *BMJ Open*, 2015.
- [24] Stephen K Gruschkus, Joseph M Darragh, Michael A Kolodziej, Roy A Beveridge, Michael Forsyth and Carolina Reyes. "Impact of Disease Progression On Healthcare Cost and Resource Utilization Among Follicular NHL Patients Treated within the US Oncology Network." *Blood Journal - Vol 114*, October 5, 2015.
- [25] Zhou Zhou, Paresh Chaudhari, Hongbo Yang, Anna P. Fang, Jing Zhao, Ernest H. Law, Eric Q. Wu, Ruixuan Jiang, Raafat Seifeldin. "Healthcare Resource Use, Costs, and Disease Progression Associated with Diabetic Nephropathy in Adults with Type 2 Diabetes: A Retrospective Observational Study." *Diabetes Therapy*, 2017: 555-571.

- [26] Kulkarni, S. M., & Babu, B. S., "Cloud-Based Patient Profile Analytics System for Monitoring Diabetes Mellitus", (2015), IJITR, 228-231
- [27] Zhao et al., "Learning from heterogeneous temporal data in electronic health records", (2017), Journal of Bio-medical Informatics 65 (2017) 105–119, [Stockholm University, Sweden]
- [28] Miotto et al., "Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records", (2016), Scientific Reports, [Icahn School of Medicine at Mount Sinai, New York, NY, USA]
- [29] Florio et al., "Towards a Smarter organization for a Self-Servicing Society", (2016), ACM DSAI 2016, [Morocco & Belgium]
- [30] Du H., Taveter K., and Huhns M. N., "Simulating a Societal Information System for Healthcare", (2012), Proceedings of the Federated Conference on Computer Science and Information Systems pp. 1239–1246, [University of South California, USA]
- [31] Marshall et al., "Transforming Healthcare Delivery: Integrating Dynamic Simulation Modelling and Big Data in Health Economics and Outcomes Research", (2015), Pharmacoeconomics Springer, [Clinical Engineering Learning Lab, Mayo Clinic Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery, Rochester, MN, USA]