January-February 2020
ISSN: 0193-4120 Page No. 8684 - 8695

Machine Learning Prediction Models Through the Performance Evaluation of Diabetic Classification: A Survey

G Vijaya Kumar, G Pulla Reddy Engineering College, gvjykumar@gmail.com
Patil Manogna K, G Pulla Reddy Engineering College, patilmanognareddy95@gmail.com

Article Info
Volume 82
Page Number: 8684 - 8695
Publication Issue:
January-February 2020

Article History
Article Received: *5 April 2019*
Revised: *18 Jun 2019*
Accepted: *24 October 2019*
Publication: *08 February 2020*

Abstract:
In the field of health research, the health system utilizing advanced computer technology is the most essential research requirement. Researchers in the field of informatics and healthcare are constantly working together to develop more advanced technologies for such systems. Recent World Health Organization studies show an increase in the number of diabetic patients and their mortality. Diabetes is one of the major diseases and has far-reaching complications. To make the medical information a great deal, it is important to collect, store, study and evaluate the health of these patients through continuous monitoring and technological innovation. The alarming rise in the number of diagnostic patients has becomes a major concern. Through innovation, it is important to set up a system for storing and monitoring diabetes information, and more serious dangers can be discovered. Early detection and analysis remain a challenge for researchers. This survey paper provides an ongoing study on diabetes detection and suggested structures.

Keywords: *classification, diabetes, machine learning, performance, predictive analytics.*

I. INTRODUCTION

Diabetes is a collection of metabolic diseases. These diseases are categorised by hyperglycemia, which is the result of any of the defects in secretion, the action of insulin or both. Diabetes hyperglycemia is connected with the continuing damage, or multi-organ failures such as heart, blood vessels, nerves, eyes and kidneys. The involvement of various pathogenic processes develops diabetes [1]. These processes range from the destruction of the pancreas's β-cells, which helps in autoimmunity resistance irregularities in the insulin action. Fundamental reason for these irregularities is the absorption of carbohydrate, fat and protein. This creates a misfunction of insulin action over the target issues. Lack of insulin action is the result of insufficient discharge of insulin and tissue responses being weak. Damage in insulin discharge coexists very often in the same patient. However, it is not clear which abnormality is the fundamental reason for hyperglycemia.

The process of data classification for pattern recognition is valuable over a long period of time. People have a great deal of environmental experience. Action would be performed on the perceived situations from the environment [2]. Moreover, big data is becoming a part of multidisciplinary machine learning, database and statistical efforts. Nowadays diagnostic tests in the medical sciences are a serious task. It is important to understand the exact diagnosis of patients through clinical examination and evaluation. Decision support computer systems play an important role in effective diagnostics and cost-effective management. In the field of health, extensive data are available on clinical evaluation, patient reporting, prevention, follow-up and medication. Appropriate measures can be difficult. Incorrect processing of data affects the quality of the data organization. Adequate ways of efficient and effective data collection and processing

January-February 2020
ISSN: 0193-4120 Page No. 8684 - 8695

are needed for the growth of data compilation [3].

Among the many machine learning applications one can be used to create such a classification, which can be broken down by properties. A dataset can be classified into two or more. These classifications are utilized to analyse medical data to diagnose the disease. Primarily, ML algorithms were developed and used to test medical data sets. Nowadays ML recommends various tools for efficient data analysis. Particularly in recent years, the digital revolution has provided a cheap and accessible way to collect and store data. Data collection and testing facilities are installed in the recent and modern hospitals to collect and share the information in big data systems. Subsequently, ML methods have proved to be very effective for analysing medical data and a considerable work is being done on diagnostic problems. A correct diagnostic information is displayed in medical records or reports of advanced hospitals or their specialized data units.

To implement the algorithm, the appropriate diagnostic patient record is stored as input to the computer. Previously obtained case results are automatically obtained. Doctors benefit from this classification of derivatives to identify a new patient quickly and improve accuracy. To train the professionals as well as the students, these classifiers can be used [4] to diagnose a problem. Recently, technologies have started incorporating the ML techniques to provide the assistance for autonomous vehicles, speech detection, effective web-based research and a better understanding of the humankind. Nowadays, machine learning applications are being used by the users ubiquitously and are being used multiple times a day without knowing it. Researchers are considering it as a wonderful way to reach the human users. Machine learning methods usually find an electronic health record that has large models and multiple data sets. Sample identification is an MLT theme that supports diagnosis and decision making as well as treatment planning. Machine learning algorithms can process huge amounts of data, by combining the data from different sources, and integrate basic information into research.

II. Related Work

The availability of health care data is widely believed by many researchers, doctors and patients to open up access to consent. While this may be true, in the foreseeable future these data, by their very nature, may become a curse, posing a challenge to human society. In recent years, many publications have documented the challenges of collecting, storing, accessing, analyzing, and storing healthcare data.

In the research study proposed by Lai Hang et al. [5](2019), two models have been used utilizing the machine learning techniques, such as gradient augmentation machine, and logistic regression for identifying patients at high risk of developing diabetes. For their samples, they have used both the classical statistical model and advanced machine learning techniques. They have solved the data mismatch using the correction threshold method and the class weighting technique. As claimed by the authors, the capability to identify diabetic patients using their models was high with reasonable sensitivity. These models have been developed and validated for Canadian populations that reflect diabetes risk profiles in Canadian patients. The ease of these models is that they can be created in an online computer program to help doctors assess the risk of developing diabetes in Canadian patients.

The stratification of people at risk for diabetes can target intervention programs for high-risk individuals, by avoiding the efforts and prevention and treatment cost for low-risk individuals. Perveen [6](2019) et.al. have proposed a study to investigate the potential role of machine learning technique hidden Markov Model (HMM) in the performance validation of the well-followed prognostic model, the Framingham Diabetes Risk Assessment Model (FDRSM). Can HMM effectively predict the

January-February 2020
ISSN: 0193-4120 Page No. 8684 - 8695

symptoms of diabetes over the next 8 years in humans? The authors have an opinion that, no study exists which has attempted to use HMM to test FDRSM. Authors have used electronic health record (EHR) data for 172,168 patients at risk of developing 8-year diabetes in a person using HMM. In one study, the area under the receiver operating characteristic (AROC) was reported in the authors' study sample of 911 people, compared to 86.9% of all risk factors and 86.9% of the available AROC. Previous validation of 78.6% and 85% FDRSM was observed in the same dataset of Canadian population and Framingham study. These results indicate that our proposed HMM has a higher discriminatory capacity than the validation study using FDRSM in the Canadian and Framingham populations. We conclude that HMM has the potential to identify patients at risk for diabetes over the next 8 years.

The aim of the work by Zhu et.al. [7](2019) is to develop an effective diabetes prediction model. After a careful review of other published works, they proposed a new model that incorporates the use of PCA to reduce the dimensionality, k-averages in clustering, and logistic regression-based classification. Other researchers' tools are designed to improve results First, authors have applied the PCA method to our dataset. Although PCA is a popular technique, its ability to improve the K-Means clustering and logistic regression classification model Not enough attention has been paid. A better logistic regression model has been shown to predict diabetes by integrating PCA and K-agents. The novelty of the study is that it is able to obtain a better set of k-means than other researchers in similar studies. Compared to the results obtained in our study and other algorithms used in our research, the logistic regression model was superior in predicting the onset of diabetes. Another advantage of their model is that it is able to successfully model a new data set.

Diabetes mellitus is considered to be one of the deadliest and most chronic diseases that cause high blood sugar. Most complications occur when diabetes is not treated and are not diagnosed. Patient attendance diagnosis is a thorough identification process and medical consultation. But the growth machine learning methods provide solutions to this critical problem. Aim of the study conducted by Sisodia et.al. [8](2018) was to develop a model with maximum accuracy could predict the likelihood of diabetic patients. Hence, the three machines used in this experiment are Decision Tree, SVM, and Naive Bayes Learning Classification Algorithms for Diabetes Detection. Pima Indian Diabetes Database (PIDD) Training Database obtained from UCI machine. Performance of the three algorithmic operations were evaluated over accuracy, precision, recall and F-measure. For the cases of correctly and incorrectly classified instances accuracy is measured. The results show Naïve Bayes outperforms others by 76.30% over other algorithms. These results have been properly and systematically verified using receiver operating characteristics (ROC) curves.

K. Deepthi Krishnan et.al. [9] (2018) conducted a disease assessment survey using machine learning to convert incomplete data into complete data and to quickly predict disease using comprehensive data. . With the use of various work-based methods, such as CND-based One-Dimensional Disease Prediction (MDRP) and MDRP (Multimodal Disease Prediction), there is no effective method for predicting disease. In the end, his work suggested that high performance requires a lot of machine learning algorithms.

Razak et al (2017) [10] carried out their work on the applicability of deep learning in medical image processing. Beneficiaries of the healthcare industry expect a lot to be a premium. The authors estimate that, despite most of the spent state budget, the industry does not meet public expectations. Unlike the medical level analysis of medical images, the authors recommend continuing education based on the skill level of the experts or the heavy workload.

January-February 2020
ISSN: 0193-4120 Page No. 8684 - 8695

Based on the real-world success, deep learning plays a crucial role in addressing the subjectivity and complexity of medical images. This success would further produce the high-quality results to improves the reliability of analytical tools in data and physician-patient communication. In his review article, William W. Steed (2018) [68] expressed similar views on the clinical implications and challenges of AI, Deep neural network, convolutional neural network, repetitive neural network, deep conventional deep learning machine, deep Boltzmann machine, deep belief network, deep auto encoder.

B. Seligman et al (2018) [12] conducted a study on the use of machine learning methods in the social determinants of health during health and retirement. These authors have attempted to address the socio-economic factors affecting public health, particularly in retirement age. Although researchers have made great efforts to use machine learning approaches to assess health problems, little research has been done on the relationship between public and health. The method used by the authors was used to estimate age, gender and income, wealth and regression based on educational evidence, systolic BP, BMI, waist circumference and length. telomeres. Linear regression, random forests, fine regression, and neural networks were used to compare prediction, fit, and interpretation. Although all machine learning methods have poor unintended sampling prediction, neural networks go beyond this to provide good data fit. Nine variables were selected as common factors, such as current smoking, dental visits, and transmission probability.

In an innovative study, Samant, Piyush et al (2018) [13] proposed a new method for the detection of diabetes using human iris images in combination with machine learning methods. Their aim is to diagnose type 2 diabetes by evaluating the diagnostic validity of old complementary techniques and alternative medicine, namely iridology. The dataset contains medical data on 180 diabetes subjects and 150 non-diabetic patients, for a total of 338 subjects. The medical information includes infrared images of both eyes to obtain an iris image. According to the author and the iridology chart, the area of the iris corresponds to the pancreas. Signs extraction, feature selection, and iris classification were then performed for further analysis. In the analysis, the authors noted that the statistical, structural, and DWT characteristics have the greatest potential to classify diabetes and non-diabetic images by Iris.

Manogaran.G et.al. [14](2017) conducted a study on existing big data structures and machine learning in healthcare. They claim that the estimated size of the data is 44 times larger than the 2009 data. The problems they have observed in various research literature are data mining, data processing, data storage, acceleration, data visualization, data interpretation, problem determination, problems with strangers, data quality, confidentiality and data security. Their work focuses on the origin and types of data: health data - EHR, medical image data, genetic data; Text, graphic data for social networking data; Structured data sensor; And data recording files generated by the machine. Lambda architecture is implemented in data processing for data analysis Twitter with three layers: the batch layer, the service layer and the speed layer. NIST has developed a Big Data reference architecture for data scientists, software developers, data architects, engineers and senior decision makers to develop solutions that focus on Big Data functions.

Lena Zhou et.al. [15](2017) presents in writing the key opportunities and challenges facing industry and researchers. They mention the classification of ML algorithms and big data platforms, which classify them as parallel and non-parallel targets, and the methods they use to optimize non-parallel targets; Map Reduce, split graph, multi-thread, MPI / Open MP and GPU systems for parallel purposes. Open search problems are also addressed in ML big data, such as cleaning and compression of big data, large-scale distributed functionality, real-time online

January-February 2020
ISSN: 0193-4120 Page No. 8684 - 8695

learning for data streaming, data learning. unreliable or conflicting, and large-scale collaborative support for multi-user decisions. Intelligent user interfaces for data analysis and interactive ML take first place through solutions. All these opportunities and challenges are the same for healthcare health care data, which adds to the privacy and protection of the patient's personal and health care.

M. Chen et. al. [3](2017) conducted experimental work in the field of medical data analytics for big data using machine learning techniques, convolutional neural networks (CNN). Their work focuses on analyzing structured and unstructured data from hospitals to develop a multimodal CNN-based disease prediction algorithm. They claim that none of the current methods work their way through both types of big data medical analysis data. In addition, their algorithm can be seen to be faster, converging to a convergence rate of 94.8% compared to CNN-based unimodal disease prediction algorithm.

The need to develop any algorithm or refine existing algorithms is just data. However, there are cases where the evaluation of pre-processed data is appropriate for the analysis process. M. Zhu et.al. (2018) [16] worked on the resolution of class mismatch data by developing a random forest algorithm. The authors address the problem of bias in current methods by classifying models that neglect minority groups. In conclusion, the authors expressed the importance and urgency of addressing the classification of class mismatch data. An algorithm is proposed to process this data, which returns input-class imbalance data to a random forest model, making the data more sensitive to majority and minority classes. In the next step of creating a class weight vote (CWSV), the algorithm works in two steps: calculating different weights for one class and adding the votes. The next step is to rank these amplified voices using the built-in probability (AP). To obtain the class weights of the classifier, a measure of empirical error is made. The authors emphasize that their algorithm gives higher ranking performance than other circuits.

Significant changes in people's lifestyles have affected their health, especially as the number of diabetes diagnoses has increased in the last decade.

F. Mercaldo et. al. [17](2017) and Razih Asgharnezad et. al. [18](2017) have attempted to develop a system for diagnosing diabetes in humans using a machine learning algorithm that meets global health criteria. Company. The research question that these authors were working on was: “Is it possible to distinguish between diabetic patients and not be influenced by a set of features selected by the World Health Organization as a vector for the symptoms?” To answer this question, the authors considered an eight-character vector from F1 to F8 Data Set of Pima Indian women. Authors have worked on the algorithm on the data of the patients suffered from diabetes for five years. All patients in the data set were at least 21 years old and 768 different cases. The authors have learned many advanced classifications with eight characteristics. The assessment consists of three different stages: detailed statistical comparison of diabetes versus affected population; test hypothesis to see if the diabetic carrier has a different distribution between affected and affected populations; and categorical analysis to assess whether the eight symptoms have the potential to differentiate between diabetic patients and affected patients. A study was conducted by Rigla et.al. [19](2017) to propose their results on AI methods applicable to the diagnosis of diabetes.

A. Oliveira et. al. [20](2017) presented their research on various machine learning algorithms, to develop a model to diagnose undiagnosed diabetes in Brazil affecting adults with severity and general health. After extracting from a variable subset of 27 different candidate candidates, that regular habits are classified into a four-step model: re-validated the parameters with cross-validation for dozens of times. To evaluate each subset of variables, automatic

January-February 2020
ISSN: 0193-4120 Page No. 8684 - 8695

selection of variables was preformed using the direct option with ten times mutual validation (repeated three times). Model parameter error is estimated with ten times the cross-confidence by repeating iterations, and normalization tests in an independent dataset. These models were created using machine learning algorithms such as RF, LR, ANN, naïve Bayes, and K-neighbour. Among them random forest, among which the authors observed that the best models were developed using artificial neural networks and logistic regression.

In a research paper, Jianfeng Zhang et. al. [21](2017) proposed a diagnostic method based on machine learning methods, support vector machine and language representation in diabetes. During this process, they collected language images of 296 diabetes subjects and 531 non-diabetic subjects using the TDA-1 digital language device. Using two methods, split fusion and chromatic threshold, they separate the body from the tongue and tongue overlay. Their diabetes model was created by combining input variables, texture characteristics, and language image color. The system used SVM during the training. After optimizing the combination of input variables and SVM kernel parameters, the effect of model concatenation was analyzed. Diabetes prediction accuracy increased from 77.83% to 78.77%. There was no decrease in AUC after APC administration after reduction of language symptoms. When learning to select genetic algorithm SVM parameters, the cross-validation accuracy rate increased from 72% to 83.06%. The authors claim that the experimental results of their model gave better prediction accuracy than sophisticated algorithms.

III. Background

Diabetes is a condition called glucose and blood sugar. Blood glucose is the main source of energy which comes from the foods the people consume as part of their diet. A hormone called glucose, which is made by the pancreas, is used by the cells to produce energy [22]. Sometimes your body is not in a position to make sufficient insulin or to use insulin too well. After that, glucose which is in the blood does not reach the cells. Meanwhile, too much blood glucose can lead to health problems. While diabetes cannot be cured, people can take appropriate preventive measures to manage diabetes and stay healthy. Sometimes people with diabetes call it "a touch of sugar" or "borderline diabetes". These terms indicate that no one really has a diabetes or a less serious case, but each case of diabetes is equally serious.

A. Most Common Types of Diabetes Disease Symptoms and Causes

Types of diabetes that medical community has classified are: Type 1 diabetes is an autoimmune disease. The immune system attacks and destroys the pancreatic cells that make insulin. It is knowns exactly what caused the attack. About 10 percent of people with diabetes have this type. Type 2 diabetes occurs when your body becomes insulin resistant [23] and has high blood sugar. Prediabetes occurs when your blood sugar is higher than normal, but it may not be enough to diagnose type 2 diabetes. Gestational diabetes is caused during pregnancy who have high blood glucose levels which is responsible for the insulin-blocking hormones produced by the placenta.

B. Diabetes Disease Symptoms and Causes

Symptoms of diabetes slightly vary in males and females though the study about other genders is underway. General symptoms of diabetes among all the genders are: increase in hunger and/or hunger, thirst, fatigue, urination, added with weight loss, and blurred vision. However, diabetes symptoms among specific genders are: poor muscle strength and erectile dysfunction among males; and urinary tract infections, and dry and itchy skin among females.

C. Machine Learning

Artificial intelligence makes the computer think. AI makes the computer smarter. Machine learning is an area of study for AI. Different researchers think that

January-February 2020
ISSN: 0193-4120 Page No. 8684 - 8695

it is impossible to develop intelligence without learning. Figure 1 shows many types of machine learning techniques.

Types of supervised, unattended, semi-supervised, reinforcing, evolutionary, and deep learning methods. These methods are used to classify a dataset [24]:

a) Supervised Learning: Provides a set of training with appropriate goals, and based on this training set, the algorithms respond correctly to all possible entries. Another name for supervised learning is learning from examples. Types of classified training regression and regression. Classification: it gives a rating yes or no, eg "Is this a cancer tumor?", "Does this cookie meet our quality standards?" Regression: the answers could be "how many" and "how much".

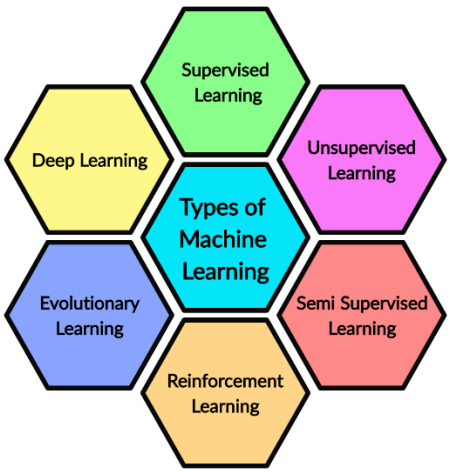

Figure 1. Types of machine learning techniques

b) Unsupervised Learning: no answers or proper targets given. The purpose of Unsupervised learning technique is to detect commonalities among input data and to classify the data as Unsupervised learning technology based on these similarities. This is also called density estimation. This technique involves clustering, a process to create the groups based on similarity.

c) Semi-supervised learning: Semi-supervised learning technique is one of the classes of supervised learning methods. This practice also uses untagged data for training purposes (usually the minimum volume of tagged data with the smallest amount of tagged data). Between unsupervised learning (data untagged) and unsupervised learning (data tagged), there is partially supervised learning.

d) Reinforcement Learning: Behavioural psychology facilitates this practice. The algorithm is announced if the answer is incorrect, but it does not provide the way for how to fix it. Until the right answer is found, various options should be explored and tested. This is also called learning with the critic. This technique does not recommend improvements. Reinforcement learning differs from supervised learning in which no specific input and output sets or intermediate measures are directly evaluated. In addition, it focuses on online performance.

e) Evolutionary learning: This biological evolution can be considered as a learning process because biological organisms have the capability to improve in their survival rate and have springs. Using the idea of fitness, this model can be used on a computer to test the accuracy of the solution.

f) Deep Learning: Deep learning technique is one of the machine learning techniques based on a set of algorithms. In the data, these learning algorithms model a high level of abstraction. It uses a deep graph with multiple layers of processing, making it a very linear and linear transformation.

D. Machine Learning Role in Predicting diabetes

The idea of using automated tools to predict a non-communicable disease is not a recent one. However, with the evolution of implementing AI techniques [25], especially machine learning techniques predictive analytics have proved its strength to analyse the existing data to predict the future.

a. Traditional Methods to Identify diabetes

In the process of diagnosing diabetes mellitus type 2 various tests are performed: i) Fasting Plasma Glucose (FPG) Test: FPG test [26] would be performed by asking the patient to attend the process

January-February 2020
ISSN: 0193-4120 Page No. 8684 - 8695

while fasting, especially in the morning. ii) Oral Glucose Tolerance Test (OGTT): Unlike FPG, OGTT [27] is conducted in two phases: Phase 1 requires the person to observe fasting for at least 8 hours before the blood sample is offered to the pathology. Immediately after the blood sample collection, water dissolved with 75 grams of glucose shall be consumed by the person. In the Phase 2, another blood sample would be collected from the person after 2 hours of consumption of the liquid. In both the instances the Plasma glucose levels are tested.

b. Disadvantages of Traditional Methods

Though the traditional methods are being used even today, there have been many reports for erroneous diagnoses. The possible errors are diagnosing for diabetes for those who are not in real-time, and vice-versa. The causes for such errors range from human negligence at labelling the samples to chemical preparations for uniform calibration.

c. Scope of Machine Learning Techniques

The most valuable and helpful technique for pattern recognition is the data classification. The process of classification has been in use over a long period of time. Researchers have utilized the classification technique in many areas, especially image processing. In the recent times, medical industry has started using various automated diagnosis approaches for a better outcome. The process has been widely accepted for the reason that the diagnosis would be accurate and further treatment becomes more appropriate. When the above techniques are integrated with decision support system, it can play an important role for an efficient diagnosis and the system management become cost-effective. For the healthcare industry, clinical evaluation, patient reporting, disease prevention and treatment [28] follow-up would become very easy with adoption of such decision support systems. All this is possible because of the availability of extensive health data.

To prepare a classification process, systems can use one of the many machine learning techniques. The process works with the initial task of dividing the dataset into two or more classes. To analyse the medical and diagnose the disease these classification processes would help in a better way. Fundamentally, development of ML algorithms was to use them to test the medical data sets. Moreover, ML algorithms endorse various tools for efficient analysis of the health data. ML algorithms have proved to be very efficient for the effective analysis of medical data. In this direction, there has been a tremendous work which was carried out by the research community. When the diagnosis information is found correct it would be recorded in the medical databases or the reports.

For the implementation of the ML algorithms, the accurate records of the diagnostic patient would be stored and would be supplied as input to the computing system. The results from this automatic system were found to be more accurate as the computing systems were tested for more accuracy. Due to the accuracy in information doctors would benefit, especially when the systems use the classification techniques to identify a new patient with in very short time. Such classifiers can be conveniently used in the training process for professionals and students to diagnose a health problem.

E. Challenges in Diabetes Management

Current challenges in the diabetes management [29] consists of: i) to optimize presently available treatments to guarantee suitable glycemic, blood pressure, and lipid control and to reduce complications; ii) to educate the patients on diabetes self-management; iii) to improve patient's lifestyle and pharmacological intervention; iv) to reduce barriers to early insulin delivery; and v) to improve the delivery of health care to people with chronic illnesses.

January-February 2020
ISSN: 0193-4120 Page No. 8684 - 8695

F. Machine Learning techniques for Diabetes Diseases

A tremendous research has taken place for the appropriate of machine learning techniques (MLT). It is a continuous process to conduct research and propose new technique or to use the combination of MLT. The MLT which have gained popularity to assist for diabetes prediction are: Naïve Bayes, J48, CART, Adaboost, Logiboost, Grading, Support Vector Machine, Genetic Algorithms with Fuzzy Logic [30].

G. Challenges using Machine Learning for Diabetes Diseases

After careful research of the literature and tools, it is recognized that while data can be classified as structured, semi-structured and structured, other types of data need to be converted to structured data if tools exist. Otherwise, the tools will not work to achieve efficiency in the expected data results. In addition, predictive analysis requires a flawless concept, which must match proof of concept (POC) and proof of work (POW).

These are the challenges of big data analytics research in health care for chronic disease assessment.

i. Serious work is underway to develop big data analytics solutions. An important feature that such solutions should have is scalability. While solutions can reach a certain degree of scalability, further improvements are needed to reach the degree to which solutions can solve problems of any size.

ii. An important aspect of big data analysis is massive data processing and finding a solution in a very short time. However, this does not solve the complexity of the time, as each solution does not solve parallelism. When solutions are installed to achieve parallelism, they automatically become complex to make decisions faster.

iii. Unlike other data, big data on health care comes in many forms and can be discrete. Types include graphs / graphics such as text, image, video, scanned image, ECG and EEG.

iv. Data collection and outpatient consultation differ in the context of clinical patients. Examples of data collection are clinical diagnostic tools such as body scanners, pathology laboratories, and remotely connected devices such as smart sensors and handheld devices. The collected data must be stored prior to communication. Therefore, data must be stored for a certain period, sometimes longer

v. Stored healthcare requires the transmission of large amounts of data using various communication skills, for example in a hospital or clinical setting. Different methods of communication include Wi-Fi, Bluetooth, Internet, etc. Installations are not expensive, but protocols that accept data from sources must be properly integrated to communicate with the destination. For better communication, protocols and communication standards are being developed for the use of these body sensors. Therefore, there is a small amount of trouble-free communication that is not enough.

vi. The data received after communication with the source must be processed at the destination, that is, with the therapist, for later action. In most cases, the data received may be noisy or incomplete. Inconsistent data or knowledge processing recorded in noisy data can result in typographical errors or results in production errors. In addition, some important data records may have no values, values that may lead to incorrect results. Another example of incomplete data is lack. Therefore, appropriate tools are needed to process health data.

vii. Due to the size of the data, ie the mass, they must be compressed before transmission. However, technology adopted to maintain data integrity to meet treatment needs should not be compromised. Therefore, the system should accept lossless compression, the source of which is rare for the production of expensive equipment.

January-February 2020
ISSN: 0193-4120 Page No. 8684 - 8695

IV. Conclusions and Suggestions

Predicting diabetes is an essential activity, nowadays, as early detection would lead to a better treatment either with medicine or change in the lifestyle. This work describes machine learning classification methods to predict diabetes. The survey was conducted on various classification methods to lead identifying a better method to improve the results in terms of accuracy, precision and sensitivity. Challenges involved in the diabetes diagnoses have also been explored. Moreover, this work has explored for the challenges when machine learning techniques are implemented for diabetes prediction. Study of the performance can be extended with more parameters and effective diagnosis of diabetes. The work can be improved by expanding the study in the area of automation of diabetes Analysis.

V. References

[1] Puchulu, Félix. "Definition, Classification and Diagnosis of Diabetes Mellitus." Cutaneous Manifestations of Diabetes, 2018, 1–1. https://doi.org/10.5005/jp/books/13050_2.

[2] Cheaito, Mohammad Ali, and Marwan Cheaito. "A Novel Framework for Electronic Global Health Record Access." Health Informatics - An International Journal, vol.4, no. 1/2 (2015): 13-33. doi:10.5121/hiij.2015.4201..

[3] Chen, Min, Yixue Hao, Kai Hwang, Lu Wang, and Lin Wang. "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities." IEEE Access, vol.5 (2017): 8869-879. doi:10.1109/access.2017.2694446.

[4] Sarkar, Bikash Kanti. "Big Data for Secure Healthcare System: A Conceptual Design." Complex & Intelligent Systems, vol.3, no. 2 (2017): 133-51. doi:10.1007/s40747-017-0040-1.

[5] Lai, Hang, Huaxiong Huang, Karim Keshavjee, Aziz Guergachi, and Xin Gao, "Predictive Models for Diabetes Mellitus Using Machine Learning Techniques," BMC Endocrine Disorders, Vol.19, No.1, 2019. https://doi.org/10.1186/s12902-019-0436-6.

[6] Perveen, Sajida, Muhammad Shahbaz, Karim Keshavjee, and Aziz Guergachi, "Prognostic Modeling and Prevention of Diabetes Using Machine Learning Technique," Scientific Reports Vol.9, No.1, 2019. https://doi.org/10.1038/s41598-019-49563-6.

[7] Zhu, Changsheng, Christian Uwa Idemudia, and Wenfang Feng. "Improved Logistic Regression Model for Diabetes Prediction by Integrating PCA and K-Means Techniques." Informatics in Medicine Unlocked, 2019, 100179. https://doi.org/10.1016/j.imu.2019.100179.

[8] Sisodia, Deepti, and Dilip Singh Sisodia, "Prediction of Diabetes Using Classification Algorithms," Procedia Computer Science No.132, 2018, pp:1578–85. https://doi.org/10.1016/j.procs.2018.05.122.

[9] K.DeepthiKrishnan. " A Survey on Disease Prediction by Machine Learning over Big Data from Healthcare Communities." IOSR Journal of Engineering (IOSRJEN), vol. 08, no. 10, 2018, pp. 53-59.

[10] Razzak, Muhammad Imran, Saeeda Naz, and Ahmad Zaib. "Deep Learning for Medical Image Processing: Overview, Challenges and the Future." Lecture Notes in Computational Vision and Biomechanics Classification in BioApps, 2017, 323-50. doi:10.1007/978-3-319-65981-7_12.

[11] Stead, William W. "Clinical Implications and Challenges of Artificial Intelligence and Deep Learning." Jama, vol. 320, no. 11 (2018): 1107. doi:10.1001/jama.2018.11029.

[12] B. Seligman, S. Tuljapurkar and D. Rehkopf, "Machine learning approaches to the social determinants of health in the health and retirement study", SSM - Population Health, vol. 4, pp. 95-99, 2018. Available: 10.1016/j.ssmph.2017.11.008

[13] Samant, Piyush, and Ravinder Agarwal. "Machine Learning Techniques for Medical Diagnosis of Diabetes Using Iris Images." Computer Methods and Programs in

January-February 2020
ISSN: 0193-4120 Page No. 8684 - 8695

Biomedicine, vol. 157 (2018): 121-28. doi:10.1016/j.cmpb.2018.01.004.

[14]Manogaran, G. and Lopez, D. (2017) 'A survey of big data architectures and machine learning algorithms in healthcare', Int. J. Biomedical Engineering and Technology, Vol. 25, Nos. 2/3/4, pp.182–211.. doi:10.1504/ijbet.2017. 10008616.

[15]Zhou, Lina, Shimei Pan, Jianwu Wang, and Athanasios V. Vasilakos. "Machine Learning on Big Data: Opportunities and Challenges." Neurocomputing, vol. 237 (2017): 350-61. doi:10.1016/j.neucom.2017.01.026.

[16]Zhu, Min, Jing Xia, Xiaoqing Jin, Molei Yan, Guolong Cai, Jing Yan, and Gangmin Ning. "Class Weights Random Forest Algorithm for Processing Class Imbalanced Medical Data." IEEE Access, vol.6 (2018): 4641-652. doi:10.1109/access.2018.2789428.

[17]F. Mercaldo, V. Nardone and A. Santone, "Diabetes Mellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques", Procedia Computer Science, vol. 112, pp. 2519-2528, 2017. Available: 10.1016/j.procs.2017.08.193.

[18]Asgarnezhad R, Shekofteh M and Boroujeni FZ: Improving Diagnosis of Diabetes Mellitus using Combination of Preprocessing Techniques. J Theor Appl Inf Technol 2017; vol. 95, no.13, pp: 2889-2895, 2017.

[19]M. Rigla, G. García-Sáez, B. Pons and M. Hernando, "Artificial Intelligence Methodologies and Their Application to Diabetes", Journal of Diabetes Science and Technology, vol. 12, no. 2, pp. 303-310, 2017. Available: 10.1177/1932296817710475

[20]A. Olivera et al., "Comparison of machine-learning algorithms to build a predictive model for detecting undiagnosed diabetes - ELSA-Brasil: accuracy study", Sao Paulo Medical Journal, vol. 135, no. 3, pp. 234-246, 2017. Available: 10.1590/1516-3180. 2016. 0309010217.

[21]Zhang, Jianfeng, Jiatuo Xu, Xiaojuan Hu, Qingguang Chen, Liping Tu, Jingbin Huang, and Ji Cui. "Diagnostic Method of Diabetes Based on Support Vector Machine and Tongue Images." BioMed Research International2017 (2017): 1-9. doi:10.1155/2017/7961494.

[22]Zou, Quan, Kaiyang Qu, Yamei Luo, Dehui Yin, Ying Ju, and Hua Tang, "Predicting Diabetes Mellitus With Machine Learning Techniques," Frontiers in Genetics, 9, 2018,. https://doi.org/10.3389/fgene.2018.00515.

[23]Fatima, Meherwar, and Maruf Pasha. "Survey of Machine Learning Algorithms for Disease Diagnostic," Journal of Intelligent Learning Systems and Applications, vol.9, no.1, 2017, pp: 1–16. https://doi.org/10.4236/jilsa.2017.91001.

[24]"Follow-up Report on the Diagnosis of Diabetes Mellitus," Clinical Diabetes vol.22, no.2, January 2004, pp:71–79. https://doi.org/10.2337/diaclin.22.2.71.

[25]Blonde, Lawrence, "Current Challenges in Diabetes Management," Clinical Cornerstone, 7, 2005. https://doi.org/10.1016/s1098-3597(05)80084-5.

[26]Hassler, Andreas Philipp, Ernestina Menasalvas, Francisco José García-García, Leocadio Rodríguez-Mañas, and Andreas Holzinger. "Importance of Medical Data Preprocessing in Predictive Modeling and Risk Factor Discovery for the Frailty Syndrome." BMC Medical Informatics and Decision Making, vol.19, no. 1 (2019). doi:10.1186/s12911-019-0747-6.

[27]S. Ramírez-Gallego, S. García, J. M. Benítez and F. Herrera, "Multivariate Discretization Based on Evolutionary Cut Points Selection for Classification," in IEEE Transactions on Cybernetics, vol. 46, no. 3, pp. 595-608, March 2016. doi: 10.1109/TCYB.2015.2410143

[28]Timo M. Deist, Frank J. W. M. Dankers, Gilmer Valdes, Robin Wijsman, I- Chow Hsu, Cary Oberije, Tim Lustberg, Johan Soest, Frank Hoebers, Arthur Jochems, Issam El Naqa, Leonard Wee, Olivier Morin, David R. Raleigh, Wouter Bots, Johannes H. Kaanders, José Belderbos, Margriet Kwint, Timothy Solberg,

January-February 2020
ISSN: 0193-4120 Page No. 8684 - 8695

René Monshouwer, Johan Bussink, Andre Dekker and Philippe Lambin, Erratum: "Machine learning algorithms for outcome prediction in (chemo)radiotherapy: An empirical comparison of classifiers", Medical Physics, vol.46, no.2, pp:1080-1087, 2019.

[29]F. Mercaldo, V. Nardone and A. Santone, "Diabetes Mellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques", Procedia Computer Science, vol. 112, pp. 2519-2528, 2017. Available: 10.1016/j.procs.2017.08.193.

[30]M. Chen, Y. Hao, K. Hwang, L. Wang and L. Wang, "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities," in IEEE Access, vol. 5, pp. 8869-8879, 2017. doi: 10.1109/ACCESS.2017.2694446.