A SURVEY ON DIABETES PREDICTION MODELS USING MACHINE LEARNING OVER BIG DATA

SUHAS K C , Dr. ANAND BABU J

RESEARCH SCHOLAR COMPUTER SCIENCE AND ENGINEERING MALNAD COLLEGE OF ENGINEERING HASSAN, KARNATAKA

PROFESSOR ,INFORAMTION SCIENCE AND ENGINEERING MALNAD COLLEGE OF ENGINEERING HASSAN,KARNATAKA

suhas2385@gmail.com , anand.tiptur@gmail.com

Article History: **Received:** 19.05.2023 **Revised:** 20.06.2023 **Accepted:** 10.07.2023

ABSTRACT: Healthcare domain is one of the most prominent research field with rapid technological advancement and increasing data day by day. In order to deal with large volumes of healthcare data Big Data Analytics which is needed which is an emerging approach in Healthcare domain. The size of databases increases rapidly every day but still not well explored to discover hidden knowledge. Diabetes, an incurable disease which occurs because of high blood sugar levels over a prolonged time period, requires early prediction to significantly reduce its severity. In the medical field, it is essential to predict diseases early to prevent them. Now-a-days Machine Learning (ML) community has been working on diabetes prediction and many researches have been done since decades for its prediction. Advanced machine learning techniques can be used to develop predictive models for medical data, such models could be very useful for physician to make the decision. In this work a survey on prediction models of Diabetes using Machine Learning Techniques over Big Data. Over the years, so many researches are conducted on diabetes prediction, some of them are investigated in this survey. This survey provides the definite outcomes of various researches conducted on diabetes disease and it will be helpful to researchers to work with new ideas with previous work guidance. In this paper, different diabetes disease prediction models using Big data and ML techniques are discussed in detail.

Keywords: Healthcare, Diabetes Disease, Machine Learning, Big Data, Prediction of Diabetes.

I.INTRODUCTION

The use of big data in daily life is increasing from health care, social networks, banking systems, entry into the banking system, use of sensors and smart devices, leading to large amounts of data.

That's why, it is necessary to develop model and device that handles data in optimized form. Electronic Health Records (EHR) is growing at an exponential rate that is being stored in enterprise databases or cloud storages. These records have now grown to be called as Big Data. Most of these data are unstructured. The data can be efficiently processed on cloud for lowering the processing costs. Medicine as a big resource of information is closely concerned because healthcare industry has always been a big generator of biomedical data with the electronic health records, scientific instruments, clinical decision support systems or even research articles. New computing sources and biotechnology advances help to generate trillions of data points coming from biomedical enterprises, mHealth, Telehealth and telemedicine [9].

Diabetes mellitus also known as diabetes, a ubiquitous disease and has no permanent treatment [1]. Diabetes is a common chronic disease which poses a significant

danger to human health. This disease appears in people when glucose in the blood is higher than the average level, and it is functionally caused by defective insulin secretion or its impaired biological effects, or both. Diabetic patients cannot effectively convert consumed carbohydrates into glucose sugar that produces energy for day-to-day activities. This leads to a gradual increase in sugar in the bloodstream. Therefore, glucose remains in the bloodstream and will not reach all body cells. Thus, diabetes is considered one of the most significant health challenges globally in the 21st century. Besides, it is a major cause of blindness, kidney failure, heart attacks, stroke, and lower limb amputation [10]. There are 3 main types of DM: Type 1(Occurs in children and adolescents), Type 2 (occurs in 20+ year person but it is also seen in adolescents that are obese) and Gestational DM (GDM) only seen in pregnant women. GDM usually disappears after pregnancy, but the infected women and their children may develop to type 2 diabetes later in life. Typical symptoms of Type 1 and 2 DM can be frequent urination, excessive thirst, blurred vision, constant hunger, bed sweating, sudden weight loss and lack of energy. In Long run, diabetes may lead to damage eyes, heart, kidneys and nerves of diabetes patient if improper medication is done which also leads to death.

However, there is no long-term cure for diabetes, but it can be controlled and prevented if an early prediction is accurately possible. An accurate and timely diagnosis will help patients prevent the diabetes, and it helps the patients find out whether they get diabetes in the early stage. However, the medical resource is limited, and doctors can only make diagnoses for certain number of patients in the limited time. Therefore, most people make an assessment based on their experience and symptoms. However, most patients lack professional medical knowledge, and they are just based on what they know and what they hear so it is inaccurate for patients to make diagnoses for themselves. Hence, it is necessary to make an efficient prediction model, which can save medical resources and help patients make a self-test accurately [12]. The prediction of diabetes is a challenging task, as the distribution of classes for all attributes is not linearly separable [13]. Early detection and symptomatic treatment are essential to ensure the healthy life and well-being of pre-diabetic patients. An intelligent medical diagnosis system based on symptoms, signs, laboratory tests, and observations will be helpful in disease detection and prevention [3]. Prevention is Better Than Cure", if we apply this to medico and health field we can save people from Diabetes.

Identifying the patients with high risk of being admitted to the hospital in nearby future will help the physicians, doctors to take decisions accurately. Hospitals can provide better healthcare to the patients by providing the needed infrastructure at right time. This can only be possible by providing the health care providers with accurate data analysis and better predictions using the available Big data sets. But with the availability of huge unstructured EHR data sets and hundreds of patient attributes it is challenging to find how best the indicators help in predicting the accuracy for risk of hospitalization. New programming framework such as Apache Hadoop which implements MapReduce computational paradigm is good for data intensive applications.

Machine learning is another emerging and trending approach which closely works to solve the real time problems. In generally machine learning is subset of artificial intelligence (widely known as AI) and the

AI is a special field in the computer science approaches. The main goal of machine learning is to reduce burden on programmers. Machine learning models works with data's which are provided by user or initial programmer and the machine analyses our input data set and produces our desired output. Currently in the health care domain various data mining methods are used to find interesting pattern of disease using statistical medical data with the help of machine learning algorithms. Machine learning classification methods are the most commonly used methods for computer-aided diagnostics. Machine learning approach can be applied for prediction of diseases and provide automated diagnosis under the validation of professional doctor. Big data analysis is very handy and useful in healthcare to use the data to generate the medical data with spark and machine learning algorithms to predict health issues. This can help the people to get information about the health risks earlier and to get alerts about health issues. It can help physicians also to monitor the patients in real-time by using wearable devices. It also supports the diagnosis of diseases by using machine learning recommendation systems [11]. Due to the importance of the research related to diabetes, many data mining researchers attempted to apply machine learning methods to large datasets of diabetic patients' records [18]. Hence in this research, A survey on diabetes prediction models using machine learning over big data is presented.

II. LITEARTURE SURVEY

In recent years, plenty of methods have been proposed and published for diabetes prediction. Several researchers have attempted to construct an accurate diabetes prediction model over the years. However, this subject still faces significant open research issues due to a lack of appropriate data sets and prediction approaches, which pushes researchers to use big data analytics and machine learning (ML)-based methods. Of course, many strategies should be combined to solve a such difficult situation, but one concept can help reducing its impact; its the integration of Big Data and Machine learning techniques for personalizing healthcare an predicting outcomes. Otherwise, if one can predict which prostate cancer patient needs a different healthcare then one can increase the precision of outcomes and save lives, time and money.

D. Sharathchandra, M. Raghu Ram et. al., [4] describes ML Based Interactive Disease Prediction Model. The application of Machine learning algorithms to predict diseases is one of the finest methodology to reduce heavy work load on doctors and related medical staff. The accurate estimation and analysis of heart & diabetes disease patients reports data may help in predicting future heart problems including diabetes. . The proposed Dual disease prediction technique is user interactive based method. The proposed method observe inputs from the end user with realistic data to predict heart and diabetes disease. In the presented work, we used Logistic regression model (LR) and Support vector machine (SVM) model for prediction of diseases. The proposed model works with 85 and 78 percent accuracy in prediction of heart and diabetes diseases respectively.

Pınar Cıhan, Hakan Coşkun et. al., [6] describes Performance Comparison of Machine Learning Models for Diabetes Prediction. The aim of this study is to design a model to detect the probability of diabetes in patients at an early stage with maximum accuracy. Therefore, seven machine learning classification algorithms were used, namely Logistic Regression, K-Nearest Neighbors, Support Vector

Machine, Gaussian Naive Bayes, Decision Tree, Random Forest, and Artificial Neural Network. The study was carried out on the Pima Indians Diabetes Database (PIDD) taken from the Kaggle database. The performances of machine learning methods were evaluated according to precision, recall, ROC curve, and PRC criteria. According to the results, the Logistic Regression method is more successful than other methods in classifying diabetes disease accurately.

Simrn Gupta, Uditi Namdev, Vanshay Gupta, Vatsal Chheda, Kiran Bhowmick et. al., [7] describes Data-driven Pre-processing Techniques for Early Diagnosis of Diabetes, Heart and Liver Diseases. This paper focuses on building robust classification models for the prediction of diabetes, heart, and liver disease using a variety of pre-processing techniques to achieve optimal results. The implementation of the models is carried out on datasets sourced from the University of California, Irvine (UCI) Machine Learning Repository. Numerous pre-processing techniques such as feature engineering, data pruning, oversampling for skewed datasets, imputation of missing values, encoding categorical variables, and feature scaling are used in this analysis. These techniques help considerably augment the performance of the classification algorithms used, which include Random Forest, K-Nearest Neighbours (KNN), and Support Vector Machine (SVM) among others. The performance of these algorithms is further improved by hyperparameter tuning, significantly improving the accuracy scores. The maximum accuracies obtained for heart disease, liver disease and diabetes prediction are 90.16%, 73% and 93.23% respectively.

Kamlesh Lakhwani, Sandeep Bhargava, Kamal Kant Hiran, Mahesh M. Bundele, Devendra Somwanshi et. al., [14] presents Prediction of the Onset of Diabetes Using Artificial Neural Network and Pima Indians Diabetes Dataset. An automatic diagnosis system is introduced and analyzed. For this purpose, a Three-Layered Artificial Neural Network (ANN) and Pima Indians Diabetes dataset are used. In this analysis, an ANN inspired diabetes prediction model has been proposed. In this ANN based prediction model, a logistic-activation function for activation of neurons, and the Quasi Newton method is used as the algorithm for the training. As a result cumulative gain plot and as a measure of the quality of this model the maximum gain score is used. However, Measurable testing isn't suitable.

2.1 DIABETES PREDICTION TECHNIQUES

Lots of research has been done in disease prediction such as diagnosis, prediction, classification, therapy etc. Recent research shows that various ML (Machine Learning) algorithms have been used for disease identification and prediction. They have resulted in remarkable efficiency and improvement in profound conventional and ML methods. ML has shown their abilities to efficiently and strongly deal with high numbers of variables while making strong predictive models. Some of them are discussed in detail as follows:

Raja Krishnamoorthi, Shubham Joshi, Hatim Z. Almarzouki, Piyush Kumar Shukla, Ali Rizwan, C. Kalpana and Basant Tiwari et. al., [2] describes A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques. Applying four different machine learning methods, the research tries to overcome the problems and investigate healthcare predictive analytics. This study's primary goal was to

see how big data analytics and machine learning-based techniques may be used in diabetes. In this study, the authors utilize this framework to develop and assess decision tree (DT)-based random forest (RF) and support vector machine (SVM) learning models for diabetes prediction, which are the most widely used techniques in the literature at the time of writing. It is proposed in this study that a unique intelligent diabetes mellitus prediction framework (IDMPF) is developed using machine learning. The Fig. 1 shows the framework of described ML techniques.

Pima Indian Diabetes Database is a familiar and commonly used data set for the prediction of diabetes. This data set consists of 768 rows and 9 columns. The attributes included in the column are glucose, pregnancies, skin thickness, blood pressure, BMI, insulin, age, and outcomes. The outcome variable predicts whether the patient is diabetic positive or diabetic-negative.

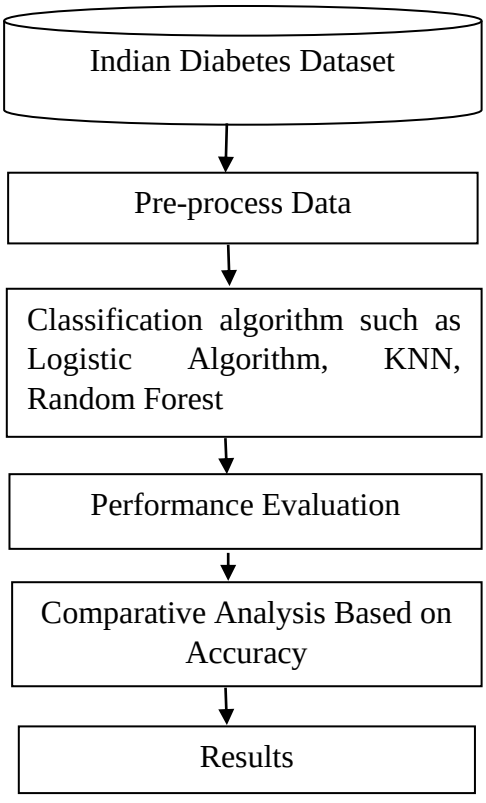


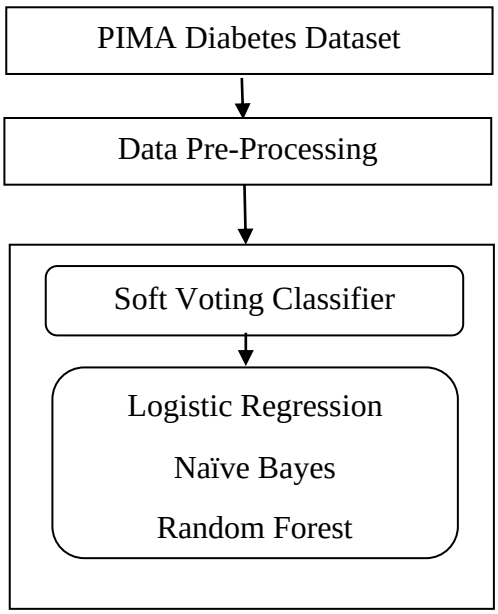
Fig. 1: Framework of ML Techniques

Pre-processing includes the removal of outliers and standardizing the data. The processed data have been used for creating a model. The data should be pre-processed and arranged properly before applying classifiers to the data index. 90%, of the data set, is used for training purposes and the remaining 10% is used for testing by selecting the data randomly. Then, different classifiers such as ML algorithms are applied to diagnose diabetes.

ML classifiers are adapted because of their simplicity and popularity. Hyper-parameter tuning is used to evaluate the ML models. The process of choosing a set of optimal hyper-parameter is known as hyper-parameter tuning. The value of the hyper-parameter's model is fixed before starting the ML task. The ML classification algorithms Logistic Algorithm, K-nearest Neighbour (KNN) and Random Forest are compared based on accuracy. After the evaluation process, one of the best ML classifiers is identified and hyper-parameter tuning has been applied. Using the framework, the authors describe the training procedures, model assessment strategies, and issues associated with diabetes prediction, as well as solutions they provide. The findings of this study may be utilized by health professionals, stakeholders, students, and researchers who are involved in diabetes prediction research and development. This work gives 83% accuracy with the minimum error rate. The drawback of the study is that authors have selected a structured data set, and the unstructured data will be considered for the future.

Saloni Kumari, Deepika Kumar, Mamta Mittal et. al., [5] presents An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. The main objective of

this study is to predict diabetes mellitus with better accuracy using an ensemble of machine learning algorithms. The Fig. 2 shows the flow diagram of described approach. The Pima Indians Diabetes dataset has been considered for experimentation, which gathers details of patients with and without having diabetes. Data Pre-processing is an important step that is used to transform the data in a useful and efficient format so that it can be fed to the machine learning algorithm. Described ensemble soft voting classifier gives binary classification and uses the ensemble of three machine learning algorithms viz. random forest, logistic regression, and Naive Bayes for the classification. Experimentation has been conducted using PIMA diabetes dataset. The dataset has 769 data points and 10 feature columns where zero has been replaced with their median values. The dataset has been divided into testing and training datasets with 20% and 70% respectively. Empirical evaluation of the presented methodology has been conducted with state-of-the-art methodologies and base classifiers such as AdaBoost, Logistic Regression, Support Vector machine, Random forest, Naïve Bayes, Bagging, GradientBoost, XGBoost, CatBoost.



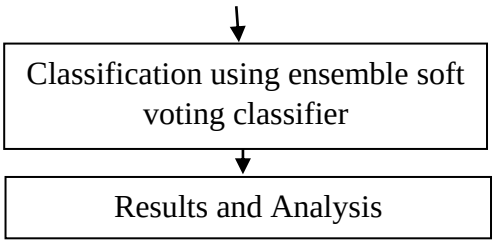
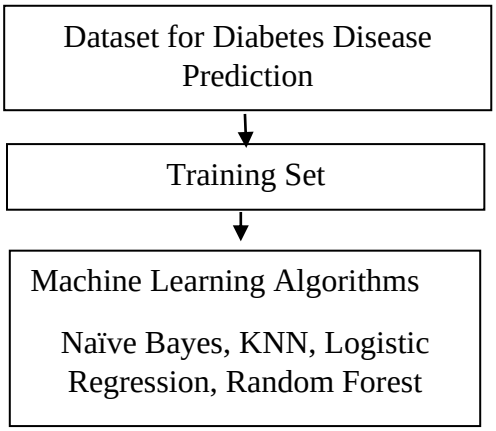


Fig. 2: Flow Diagram of Described Approach

Accuracy, Precision, Recall, F1 score are the most common evaluation metrics adopted for checking the robustness and efficiency of the algorithms. The ensemble soft voting classifier has given 79.08% accurate results on the Pima Indians diabetes dataset.

Srinivasa Rao Swarna, Sumati Boyapati, Pooja Dixit, Rashmi Agrawal et. al., [8] describes Diabetes prediction by using Big Data Tool and Machine Learning Approaches. The flowchart of described methodology is shown in Fig. 3. The main objective of the paper is to observe diabetes disease with the help of big data tools and machine learning model. For doing this, the authors can select more accurate model with the help of some matrices. This work predicts diabetes disease using four machine learning models and then compare their performance among themselves. Machine learning provides more flexible and scalability than the older bio statistical method that helps it perform a variety of tasks such as risk detection, diagnosis, classification, and prediction.



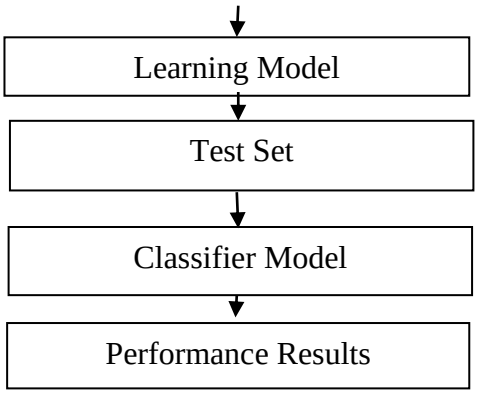
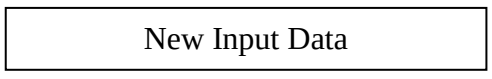


Fig. 3: Flowchart of described Methodology

Described classifier model primarily cautions patients with diabetes, and takes input into the data set for diabetes. Different models of machine learning algorithms such as random forest, KNN algorithm, logistic regression, naive bayes have been tested on the data sets taken in the input and the results generated from them have been collected based on the experimental results. The best performing algorithm has been used on the basis of its accuracy which can make accurate prediction of the disease preceding diabetes disease

K.VijiyaKumar et. al., [16] describes Random Forest Algorithm for the Prediction of Diabetes. The objective of this work is to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by using Random Forest algorithm as a machine learning technique. The datasets are collected from the database. Data pre-processing is one vital step in data discovery methodology. Most health care information contain missing value, wheezy and inconsistency information. In phase two the data will be pre-processed which will include data cleaning, integration and transformation. The Fig. 4 shows the system architecture.



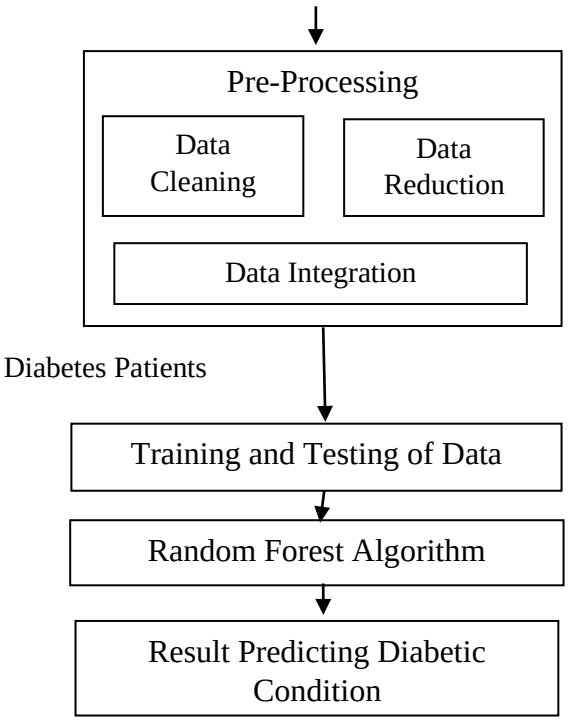


Fig. 4: System Architecture

Data cleaning is that the tactic of detection and correcting (or removing) corrupt or inaccurate records from a record set, table, or data and refers to distinguishing incomplete, incorrect, inaccurate or tangential parts of the knowledge some substitution, modifying, or deleting the dirty or coarse data. Data integration could be a method within which heterogeneous knowledge is retrieved Associate in Nursing combined as an incorporated kind and structure. Data reduction is that the transformation of numerical or alphabetical digital data derived through empirical observation or by experimentation into a corrected, ordered, and simplified type. Random Forest algorithms are often used for each classification and regression tasks and also it is a type of ensemble learning method. The accuracy level is greater when compared to other algorithms. The proposed model gives the best results for diabetic prediction and the result showed that the prediction system is capable of

predicting the diabetes disease effectively, efficiently and most importantly, instantly.

Ayman Mir, Sudhir N. Dhage et. al., [19] presents Diabetes Disease Prediction using Machine Learning on Big Data of Healthcare. This work aims at building a classifier model using WEKA (Waikata Enviroment for Knowledge Analysis) tool to predict diabetes disease. The workflow of presented approach is shown in Fig.5.

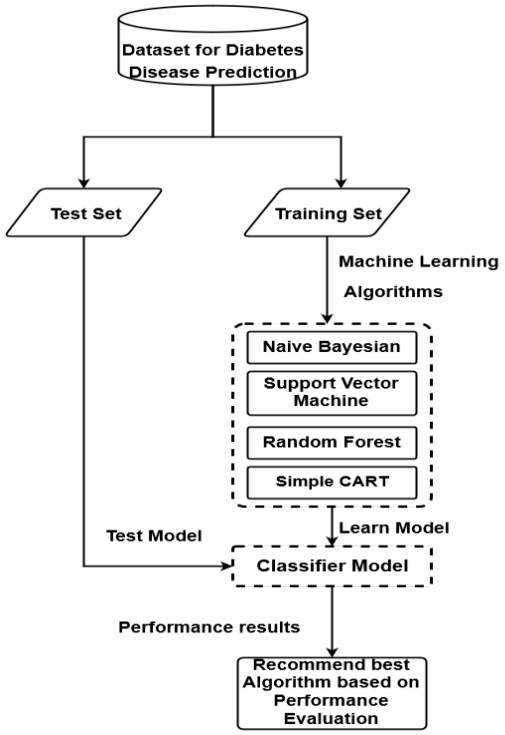
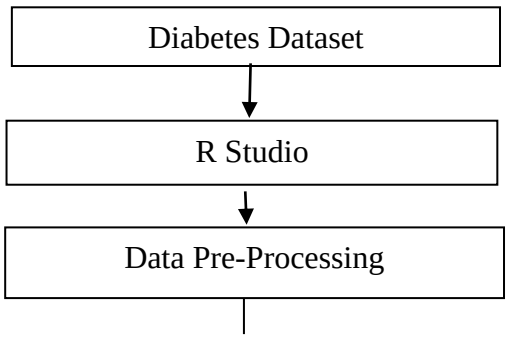


Fig. 5: Described Methodology Flowchart

WEKA is a very popular machine learning and data mining toolkit for conducting data driven researches. The version of WEKA used for experimentation in this paper is WEKA Version 3.82. The research hopes to recommend the best algorithm based on efficient performance result for the prediction of diabetes disease. The described Classifier model specifically considered diabetes disease and takes input of the dataset for diabetes. The input dataset is processed using four machine learning algorithms that are Naive Bayes, SVM, Random Forest, Simple CART and for each algorithm respective classifier model is trained and tested and the results are gathered. Based on the experimental results the best performing algorithm can be determined which will help in accurate prediction of the disease Experimental results of each algorithm used on the dataset was evaluated. It is observed that Support Vector Machine performed best in prediction of the disease having maximum accuracy.

P. Suresh Kumar, S. Pranavi et. al., [24] presents Performance Analysis of Machine Learning Algorithms on Diabetes Dataset using Big Data Analytics. The aim of this paper is to analyze and compare different machine learning algorithms to identify a best predicting algorithm based on various metrics such as accuracy, kappa, precision, recall, sensitivity and specificity. A comprehensive study is done on diabetes dataset with Random Forest (RF), SVM (Support Vector Machine), k-NN, CART (Classification and Regression Tree) and LDA (Linear Discriminant Analysis) algorithms. The architecture of presented approach is shown in Figure 6. Further Data pre-processing is done on loaded dataset with cross validation method with 10 folds and this process is repeated 3 times. This is a common configuration or standard method for comparing different models. Next to that, the pre-processed data is randomly divided into two sets namely training set and test set with the ratio of 80: 20 respectively which is commonly used ratio in literature.



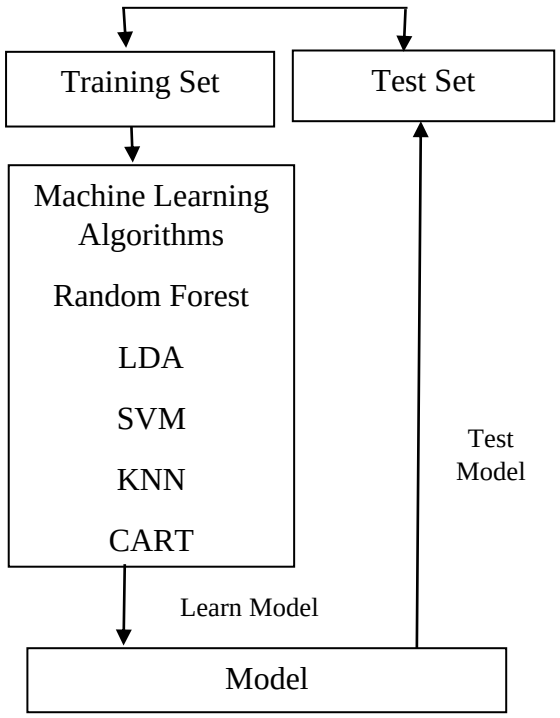


Fig. 6: Architecture of Described Approach

Apply different machine learning algorithms such as RF, LDA, CART and k-NN to learn the data patterns and train the data to get predictions. Then learn about the model to test the predictions with test dataset. After this step, analysis is performed based on accuracy and kappa metrics. The achieved results show that RF algorithm is predicting the data more correctly and accurately.

Few of the research works on diabetes disease prediction using different methods are tabulated in table 1.

Table 1: Different approaches for Diabetes disease pr

Authors	Title	Description
Hala Alshamlan, Areej Al Sahow, Hind Bin Taleb et. al., [15]	A Gene Prediction Function for Type 2 Diabetes Mellitus using Logistic Regression.	Diabetic and normal persons are preidcted by using fisher score feature selection, chi-2 feature selection and Logistic Regression supervised learning algorithm with best accuracy of 90.23%.
Geetha Guttikonda, Madhavi Katamaneni, MadhaviLatha Pandala et. al., [17]	Diabetes Data Prediction Using Spark and Analysis in Hue Over Big Data	Authors have used predictive analysis in HUE to foresee the diseases that are persistent in nature. Here, the dataset is collected from the Pima Indian database. This framework along with SVM Classification gives an effective method to count the number of persons who are suffering from diabetes.
Pahulpreet Singh Kohli, Shriya Arora et. al., [20]	Application of Machine Learning in Disease Prediction	In this work, authors applied different classification algorithms, each with its own advantage on three separate databases of disease (Heart, Breast cancer, Diabetes) available in UCI repository for disease prediction. The feature selection for each dataset was accomplished by backward modeling using the p-value test. The results of the study strengthen the idea of the application of machine learning in early detection of diseases.

Muhammad Azeem Sarwar, Nasir Kamal, Wajeeha Hamid, Munam Ali Shah et. al., [21]	Prediction of Diabetes Using Machine Learning Algorithms in Healthcare	This paper discusses the predictive analytics in healthcare, six different machine learning algorithms are used in this research work. For experiment purpose, a dataset of patient's medical record is obtained and six different machine learning algorithms are applied on the dataset. This paper aims to help doctors and practitioners in early prediction of diabetes using machine learning techniques.
S. K. Dey, A. Hossain, and M. M. Rahman et. al., [22]	Implementation of a web application to predict diabetes disease: An approach using machine learning algorithm	The main aim of this exploration is to build a web application based on the higher prediction accuracy of some powerful machine learning algorithm. We have used a benchmark dataset namely Pima Indian which is capable of predicting the onset of diabetes based on diagnostics manner. With an accuracy of 82.35% prediction rate Artificial Neural Network (ANN) shows a significant improvement of accuracy.
S. Wei, X. Zhao, and C. Miao et. al., [23]	A comprehensive exploration to the machine learning techniques for diabetes identification	Author make a comprehensive exploration to the most popular techniques (e.g. DNN (Deep Neural Network), SVM (Support Vector Machine), etc.) used to identify diabetes and data preprocessing methods. They compare the accuracy of each classifier over several ways of data preprocessors and modified the parameters to improve their accuracy. The best technique we find has 77.86% accuracy using 10-fold cross-validation.
Chandrasegar Thirumalai, K Vamsi Krishna, G V SaiSharan, Kota Jayadev Senapathi et. al., [25]	Prediction of Diabetes Disease using Control Chart and Cost Optimization-Based Decision	This approach is used to detect how diabetes is occurring in different age groups, genders and also can be implemented in the future to include prediction based on region. This metric analysis helps us in the new age of Smart phones where a phone accepts very fewer inputs and gives us an out of the bound output. In this scenario, Predictive analysis on diabetes can be determined with just 9 attributes.
Thanga Prasad. S, Sangavi. S, Deepa. A, Sairabanu. F, Ragasudha. R et. al., [26]	Diabetic Data Analysis In Big Data With Predictive Method	Big data analystics in hadoop implementation present methodical approach designed for reach in good health result similar to availability and affordability of healthcare once-over on the way to every residents. This project is mainly focused for both use of rural and urban neighborhood.
Ihsan Salman Jasim, Adil Deniz Duru,	Evaluation and Measuring Classifiers	In this paper classification performed on diseases diagnoses by choosing to work with

Khalid Shaker, Baraa M. Abed, Hadeel M. Saleh et. al., [27]	of Diabetes Diseases	(k-nearest neighborhood algorithm KNN) measure and evaluate the method with (Artificial Neural Network ANN).T-test used to validate choosing two different factors (K in KNN and number of hidden layers in ANN). After performing classification by changing architecture, ANN proves better results than KNN in this disease classification.
Sreekanth Rallapalli, Suryakanthi T et. al., [28]	Predicting the Risk of Diabetes in Big Data Electronic Health Records by using Scalable Random Forest Classification Algorithm	Classification algorithms like Naive Bayes, Linear Regression; generalized additive model, Random Forest, Logistic Regression, Hidden Markov Models has to be considered for developing a predictive models. In this work, author presented a predictive model using scalable Random forest classification algorithm which can accurately identify the classifier rate for risk of diabetes.
Lin Li et. al., [29]	Diagnosis of Diabetes using a Weight-Adjusted Voting Approach	The Pima Indians diabetes data set (268 diabetes patients and 500 normal subjects) was used in the work. A wrapper method was adopted to select features for classification. An experimental comparison of this method with other voting strategies and each single classifier used in our study demonstrated that this approach performed better in sensitivity.
Meng, X.-H, Huang, Y.-X, Rao, D.-P.; Zhang, Q, Liu, Q et. al., [30]	Comparison of three data mining models for predicting diabetes or prediabetes by risk factors	Three predictive models are developed using 12 input variables and one output variable from the questionnaire information; we evaluated the three models in terms of their accuracy, sensitivity and specificity. The decision tree model (C5.0) had the best classification accuracy, followed by the logistic regression model, and the ANN gave the lowest accuracy.

III. CONCLUSION

In this work, A survey on diabetes prediction models using machine learning over big data is done. Diabetes is a common disease and its early symptoms are not very noticeable, so an efficient method of prediction will help patients make a self-diagnosis. Machine Learning is a very promising approach which helps in early diagnosis of disease and might help the practitioners in decision making for diagnosis. The health care industry needs to incorporate the Big data tool and Machine Learning to develop a system which will ease the prediction and treatment of diabetes for doctors. Several authors have been described so many methods for the prediction of diabetes using machine learning and big data. This research work has made sufficient investigations on diabetic prediction models in health care industry that initiates huge information examination process. The outline of prescient investigation arrangement for diabetes prediction may deliver progressive information and analysis that capitulate the best outcomes

in medicinal services. In this survey different methods for diabetes prediction are investigated. However, most of them are inaccurate and the accuracy need to be improved for accurate diabetes prediction and diagnosis. Hence as a future work, a multimodal early diabetes disease prediction using Hybrid Machine Learning Algorithms over Big data will be presented which will achieve high prediction accuracy than existing approaches.

IV. REFERENCES

[1] S. Reshmi, Saroj Kr. Biswas, Arpita Nath Boruah, Dalton Meitei Thounaojam, Biswajit Purkayastha, "Diabetes Prediction Using Machine Learning Analytics", 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON), DOI: 10.1109/COM-IT-CON54601.2022.9850922

[2] Raja Krishnamoorthi, Shubham Joshi, Hatim Z. Almarzouki, Piyush Kumar Shukla, Ali Rizwan, C. Kalpana,6 and Basant Tiwari, "A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques", Hindawi Journal of Healthcare Engineering Volume 2022, Article ID 1684017, 10 pages, doi:10.1155/2022/1684017

[3] Usama Ahmed, Ghassan F. Issa, Muhammad Adnan Khan, Shabib Aftab, Muhammad Farhan Khan, Raed A. T. Said, Taher M. Ghazal And Munir Ahmad, "Prediction of Diabetes Empowered With Fused Machine Learning", IEEE ACESS, Volume 10, 2022, doi: 10.1109/ACCESS.2022.3142097

[4] D. Sharathchandra, M. Raghu Ram, "ML Based Interactive Disease Prediction Model", 2022 IEEE Delhi Section Conference (DELCON), DOI: 10.1109/DELCON54057.2022.9752947

[5] Saloni Kumari, Deepika Kumar, Mamta Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier", International Journal of Cognitive Computing in Engineering 2 (2021)- 40-46, doi:10.1016/j.ijcce.2021.01.001

[6] Pınar Cıhan, Hakan Coşkun, "Performance Comparison of Machine Learning Models for Diabetes Prediction", 2021 29th Signal Processing and Communications Applications Conference (SIU),DOI: 10.1109/SIU53274.2021.9477824

[7] Simrn Gupta, Uditi Namdev, Vanshay Gupta, Vatsal Chheda, Kiran Bhowmick, "Data-driven Preprocessing Techniques for Early Diagnosis of Diabetes, Heart and Liver Diseases", 2021 Fourth International Conference on Electrical, Computer and Communication Technologies (ICECCT), DOI: 10.1109/ICECCT52121.2021.9616835

[8] Srinivasa Rao Swarna, Sumati Boyapati, Pooja Dixit, Rashmi Agrawal, "Diabetes prediction by using Big Data Tool and Machine Learning Approaches", Proceedings of the Third International Conference on Intelligent Sustainable Systems [ICISS 2020] IEEE Xplore Part Number: CFP20M19-ART; ISBN: 978-1-7281-7089-3

[9] Adil LAABIDI and Mohammed AISSAOUI, "Performance analysis of Machine learning classifiers for predicting diabetes and prostate cancer",2020 1st International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET), DOI: 10.1109/IRASET48871.2020.9092255

[10] Hager Ahmed, Eman M.G. Younis, Abdelmgeid A. Ali, "Predicting Diabetes using Distributed Machine Learning based on Apache Spark", 2020 International

.Conference on Innovative Trends in Communication and Computer Engineering (ITCE), DOI: 10.1109/ITCE48509.2020.9047795

[11] Ahmed Ismail Ebada, Samir Abdelrazek, Ibrahim Elhenawy, "Applying Cloud Based Machine Learning on Biosensors Streaming Data for Health Status Prediction", 2020 11th International Conference on Information, Intelligence, Systems and Applications (IISA), doi: 10.1109/IISA50023.2020.9284349

[12] Juncheng Ma, "Machine Learning in Predicting Diabetes in the Early Stage", 2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), DOI 10.1109/MLBDBI51377.2020.00037

[13] Md. Kamrul Hasan, Md. Ashraful Alam, Dola Das, Eklas Hossain, Mahmudul Hasan, "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers", IEEE ACCESS, Volume 8, 2020, doi: 10.1109/ACCESS.2020.2989857

[14] Kamlesh Lakhwani, Sandeep Bhargava, Kamal Kant Hiran, Mahesh M. Bundele, Devendra Somwanshi, "Prediction of the Onset of Diabetes Using Artificial Neural Network and Pima Indians Diabetes Dataset", 5th IEEE International Conference on Recent Advances and Innovations in Engineering-ICRAIE 2020 (IEEE Record#51050), doi: 10.1109/ICRAIE51050.2020.9358308

[15] Hala Alshamlan, Areej Al Sahow, Hind Bin Taleb, "A Gene Prediction Function for Type 2 Diabetes Mellitus using Logistic Regression", 2020 11th International Conference on Information and Communication Systems (ICICS) 978-1-7281-6227-0/20, 2020 IEEE 10.1109/ICICS49469.2020.23954

[16] K.VijiyaKumar, "Random Forest Algorithm for the Prediction of Diabetes", Proceeding of International Conference on Systems Computation Automation and Networking 2019, 978-1-7281-1524-5, 2019 IEEE

[17] Geetha Guttikonda, Madhavi Katamaneni, MadhaviLatha Pandala, "Diabetes Data Prediction Using Spark and Analysis in Hue Over Big Data", Proceedings of the Third International Conference on Computing Methodologies and Communication (ICCMC 2019) IEEE Xplore Part Number: CFP19K25-ART; ISBN: 978-1-5386-7808-4

[18] Ayman Alahmar, Emad A. Mohammed, Rachid Benlamri, "Application of Data Mining Techniques to Predict the Length of Stay of Hospitalized Patients with Diabetes", 2018 4th International Conference on Big Data Innovations and Applications, 978-1-5386-7793-3/18, 2018 IEEE DOI 10.1109/Innovate-Data.2018.00013

[19] Ayman Mir, Sudhir N. Dhage, "Diabetes Disease Prediction using Machine Learning on Big Data of Healthcare", 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), DOI: 10.1109/ICCUBEA.2018.8697439

[20] Pahulpreet Singh Kohli, Shriya Arora, "Application of Machine Learning in Disease Prediction", 2018 4th International Conference on Computing Communication and Automation (ICCCA), 978-1-5386-6947-1/18, 2018 IEEE

[21] Muhammad Azeem Sarwar, Nasir Kamal, Wajeeha Hamid, Munam Ali Shah, "Prediction of Diabetes Using Machine Learning Algorithms in Healthcare", Proceedings of the 24th International Conference on Automation & Computing, Newcastle University, Newcastle upon Tyne, UK, 6-7 September 2018

[22] S. K. Dey, A. Hossain, and M. M. Rahman, ''Implementation of a web application to predict diabetes disease: An approach using machine learning algorithm,'' in Proc. 21st Int. Conf.

Comput. Inf. Technol. (ICCIT), Dec. 2018, pp. 21–23, doi: 10.1109/ICCITECHN.2018.8631968.
[23] S. Wei, X. Zhao, and C. Miao, ''A comprehensive exploration to the machine learning techniques for diabetes identification, in Proc. IEEE 4th World Forum Internet Things (WF-IoT), Feb. 2018, pp. 291–295, doi: 10.1109/WF-IoT.2018.8355130.
[24] P. Suresh Kumar, S. Pranavi, "Performance Analysis of Machine Learning Algorithms on Diabetes Dataset using Big Data Analytics", 2017 International Conference on Infocom Technologies and Unmanned Systems (ICTUS'2017), Dec. 18-20, 2017, ADET, Amity University Dubai, UAE, 978-1-5386-0514-1/17.S
[25] Chandrasegar Thirumalai, K Vamsi Krishna, G V SaiSharan, Kota Jayadev Senapathi, "Prediction of Diabetes Disease using Control Chart and Cost Optimization-Based Decision", International Conference on Trends in Electronics and Informatics ICEI 2017, 978-1-5090-4257-9/17,2017 IEEE
[26] Thanga Prasad. S, Sangavi. S, Deepa. A, Sairabanu. F, Ragasudha. R, "Diabetic Data Analysis In Big Data With Predictive Method", 2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET),DOI: 10.1109/ICAMMAET.2017.8186738
[27] Ihsan Salman Jasim, Adil Deniz Duru, Khalid Shaker, Baraa M. Abed, Hadeel M. Saleh, "Evaluation and Measuring Classifiers of Diabetes Diseases", ICET2017, Antalya, Turkey, 978-1-5386-1949-0/17, 2017 IEEE
[28] Sreekanth Rallapalli, Suryakanthi T, "Predicting the Risk of Diabetes in Big Data Electronic Health Records by using Scalable Random Forest Classification Algorithm", 2016 International Conference on Advances in Computing and Communication Engineering(ICACCE),DOI: 10.1109/ICACCE.2016.8073762
[29] Lin Li, "Diagnosis of Diabetes using a Weight-Adjusted Voting Approach", 2014 IEEE 14th International Conference on Bioinformatics and Bio-engineering
[30] Meng, X.-H.; Huang, Y.-X.; Rao, D.-P.; Zhang, Q.; Liu, Q, "Comparison of three data mining models for predicting diabetes or prediabetes by risk factors", Kaohsiung J. Med. Sci. 2013, 29, 93–99