DiabDoc : Diabetes Prediction System

Sujal Pose
Department of Computer Engineering
School of Engineering & Applied Sciences

Dr. Uttara Gogate
Department of Computer Engineering
SSJCOE, Dombivli

Abstract— **Diabetes is one of the most surprisingly deadly diseases, serious illnesses, and many people suffer from it intentionally or unknowingly around the world. Diabetes is caused by the condensation of excess sugar in the blood. Age, obesity, lack of exercise, hereditary diabetes, lifestyle, poor diet, and high blood pressure. People with diabetes are at increased risk of heart disease, kidney disease, stroke, eye problems, and nerve damage. Numerous computer-based detection systems have been designed but normal identification process for diabetics requires more time and money. With the rise of machine learning, we have the ability to develop solutions to this serious problem. Medical professionals need a reliable predictive framework for analyzing diabetes. Machine learning can be used to study large datasets, find hidden information and patterns, discover knowledge from the data, and predict outcomes accordingly.**

In this paper, we studied the effectiveness of machine learning algorithms for two different diabetes datasets. We have developed a system that incorporates machine learning algorithms so that users can interact with the system in an effective way, provide the necessary information and observe predictions in their applications. We feeded the data in to various models and tried to compare the better accuracy among then we found Random Forest model promising and embedded our application with that model.

Keywords—diabetes, prediction, machine learning, healthcare, dataset

I. INTRODUCTION

Health has always been a priority, even before technology exists. The healthcare domain is so evolved that it offers a lot of research scope. Existing medical technologies need to be upgraded by adopting the digitization of medical information, both in terms of patient-provided data and medical outcomes generated from advanced equipment. The general consequence of this information revolution is that we are faced with the difficult task of interpreting and understanding the vast amount of data collected. Machine learning is useful because of the large amount of data. [1]

Today, so many chronic diseases are prevalent in the world, and such serious diseases are prevalent in both developing and developed countries. Of these serious illnesses, diabetes is the most life-threatening chronic illnesses in the world. Diabetes is a condition that causes deficiency due to low levels of insulin in the blood. High blood sugar warning signs result in frequent urination, thirst, and increased hunger. Not taking medicine can lead to many difficulties. This difficulty leads to death. Serious problems lead to cardiovascular disease, foot pain and blurred vision. To sumarize if your blood sugar is elevated, it's called diabetes.

According to the World Health Organization, diabetes is one of the leading causes of death in the world, with approximately 422 million people worldwide. In fact, 1.6 million people died in 2016 [2]. There are two main types of diabetes, type 1 and type 2. Type 1 diabetes accounts for 5-10% of all diabetes cases. This type of diabetes most often appears in childhood or adolescence and is characterized by partial functioning of the pancreas. The disease becomes apparent only when 80-90% of the insulin-producing cells in the pancreas have already been destroyed [3]. Type 2 diabetes accounts for 90% of all diabetes cases. This type of diabetes is characterized by chronic hyperglycemia and the inability of the body to regulate blood sugar levels.[4] In medicine, doctors and current research have confirmed that early detection of the disease has a high chance of recovery. With continuous advances in technology, machine learning and deep learning techniques are extremely useful for early prediction and provide automated diagnosis under expert validation.

Machine learning is a new trend approach that works closely to solve real-time problems. It is seen as an urgent need for today's situation to eliminate human effort by supporting automation with minimal flaws. A technique called predictive analytics incorporates a variety of machine learning algorithms, data mining techniques, and statistical techniques that use current and historical data to find knowledge and predict future events. It aims to diagnose disease with the highest possible accuracy, enhance patient care, improve clinical outcomes, and optimize resources.[5] Various information mining algorithms provide different decision support systems to assist healthcare professionals. The effectiveness of a decision support system is recognized by its accuracy. The existing method for detecting diabetes is to use laboratory tests such as fasting blood glucose and oral glucose tolerance tests. However, this method is time consuming.[6] Therefore, the goal is to build a decision support system to predict and diagnose a particular disease with great accuracy.

II. RELATED WORK

Amani Y, Akhtar J, Jawad R, and Mirasat Y performed a comparative analysis of machine learning and deep learning-based algorithms to predict diabetes. The results showed that RF was effective in classifying diabetes in all experimental rounds, with an overall accuracy of diabetes prediction of 83.67%. The SVM prediction accuracy reached 65.38%, but the DL method generated 76.81% on the dataset. [7] Aada A. and Sakshi Tiwari, various classifiers and decision trees for past research achieve the most notable accuracy of 94.44%. The selection tree is simple and is an excellent classifier for expected diabetes. An inspection of the accuracy created by all classifiers before applying the re-template and the accuracy provided by them after applying the similarities. [8] Alehegn, Minyechil, Rahul Joshi, and Preeti Mulay show that the method proposed in this study provides high accuracy with an accuracy value of 90.36%, and decision Stump is lower than others by providing 83.72% accuracy. Therefore, use the ensemble method used to provide better predictive performance or accuracy than a single one.[9] Maniruzzaman, Md, et al. The hypothesis was used in an ML-based system that used a combination of LR-RF for feature selection techniques, and the classifier showed the highest classification accuracy. The results showed that the proposed combination reached 94.25% accuracy with the K10 protocol. [10] Mir, Ayman, and Sudhir N. Dhage. Four classifiers based on the machine learning algorithms of, Naive Bayes, Support Vector Machine, Random Forest, and Simple CART have been used to experiment with WEKA tools for predicting diabetes. [11] Sonar, Priyanka, and K. Jayamarini. Comparative studies conducted at SVM, Decision Tree, and Naive Bayes to predict diabetes [12] Hasan, Md Kamrul, et al. comparison results show that the proposed framework outperforms other frameworks in AUC, showing great potential for predicting diabetes from PID datasets. An ensemble of two boosting type

classifiers (AB and XB) is the perfect combination for diabetes prediction. This is because the basic classifier needs to have the least correlation between them.. [13] Mujumdar, Aishwarya, and V. Vaidehi were categorized using various machine learning algorithms applied to the dataset and logistic regression providing the highest accuracy of 96%. The application of the pipeline gave the AdaBoost classifier as the best model with 98.8% accuracy. [14] This study by Joshi, Tejas N., and P.P.M. Chawan describes a machine learning approach to predicting diabetes levels. [15] Dey, Samrat Kumar, Ashraf Hossain, and Md Mahbubur Rahman have proposed the Web as a base application for successful diabetes prediction. From a variety of machine learning algorithms, Artificial Neural Networks (ANNs) provide the highest accuracy with the minimum and maximum scaling methods for Indian pima datasets. [16]

III. Proposed System

Based on the problems explained in the introductory part, we propose an interactive system that can predict the presence or absence of diabetes more accurately. This model uses various classifiers such as SVM, Random Forest, Logistic Regression, Decision Tree, KNN, etc. The main focus is the famous Benchmark Diabetes Dataset from the PIMA Indian Diabetes Dataset of the UCI Machine Learning Repository datasets with 8 and 17 attributes and another dataset from the early stage diabetes risk prediction were downloaded. It was collected and approved by a physician using a direct questionnaire from patients at the Sylhet Diabetes Hospital in Sylhet, Bangladesh. We used this dataset for symptom-based prediction. This dataset contains 520 patient records with 17 attributes for maximum accuracy in machine learning techniques. The framework consists of the following important phases, as shown in the following figure. The data in the dataset is first preprocessed to avoid null or redundant data that can affect the prediction, then split later, and then features that make the prediction even more annoying are selected. Machine learning algorithms have been applied to the training and test sets to get the accuracy you need. You can later implement those models and use them to develop your system.

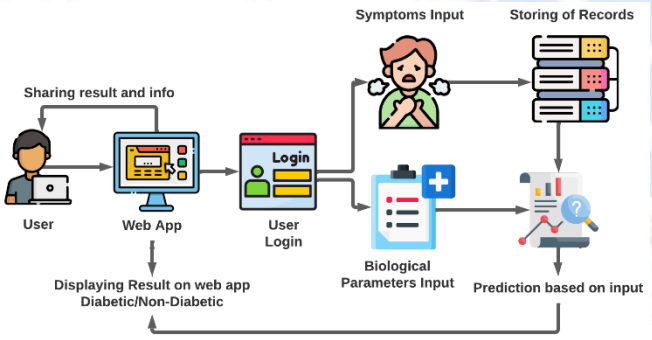

Fig. 1. Process followed for evaluation of algorithms Diabetes Datasets

IV. Methodology

ML provides methods and tools to help solve diagnostic and prognostic problems in a variety of medical environments. It is used to analyze clinical parameters regarding the importance of prognosis and their combination. for example. Predict disease progression and extract medical knowledge for outcome studies, treatment planning and support, and overall patient management. ML is intelligent to detect data regularity by properly processing incomplete data, perform data analysis such as interpretation of continuous data used in intensive care units, and monitor effectively and efficiently. Successful implementation of the ML approach has been claimed to help integrate computer-based systems into the medical environment. This facilitates and enhances the work of health professionals and ultimately improves the efficiency and quality of health care. In medical diagnosis, the main concern is to establish the presence of the disease, followed by its accurate identification. There is a separate category for each disease under consideration, and one category if the disease does not exist. Here, machine learning improves the accuracy of medical diagnosis by analyzing patient data. Measurements in this machine learning application are usually the result of certain medical tests (blood pressure, temperature, various blood tests, etc.). It is also a medical diagnosis (medical images, etc.), the presence / absence / intensity of various symptoms, and basic physical information about the patient (age, gender, weight, etc.). Based on these measurements, doctors narrow down the illnesses that are afflicting the patient..

- ML can learn features from large amounts of medical data and use the insights gained to support clinical practice in treatment design or risk assessment.
- The ML system extracts useful information from large patient populations to support real-time inference for warning of health risks and prediction of health outcomes.
- ML system helps reduce unavoidable diagnostic and therapeutic errors in human clinical practice.
- ML can assist doctors by providing up-to-date medical information from journals, textbooks, and clinical practice to inform them about appropriate patient care.
- ML can support precision medicine and new drug development based on faster processing of mutations and links to diseases.

Despite the various applications of ML in clinical research and healthcare services, they fall into two major categories. Analysis of structured data such as images, genes, and biomarkers, and analysis of unstructured data such as memos, medical journals, and patient surveys complement the structured data.

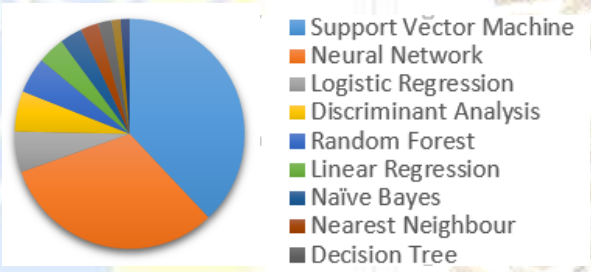

Fig 2: The popular Machine Learning algorithms used in the medical domain.

For a long time, healthcare ML was dominated by logistic regression, the simplest and most common algorithm when things need to be categorized. Easy to use, quick to complete, easy to interpret. Below are some other models and their usage in healthcare domain.

A. *Support Vector Machine*

Support Vector Machines (SVMs) can be used for classification and regression, but this algorithm is primarily used for classification problems that require the division of a data set by a hyperplane into two classes. The goal is to choose a hyperplane with the largest possible margin, or the distance between the hyperplane and any point in the training set so that the new data can be correctly classified.

SVM is widely used in clinical research. For example, it is used to identify imaging biomarkers, diagnose cancer and neuropathy, and generally classify data from imbalanced or missing values.

B. Neural networks

In neural networks, the association between the result and the input variable is represented by a combination of hidden layers of pre-specified functionals. The goal is to estimate the weights from the input and result data so that the mean error between the result and its prediction is minimized.

Neural networks have been successfully applied in a variety of medical environments, including diagnostic systems, biochemical analysis, image analysis, and drug development, using examples from textbooks on breast cancer prediction from mammography images.

C. Logistic Regression

Logistic regression is one of the basic and still popular multivariable algorithms for modeling the results of dichotomy. Logistic regression is used to get the odds ratio when there are multiple explanatory variables. The procedure is similar to multiple regression, except that the response variable is a binomial. This shows the effect of each variable on the odds ratio of the observed event of interest..

In health care, logistic regression is widely used to solve classification problems and predict the probability of a particular event. It will be a valuable tool for risk assessment of illness and improvement of medical decisions.

D. Natural Language Processing

In healthcare, most clinical information is in the form of descriptive text and there are no special unstructured and computerized text processing methods that the program does not understand.Natural language processing addresses these issues by identifying a set of disease-related keywords in clinical notes based on a history database that inputs and extends structured data after validation to support clinical decision making.

E. Naïve Bayes

A naive Bayes classifier is a baseline method of text classification and is the problem of determining a document as belonging to one of the categories. The naive Bayes classifier assumes that the presence of a particular feature in a class is independent of the presence of other features. All of these properties contribute independently to the probability of belonging to a particular category, even if these features are interdependent.

It is one of the most effective yet efficient classification algorithms and is applied to many medical cases, such as classification of journal articles and medical reports.

F. Deep Learning

Deep learning is an extension of classical neural network technology, simply put, a neural network with many layers. Deep learning has more capacity than traditional ML algorithms and can explore more complex nonlinear patterns in the data.

In medical applications, deep learning algorithms handle both machine learning and natural language processing tasks appropriately. Commonly used deep learning algorithms include convolutional neural networks (CNNs), recurrent neural networks, deep belief networks, and multi-layer perceptrons, and CNN has been a competitor since 2016.

G. Convolutional Neural Network

CNN was developed to process data with many properties, such as high-dimensional data and images. The CNN transfers the pixel values of the image by weighting the image in the convolution layer and sampling the image in the subsampling layer. The final output is a recursive function of the weighted input values.

Recently, CNN has been successfully implemented in the medical field to assist in the diagnosis of diseases such as skin cancer and cataracts.

V. Result and Discussion

This section describes the results achieved after design of experiments. Table I describes the insights of the Pima Indian dataset. This dataset is based primarily on women who lived in the Pima Indian heritage. The following eight features (1-8) of the Pima Indian dataset will help predict diabetes in an individual with the help of the proposed methodology.

TABLE I. Pima Indian Diabetes Dataset Attributes

Sr No	Attribute	Description
1	Pregnancies	Number of times pregnant
2	Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3	BloodPressure	Diastolic blood pressure (mm Hg)
4	SkinThickness	Triceps skin fold thickness (mm)
5	Insulin	2-Hour serum insulin (mu U/ml)
6	BMI	Body mass index (weight in kg/(height in m)^2)
7	DiabetesPedigree Function	Diabetes pedigree function (Numeric)
8	Age	No. of years (Numeric)
9	Outcome	Class variable (0 or 1) 268 of 768 are 1, the others are 0

Table II describes another dataset of insights for early-stage diabetes risk prediction. It was collected and approved by a physician using a direct questionnaire from patients, can use this dataset for symptom-based prediction. This dataset contains 520 patient records with 17 attributes (1-16) to help predict diabetes in an individual with the help of the proposed methodology.

TABLE II. Early Stage Diabetes Risk Prediction Dataset Attributes

Sr No	Attribute	Description
1	Age	Age in years ranging from (20years to 65 years)
2	Gender	Male / Female
3	Polyuria	Body urinates more than usual (Yes / No)
4	Polydipsia	Medical name for the feeling of extreme thirstiness. (Yes / No)
5	Sudden weight loss	Losing weight frequently (Yes / No)
6	Weakness	Feeling weak physically (Yes / No)
7	Polyphagia	Medical term for excessive or extreme hunger. (Yes / No)

8	Genital Thrush	Affects the vagina, can be irritating and painful. (Yes / No)
9	Visual blurring	Disturbance in a person's eyesight (Yes / No)
10	Itching	Irritating sensation that makes you want to scratch your skin (Yes / No)
11	Irritability	Feeling of anger(Yes / No)
12	Delayed healing	Wound takes time to heal (Yes / No)
13	Partial Paresis	Weakening of muscle (Yes / No)
14	Muscle stiffness	Muscles feel tight (Yes / No)
15	Alopecia	Causes hair to fall out in small patches (Yes / No)
16	Obesity	Complex disease involving an excessive amount of body fat (Yes / No)
17	Class	Diabetes outcome (Positive / Negative)

Several machine learning algorithms were used in this experimental study. These algorithms are KNN, SVM, LR, DT, RF, etc. All of these algorithms have been applied to the PIMA Indian dataset and the early diabetes risk prediction dataset. The data is divided into two parts, training data and test data, respectively. All of these algorithms were applied to the same dataset using the Jupyter Notebook and the results were obtained. Predicting accuracy is the main evaluation parameter used in this task the overall success rate of the algorithm.

According to Table III. Random Forest proves to be efficient with 86.21% detection accuracy based on the diagnostic nature of the Indian Pima dataset other based on medical parameters achieved 96.49% accuracy based on user symptoms. Therefore, we built a web application using an RF model that can predict whether a patient has diabetes.

TABLE III. ACCURACY OF DIFFERENT ML ALGORITHMS BASED ON SYMPTOMS AND MEDICAL PARAMETERS

Sr. No	Model	Diabetes Prediction Accuracy based on	
		Symptoms	Medical Parameters
1	Logistic Regression	90.23	85.34
2	Random Forest	96.49	86.21
3	Support Vector Machine	88.15	84.48
4	K-nearest neighbors	91.02	79.31
5	Decision Tree Classifier	90.10	81.03
6	XGB Classifier	90.34	82.76
7	Naive Bayes	90.18	85.34
8	AdaBoost Classifier	88.42	76.72
9	Gradient Boosting Classifier	95.17	80.17
10	ExtraTrees Classifier	96.34	82.76

Figure 7 below. Represents a web application developed by us that was imbedded based on the highest accuracy of the model. We used Streamlit, an open source Python library for creating and sharing web apps for data science and machine learning projects. Libraries help you create and deploy data science solutions in minutes with just a few lines of code. Streamlit integrates seamlessly with other popular Python libraries used in data science, such as NumPy, Pandas, Matplotlib, and Scikit-learn, and can be used to code implement machine learning models. The architecture we propose typically collects dataset values from a database and trains the model during a training session. During the forecast period, the user will need to provide some information as input, as shown in Figures 7 and 8 below. This allows the developed web application to predict whether the test result will be positive. To test for diabetes, users need to provide the following information in their web application: Some of the most necessary information, such as blood pressure, body mass index (BMI), serum insulin, and oral glucose tolerance test, is needed for parameter-based and symptom-based predictions. The user simply needs to answer "yes / no" to the quaestionaire. Predictions can be made for symptoms such as itching, muscle stiffness, and sudden weight loss.

Fig. 7. Our developed web application based on detection Model.

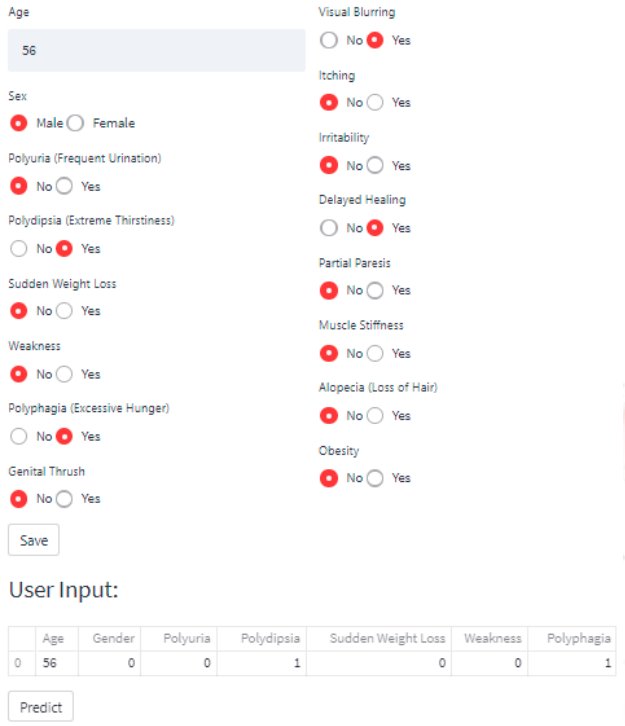

Fig. 8. Our developed web application based on detection Model.

All predicted true positives and true negatives divided by all positives and negatives. The true positives, true negatives, false negatives, and false positives predicted by all algorithms are shown in Figures 9 and 10. In this case, True Pos means real diabetes and predicted diabetes. False Neg, real diabetes, but not predicted to be diabetic. It predicted False Pos and diabetes, but I'm not actually diabetic. True Neg, not really diabetic, prediction is not diabetic.

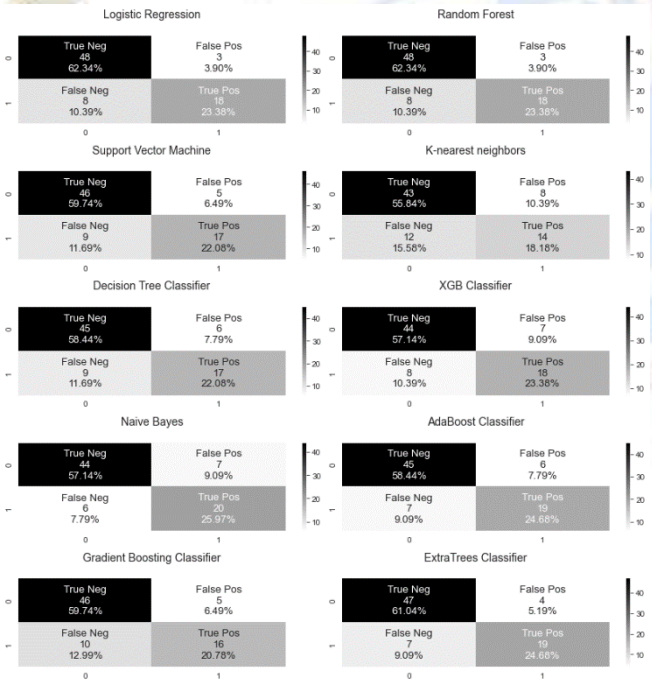

Fig. 9. Confusion matrix for medical parameter based model prediction.

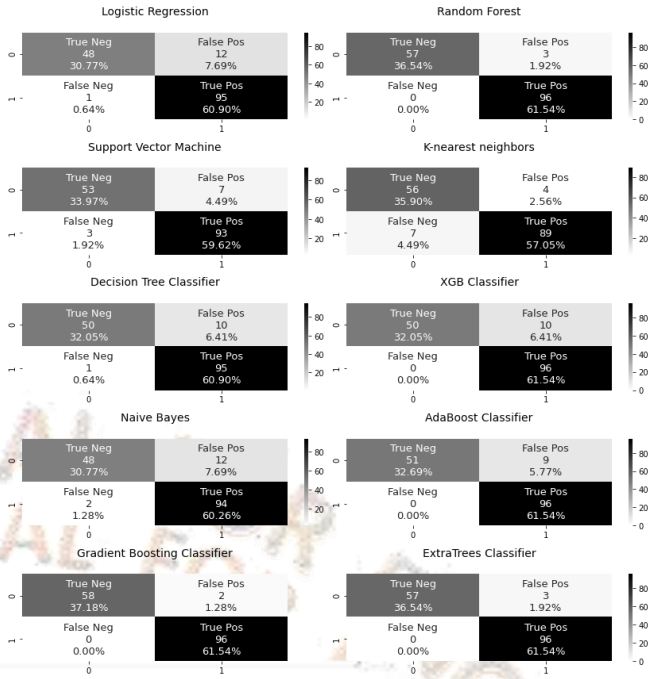

Fig. 10. Confusion matrix for symptoms based model prediction.

To summarize the results of the algorithms on both the dataset we have plotted a relational bar graph which demonstrates the algorithms with their respective accuracies.

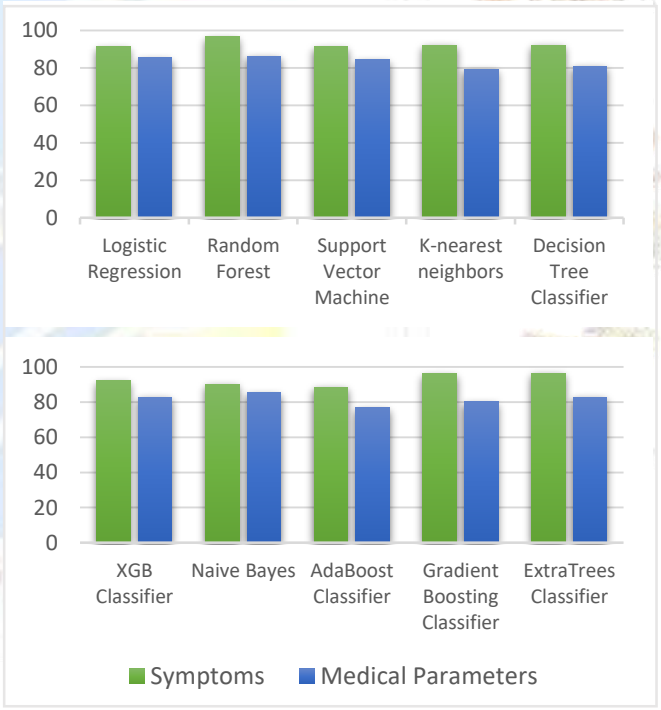

Fig. 11. Graphical representation of Accuracy of different machine learning algorithm based on symptoms and medical parameters.

CONCLUSION

In healthcare predictive analytics can change the way medical researchers, practitioners students gain insights from data to make decisions. This paper used ten common machine learning algorithms including SVM, KNN, LR, DT, RF, NB, etc. Predictions were firstly made for diabetes in the PIMA Indian dataset consisting of 768 records. Eight attributes were selected for training and testing the predictive model. Later, the predictions were made on Early risk prediction dataset for which sixteen attributes were considered for training and testing the predictive model of 520 records. The experimental results obtained show that RF provide the highest accuracy for predicting diabetes in both the cases as it will be based on

medical parameters or symptoms. In both the cases algorithm provide 86.21% accuracy to predict diabetes based on medical parameters and 96.49% accuracy to predict diabetes based on symptoms. This is the best compared to the other algorithms used in this paper. Therefore, we can conclude that RF are suitable for predicting diabetes. Such that, it was being embedded into our web application to provide an interactive way for the user to get a approximate analysis. Limitations of this study missing attribute values and the size of the dataset. To build a predictive model of diabetes best accuracy, we need thousands of records with zero missing values. Future work will focus on integrating other methods into the model in order to adjust the parameters of the model for greater accuracy. Then, testing these models on large datasets with minimal or no missing attribute values reveals more insights and better prediction accuracy.

REFERENCES

[1] Ayman Mir and 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).

[2] G. Swapna, R. Vinayakumar, and K. P. Soman, “Diabetes detection using deep learning algorithms,” ICT Express, vol. 4, no. 4, pp. 243– 246, 2018.

[3] L. Lucaccioni and L. Iughetti, “Issues in Diagnosis and Treatment of Type 1 Diabetes Mellitus in Childhood,” J. Diabetes Mellit., vol. 06, no. 02, pp. 175–183, 2016.

[4] “Type 2 Diabetes: a Review of Current Trends -,” Int. J. Curr. Res. Rev., vol. 7, no. 18, pp. 61–66, 2015.

[5] Gauri D. Kalyankar, Shivananda R. Poojara and Nagaraj V. Dharwadkar,” Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop”, International Conference On I-SMAC,978-1-5090-3243-3,2017.

[6] Aishwarya Mujumdar et al. “Diabetes Prediction using Machine Learning Algorithms 292–299. International conference on recent trends in advanced computing (ICRTAC) 2019.

[7] Yahyaoui, Amani, et al. "A decision support system for diabetes prediction using machine learning and deep learning techniques." 2019 1st International Informatics and Software Engineering Conference (UBMYK). IEEE, 2019.

[8] Aada, A., and Sakshi Tiwari. "Predicting diabetes in medical datasets using machine learning techniques." Int. J. Sci. Eng. Res 5.2 (2019).

[9] Alehegn, Minyechil, Rahul Joshi, and Preeti Mulay. "Analysis and prediction of diabetes mellitus using machine learning algorithm." International Journal of Pure and Applied Mathematics 118.9 (2018): 871-878..

[10] Maniruzzaman, Md, et al. "Classification and prediction of diabetes disease using machine learning paradigm." Health information science and systems 8.1 (2020): 1-14..

[11] Mir, Ayman, and 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). IEEE, 2018..

[12] Sonar, Priyanka, and K. JayaMalini. "Diabetes prediction using different machine learning approaches." 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC). IEEE, 2019..

[13] Hasan, Md Kamrul, et al. "Diabetes prediction using ensembling of different machine learning classifiers." IEEE Access 8 (2020): 76516-76531..

[14] Mujumdar, Aishwarya, and V. Vaidehi. "Diabetes prediction using machine learning algorithms." Procedia Computer Science 165 (2019): 292-299..

[15] Joshi, Tejas N., and P. P. M. Chawan. "Diabetes prediction using machine learning techniques." Ijera 8.1 (2018): 9-13..

[16] Dey, Samrat Kumar, Ashraf Hossain, and Md Mahbubur Rahman. "Implementation of a web application to predict diabetes disease: an approach using machine learning algorithm." 2018 21st international conference of computer and information technology (ICCIT). IEEE, 2018.

[17] Ansa Jovel Kunnathettu, Satishkumar L. Varma. "Comparative Analysis of Neural Network and Machine Learning Techniques for Air Quality Prediction", 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2020