[1]Anand Viswanathan
[2]Ravikumar. K
[3]Saravanakumar. K

Diabetes Mellitus prediction using Improved Chaotic Whale Optimization and Data mining techniques

JES Journal of Electrical Systems

Abstract: - Automation of disease detection is now prevalent in healthcare systems. Diabetes mellitus is a serious issue that has become prominent worldwide. It is a hereditary illness that impairs human existence at every stage of development. Each year, millions more people develop diabetes, which also has an impact on young people. Currently, the medical industry is shifting toward automation, as the process of identifying diseases needs periodic human verification. A single model for diabetes prediction is insufficient for complicated problems since it might not be appropriate for training huge data. In this proposed approach, a novel method which combines the outcomes of more than one algorithm is suggested. An ensemble approach using data mining techniques such as Decision Trees, Random Forests, AdaBoost, Gradient Boosting and XGBoost is employed to predict diabetes mellitus. Feature selection methods such as Averaged Fisher score and Kolmogorov-Smirnov score are used to choose the prominent characteristics from PIMA Indian Diabetes dataset. Further, the most optimal features are selected using Improved Chaotic Whale Optimization algorithm. The proposed approach produces predictions that are 98.8% accurate and reliable.

Keywords: Diabetes Mellitus, Hybrid Feature Selection, Improved Chaotic Whale Optimization, Data Mining .

I. INTRODUCTION

High blood sugar levels cause a condition, diabetes mellitus, which is considered as an autoimmune disorder which provokes insufficient production of insulin and sensitivity by the body [1]. It typically manifests itself in a variety of ways: Prediabetes refers to a blood sugar level that is greater than usual, overt diabetes which exists in two different forms as type I and type II, or gestational diabetes, a form of diabetes which is caused by conception. Diabetes has been scientifically demonstrated to be associated with chronic deterioration of key organs such as the cardiovascular system, arteries and veins, the urinary tract, vision, and neurons [2]. The impact it has on conceptions is even more concerning; every year, diabetes affects 11% of gestational periods, posing a simultaneous risk to the well-being of the pregnant woman and that of the fetus. The prevalence of diabetes has been rising, and by 2035, it's predicted that 56% of people worldwide will have the disease [3].

Diabetes mellitus is a disorder in which the immune system fails to effectively convert carbohydrates into energy. Humans use the process of transformation of carbohydrates into insulin to produce energy. The tissues of the human body use the hormone insulin, which is produced by the digestive tract to absorb sugar [4]. When a person has diabetes, their body is unable to produce or consume sufficient hormones to function as a source of energy. Diabetes frequently manifests as excessive water retention, a desire to eat more, decreased physique, exuberance, prolonged recovery from injury, and increased frequency of urine excretion [5].

Diabetes has an impact on various body systems and leads to further health issues such as coronary artery disease, vision loss, and renal damage. Diabetes with insulin dependence also referred to as juvenile-onset diabetes, primarily affects young persons under the age of thirty and is brought on by immune-mediated, biological, and external factors [6]. This kind of diabetes results in the death of the beta cells that are available in the pancreas, which are in charge of producing sugar for the functioning of the body. On the contrary, Diabetes mellitus is not reliant on insulin [7]. The digestive tract generates a certain amount of insulin in people with diabetic condition, but not enough to satisfy the energy demands of their bodies. This kind of medical condition is brought on by getting older, being overweight, sedentary lifestyle, decreased ability to absorb glucose, a higher likelihood of mellitus in family members, having experienced pregnancy-related diabetes in the past, etc [8]. Diabetes during pregnancy is a condition that affects pregnant women who have never had diabetes before. Individuals with this kind of diabetes can manage it with consistent exercise and a healthy diet, but some people

[1]Professor, Department of Information Technology, Ponjesly College of Engineering, Nagercoil, Tamilnadu, India
[2]Associate Professor, Department of Computer Science and Engineering, RRASE College of Engineering, Chennai, Tamilnadu, India
[3]Associate Professor, Department of Computer Science and Engineering, Sri Eshwar College of Engineering, Coimbatore, Tamilnadu, India
itsanandmtech@gmail.com[1], ravikumarcsephd@gmail.com[2], saravanakumar.phdauc@gmail.com[3]
*Corresponding author: itsanandmtech@gmail.com

also require medication. After giving birth, this diabetes goes away, but in a small number of cases, it can later result diabetes mellitus [9].

The fourth kind of diabetes in humans is hereditary diabetes, which is caused by inherited deficiencies in the production of insulin. Pulmonary fibrosis related mellitus and glucocorticoid diabetes are caused by high doses of steroids [10]. Diabetes of any of the aforementioned types can lead to issues like coronary artery disease, stroke, lack of vision, urinary malfunctions, and venous disorders if the condition receives insufficient treatment. Because many diabetes cases are severe as a result of insufficient detection and management, prompt recognition of diabetes is crucial [11]. These days, data mining-based methods are frequently applied to the diabetes database to predict diabetes early.

A rapidly rising blood sugar level of more than 120 mg/dl or a 3-hour oral sugar bearing test with a blood sugar level of more than 180 mg/dl are the two methods used in the medical evaluation of those related to diabetes; however, the blood sugar minimum values used to identify diabetes may vary with ethnic background because various cultural groups have varying insulin resistance likelihood levels [12]. Thus, healthcare professionals are faced with the contentious challenge of establishing an insulin resistance limit for diagnosing diabetes regardless of the patient's cultural background, as well as figuring out whether there is a limit that can be specific without requiring an extended period of follow-up investigations to validate the medical condition [13].

The viability of various electronic therapies, including simulated reality instruction, platforms on the internet, and software that use innovation to improve healthcare, is being investigated. The present situation offers an advantageous possibility to create electronic devices that offer specific therapies for changing lifestyle habits in order to tackle the diabetes mellitus epidemic, given the swift advancement of genetic, interaction, and knowledge-based technology [14]. Large volumes of medical data have recently accumulated in systems connected to therapeutic health information, such as digital health records, research information systems, medical computer systems, and IT systems for hospitals. The importance of medical data has progressively emerged as a result of its ongoing increase.

A significant portion of healthcare that benefits patients is being contributed by machine learning approaches with data mining tools and methodologies in early diabetes prediction in order to assist medical professional to improve the efficiency to make appropriate decisions. To do this, a number of data mining methods have been suggested by investigators to use the diabetes dataset for early diabetes prediction [15]. It takes cognitive precision to arrive at an accurate diagnosis in a single therapeutic evaluation because multiple glucose measurements need to be performed both prior to and following a meal. However, a mathematical reduction of the screening procedure is possible. Many statistical efforts have been made in the past few decades, mostly focused on the use of data mining techniques in the field of diabetes with the goal of assisting medical professionals in reaching a timely accurate and significant diagnosis. These include support vector machine, multilayered perceptron (MLP), and artificial neural networks. Also, people can participate in customized diabetes level screenings to enhance changes in behavior owing to the constant advancement of diabetes monitoring and diagnostic instruments.

Typically, data mining tasks involving categorization are used to forecast diabetes using historical healthcare information from patient datasets [16]. A controlled learning task called categorization uses an algorithm created during the training stage to categorize a given dataset into ordained classification labels. Data which is augmented already from the diabetes database is supplied to the classifier in the first step as training data. In the second stage, the prediction tool creates a mathematical model by learning from the training data with characteristics and using that model to predict test data.

The main contributions of this work are,

1. To propose an ensemble approach using data mining techniques for efficient diabetes mellitus prediction.
2. To employ Averaged Fisher score and Kolmogorov-Smirnov score for choosing the important features for performing diabetes classification.
3. To implement Improved Chaotic Whale Optimization algorithm to optimally select the features for making reliable predictions.

The remainder of the paper is organized as follows. Section 2 discusses the recent research works on diabetes prediction employing the popular techniques in data mining. Section 3 presents the proposed approach using data mining methods with optimal feature selection using Improved Chaotic Whale optimization technique. Section 4 elaborates the results obtained on applying the proposed methodology to the PIMA Indian Diabetes dataset. Section 5 concludes the present research.

II. RELATED WORKS

Currently among the three major risks to the well-being of humans, diabetes is an international problem. Without proper care, diabetic patients risk developing cardiovascular issues, complications with the liver, damage to the neurological system, and other conditions that could have a major negative impact on their health. In this case, diabetes avoidance and prompt identification are essential. This section explores the recent works by researchers on diabetes prediction using computerized diagnostic methods.

The combination of data mining with probabilistic reasoning using fuzzy methods were applied to a diabetic diagnosis by the authors in [17]. They used a combination of categorization techniques with feature extraction to increase the precision of the algorithm. Finally, a probabilistic tree predictor produced estimates with an accuracy of 96.8%. To address the ambiguity in medical diagnosis data, a framework for fuzzy reasoning was developed, and extremely approximate language ideas were used to analyze fuzzy data. According to the authors in this research, several reasoning techniques have 96% accuracy when it comes to high accuracy and low complexity, which is crucial for the early detection and prevention of diabetes.

In [18], experts presented a concept for a medical records system that uses large-scale data analytics to monitor and study diabetes. It is built on the feature association and the elimination function in the Apache architecture. In order to forecast different types of diabetes and offer efficient therapies, they move data to various areas of the system and process it through various data units and hubs [19]. The platform uses performance indices like the sensitivity and the precision of mathematical assessment techniques to evaluate the system. The maximum index assessment value based on sensitivity reached a precision of 0.957, while the logistic regression algorithm achieved a precision of 0.974 in the system. This demonstrates even more how effective the system is compared to the current approaches [20].

The accuracy of the data mining methods with neural network approaches has kept getting better in studies on diabetes. In [21], the research scientists address data standardization, asymmetry, and pattern enhancement, respectively, using an adaptive variational with sparse coders and standard uniformity. After that, perceptron was employed to classify data with 93.5% accuracy. A synthetic reverse propagation enlarged axial slope neural network, was reported to attain 94.6% accuracy without the need for preliminary processing of data in [22], demonstrating an even greater improvement in accuracy. Another noteworthy example of neural network-based model performance may be found in the research of the authors in [23]. The researchers compared a sequential differentiator, regressor, and average value interpolation for the removal of absent values in their research. Then, a precision of 96% was obtained by classifying using artificial neural network. For choosing attributes and absence of value inference, authors in [24] used correlation between variables and average value interpolation. Using normalized ranges, they further adjusted the data and eliminated anomalies. Their neural network-based classification model, which included a number of concealed layers, had an accuracy of 93.8%.

A sophisticated hybrid neural network model attained 98.07% accuracy in [25]. The authors' claim that data perturbation and interpolation was carried out in the initial phases. In [26], attribute extraction and elimination of absent values were accomplished through the use of filter-based methods [27] of dimensionality reduction and mean value determination, respectively. After that, radial basis function networks were used to classify data with a 95.4% accuracy rate. It's noteworthy to observe that neural network-based methods and data mining approaches exhibit similar performance accuracies.

Following data preliminary processing of information, the group of researchers in [28] assesses the categorization outcomes of several data mining classifiers, including kernel-based support vector machines (SVM), longitudinal regression (LR), moderate slope enhancement model, and decision trees. At 95% accuracy, the longitudinal regression was found to be the most accurate model. Researchers in [29] utilized a variety of algorithms based on controlled, blended or collaborative learning to diagnose diabetes mellitus with a higher accuracy rate; nevertheless, the combined approach outperforms the other methods. By using a combined technique and a fuzzy voting classifier on the Pima-Diabetes, authors in [30] increased the preciseness of the diagnosis. In comparison to the other machine-learning algorithms, the fuzzy voting classifier achieved an accuracy of 96.8%, as per the results.

Presently, data mining algorithms are helpful in the diagnosis of diabetes illnesses, but the state-of-the-art research models are less accurate since they tended to concentrate on database alignment, preparatory processing methods, and other kinds of monitored and informal learning models. Finding a novel approach with algorithm integration that can combine the effectiveness of several data mining algorithms with high diagnosis precision is therefore necessary. In order to do this, an integrated data mining model—which combines four data mining techniques is put forth in this research.

III. PROPOSED METHODOLOGY

This section presents the techniques used in the proposed methodology for prediction of diabetes mellitus using data mining techniques. The proposed architecture is depicted in Figure 1.

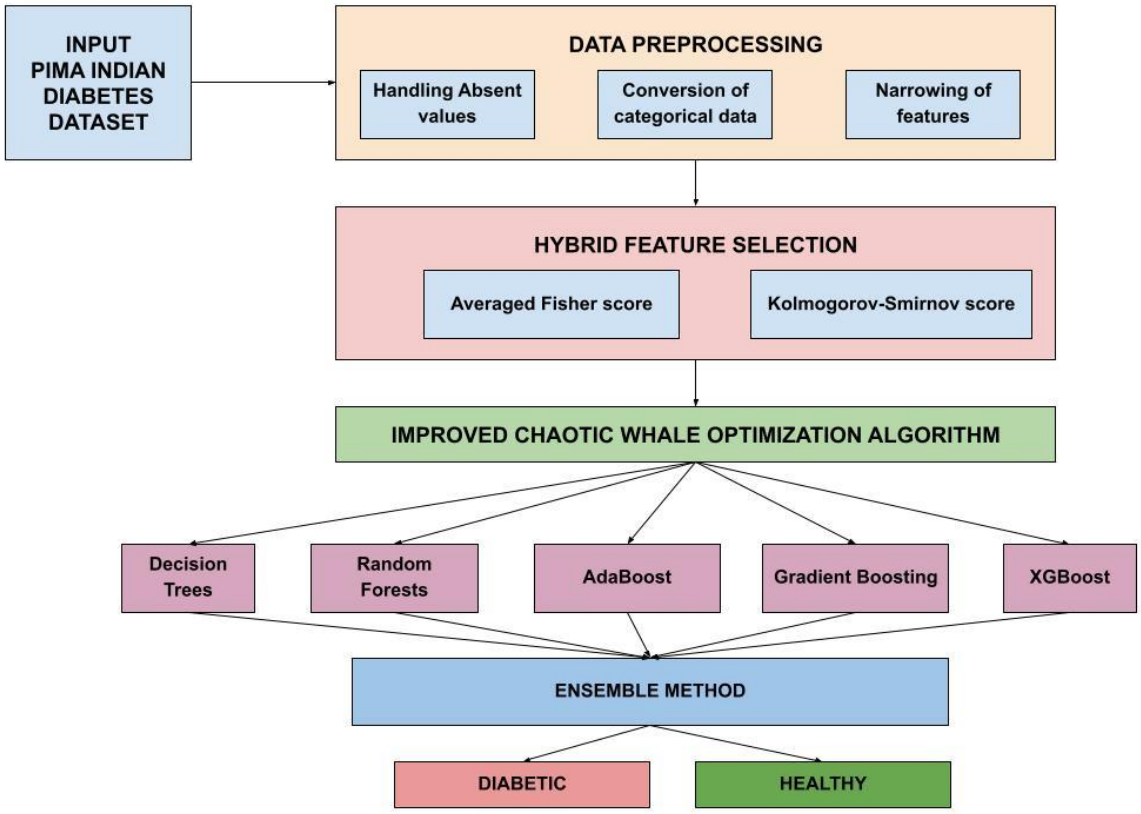

Fig. 1: Proposed Architecture

A. Data Preprocessing techniques

A crucial phase in this diabetes disease prediction problem is data preprocessing, which entails grouping and purifying the raw data to prepare it for model construction and activation. Actual observations often contain many inaccuracies and is insufficient, unreliable, and/or devoid of specific patterns or tendencies. To put it in another way, preliminary data processing is a data-mining approach that converts unprocessed data into a format that is easy to read and comprehend.

Handling Absent values

In data preprocessing, it is essential to accurately detect and treat values that are absent; otherwise, there is a high risk deriving inaccurate and misleading claims and results from the information available. The absence of a value can be treated in one of two ways: discard or update. Since the former is not entirely effective, it is advised that only the second method is adopted in situations where the dataset contains adequate sample sizes. Additionally, it is imperative to make sure that there is no distortion remaining after the data has been deleted. Because there are few values that are absent, thus the second method is implemented and updated the average in the corresponding position.

Conversion of categorical data

Since computational equations form the basis of data mining techniques, categorical data is information that is separated into discrete groups within a dataset. Since these equations only involve values in numbers, it is intuitive that retaining categorical data in the equation will lead to issues, hence it needs to be converted into value formats in integers.

Narrowing of features

Feature narrowing is a data preparation technique used to organize the feature variables of the dataset underneath an identified range. Stated differently, feature scaling helps us compare the independent variables on comparable principles by reducing their range. One of the approach for feature narrowing is the process in which minimum (D_{min}) and maximum (D_{max}) values of the features are considered as represented in equation (1),

$$D' = \frac{D - D_{min}}{D_{max} - D_{min}} \quad (1)$$

The second approach involves the incorporation of the average (θ) as well as the standard deviation (ρ) values. In the case, it is assured that the values of the and are zero and one accordingly.

$$D' = \frac{D - \theta}{\rho} \quad (2)$$

B. Hybrid Feature Selection Methods

The objective of diabetes prediction using data mining technique heavily relies on feature selection methods to cut down the dimensionality issues. Data analysis, computer vision, artificial intelligence, and machine learning all heavily rely on feature selection (Fs). The goal of feature selection is to eliminate unnecessary information and choose only the most useful characteristics. Since the choice of features can reduce the number of characteristics used for learning process in classification models, here the main goal is to use techniques for feature selection to order and choose a particular group of pertinent characteristics based on their level of relevance, inclination, or impact as defined by an application. Additionally, reducing the number of dimensions lessens the effects of the high dimensions on fitting too closely and learning time, enhances model development efficiency, and boosts interpretation of data.

Averaged Fisher score

The Fisher score is one of the effective univariate feature selection methods which assesses the most contributing characteristics in the dataset. In this research, an averaged value is included in the fisher score to enhance the linear differentiation of the characteristics. The mathematical representation of this method is denoted in equation (3),

$$F_x = \frac{\sum_{y=1}^{h} A_y.(s_y - \bar{s})^2}{\sum_{y=1}^{h} A_y\, \delta_y^2} \quad (3)$$

In the above equation, F_x is the value assigned to each feature, h denotes the count of the target classifications, A_y denotes the count of the examples in the dataset, $\bar{s}$ denotes the average value of the features, δ_y^2 denotes the disparity in a specific feature.

Kolmogorov-Smirnov score

This score is computed by determining the correlation between the arbitrary values by estimating the highest value of the variation in the variables. The repetition of characteristics is performed by first segregating the features into N groups. The repeated features in each group are computed. Based on these values the variation in the feature values is determined. The highest possible variation in the repeated values is calculated by using the equation (4),

$$\beta = \sqrt{k/2}\,(\max|X_a - X_a'|) \quad (4)$$

In the above equation, X_a and X_a' are the two arbitrary features taken for finding the inherent association.

C. Improved Chaotic Whale Optimization Algorithm

Whale optimization algorithm is based on the activities of whales that exists in groups. One important aspect about the enormous whales is their ability to perform hunting. The special hunting activity of humpback whales is termed as bubble-net foraging technique. Initially, the fishes in tiny sizes that are in near distance is hunted. These whales swirl in motion to travel in search of food. Whales begin by identifying the position of the target

which is in nearest possible distance and will encompass those targets. This activity of the whales is represented in equations (5) and (6),

$$\vec{N} = \left|\vec{M}.\vec{A} * (k) - \vec{A}(k)\right| \quad (5)$$

$$\vec{A}(k+1) = \vec{A} * (k) - \vec{S}.\vec{N} \quad (6)$$

In the above equations k denotes the present run of the algorithm, $\vec{S}$ *and* $\vec{M}$ are used as constant values, the location of the target is denoted by $\vec{A}$. The values represented by $\vec{S}$ *and* $\vec{M}$ are as shown in equations (7) and (8),

$$\vec{S} = 2\vec{s}\,.\vec{c} - \vec{s} \quad (7)$$

$$\vec{M} = 2\,.\vec{c} \quad (8)$$

In the above equation, $\vec{c}$ is an arbitrary value which ranges between zero and one. The swirling motion of the whales is represented in equation (9),

$$\vec{A}(k+1) = \overrightarrow{N'}\,.e^{rt}.\cos(2\pi t) + \vec{A} * (k) \quad (9)$$

Where r and t are arbitrary values and t takes values between -1 and 1. The whales usually swirl around the target in a circular fashion and this dual activity is represented together in equation (10),

$$\vec{A}(k+1) = \begin{cases} \vec{A} * (k) - \vec{S}\,.\vec{N} & if\ av < 0.5 \\ \overrightarrow{N'}\,.e^{rt}.\cos(2\pi t) + \vec{A} * (k) & if\ av \geq 0.5 \end{cases} \quad (10)$$

The value av in the above equation is chosen randomly and lies between zero and one. The scout for attaining the target is represented in equations (11) and (12),

$$\vec{N} = \left|\vec{M}.\overrightarrow{A_{rand}} - \vec{A}\right| \quad (11)$$

$$\vec{A}(k+1) = \overrightarrow{A_{rand}} - \vec{S}\,.\vec{N} \quad (12)$$

In this improved version of the whale optimization algorithm, chaotic nature is included to ensure that the solutions do not fall in the local optimum level. The chaotic vectors are denoted in general as shown in equation (13),

$$h_x^{(i+1)} = g\left(h_x^{(i)}\right)\ \ where\ x = 1,2,3,\ldots,k \quad (13)$$

Chebyshev chaotic map is adopted in this present research to facilitate the optimal selection of features from the dataset for making better predictions. The mathematical representation of the Chebyshev chaotic map is shown in equation (14),

$$h_{x+1} = \cos\left(x\ cos^{-1}(h_x)\right) \quad (14)$$

The detailed steps involved in the proposed optimization algorithm is presented in Figure 2.

Algorithm: Improved Chaotic Whale Optimization algorithm
Input: Initial population of whales A_k $(k = 1, 2\ 3, \ldots, N)$
Output: A^*, global optimal solution
Step 1: Compute the survival rate of every whale
Step 2: A^* = local optimal solution
Step 3: while (k< maxIter)
Step 4: for each solution
Step 5: Modify the values of s, S, N, t and av
Step 6: if (av < 0.5)
Step 7: if (|S| < 1)
Step 8: Compute location of current target using equation (5)
Step 9: else if (|S| ≥ 1)
Step 10: Pick arbitrary target as A_{rand} and modify the target location using equation (12)
Step 11: end if
Step 12: else if (av ≥ 0.5)
Step 13: Modify the target location using equation (9)
Step 14: end if
Step 15: end for
Step 16: Determine the targets that traverse far away from target boundary
Step 17: Calculate the updated survival rates of each target
Step 18: Update A^* using equation (14) to find the global optimal solution
Step 19: k = k+1
Step 20: end while

Step 21: return A^*

Fig. 2: Improved Chaotic Whale Optimization algorithm

D. Data mining techniques for Diabetes Prediction

To increase the consistency and forecasting capability of the model, different models were integrated into an ensemble approach. In cases where this combined method is employed instead of just one model, a better prediction performance is possible. The group discovers methods for merging data mining techniques to make predictions. While boosting methods lessens discrimination and stacking techniques enhances efficiency, bagging methods is used to lessen volatility. While some models perform better when predicting one component of the data, others perform better when estimating another. The ensemble model learns multiple basic models and integrates their outputs to arrive at the ultimate outcome, as opposed to learning an individual complicated model. As the effectiveness of individual classifiers is surpassed by the aggregate forecast produced by ensemble learning, the resultant forecast will be more accurate. The various data mining techniques implemented in this research for diabetes prediction are presented in this section.

Decision tree Classifier

This method of resolving the categorization issue is centered around rules. From the characteristic set, a decision tree is constructed by applying the sequence in the set with if-else conditions. In order to formulate the if-else sequence set any combination of the entropy, information gain or the Gini index are used. The most preferred technique is the the Gini index method. These techniques are applied to choose the innermost node and divide the instances for the remainder of the tree's level.

Random Forest Classifier

A categorization method based on trees is called Random Forest. As the name suggests, this method produces a forest with an abundance of trees. In this method, a set of decision trees is produced by randomly selecting certain portions of the training data. It runs the process again using a variety of sample combinations until reaching a final decision based on the vote of the greatest number. The Random Forest algorithm is helpful in handling values that are unavailable, despite the possibility of overfitting. Using the right factor modifications can help prevent excessive fitting of data.

AdaBoost Classifier

It works based on partial evaluation that first adapts a classifier on the initial set of data, after which it adapts additional repetitions of the classifier on the associated data while rearranging an immense number of poorly organized samples to make subsequent classifiers focus more on problematic situations.

Gradient Boosting Classifier

The core fundamental parts of the Gradient Boosting method are a function that suffers from loss that must be customized, an ineffective learner for forecasting, and a model to combine the ineffective learners in order reduce the function that defines the loss. The models are trained repeatedly, simultaneously, and progressively by the algorithm. Enhancing a customized loss function, which is primarily particular to the objectives, is one of main benefits of gradient boosting.

XGBoost Classifier

The XGBoost algorithm is a sophisticated adaptation of the Gradient Boosting methodology. In order to create a " robust " learner, it integrates all of the forecasts made by a group of "weakened" learners. XGBoost aims to keep the mathematical process from fitting too tightly while also preventing it. XGBoost streamlines the goal functions, enabling the combination of regularization and predictive terms while preserving the fastest possible processing rate. To address the shortcomings of a fragile learner, a second model is fitted to these remaining data after the first learner is adapted to the entire set of input data.

Voting mechanism

A voting predictor is a forecasting model that uses numerous learners. It is applied in situations when it is necessary to make predictions based on the most frequently encountered predictor among those that have been deployed. Voting comes in two flavors: gentle and stern. Every classifier is taken into consideration when voting for a class in a majority vote, also known as stern voting. When using gentle voting mechanism, every predictor

is associated with a likelihood score that indicates the chance that a certain sample of data will be placed in an identified target class.

IV. RESULTS AND DISCUSSION

A. *Dataset description*

The dataset used in this current research is the PIMA Indian Diabetes dataset which can be accessed through the provided link. Based on specific pathological metrics included in the collection, the dataset aims to accurately assess the presence or absence of diabetes in a patient. These examples were chosen from a bigger database under a number of restrictions.
https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database
Specifically, all of the patients in this facility are Pima Indian women who are at least twenty-one years old. There are a total of nine features which are used to predict between two target classes.

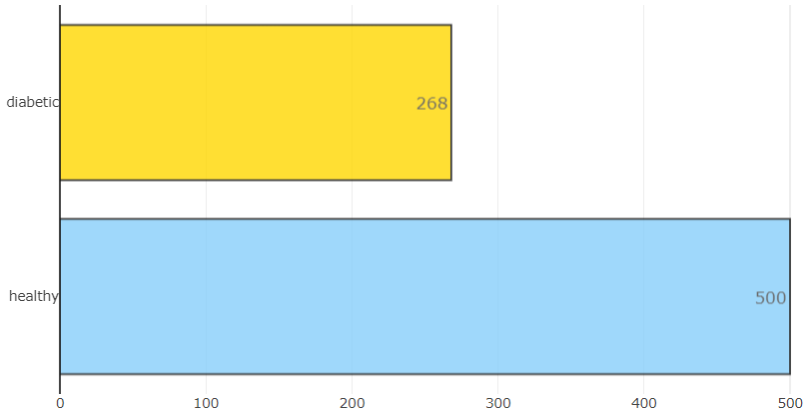

Fig. 3: Number of target classes

The total number of records in the dataset counts to 768, out of which 268 records correspond to individuals affected by diabetes and 500 for normal individuals. The distribution of records in the PIMA dataset for the categorizations diabetic and healthy are presented in Figure 3 and 4.

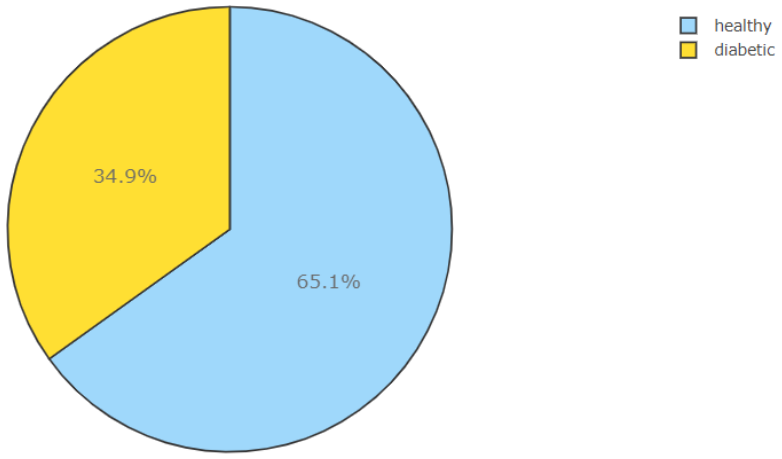

Fig. 4: Distribution of target classes

The details of the features included in the dataset for diabetes prediction are number of times individuals were conceived, the individual's glucose as well as insulin level, hypertension level, thickness of the skin layers, body mass index, family history of diabetes and age factor. The number of values that are available for the features is presented in Figure 5.

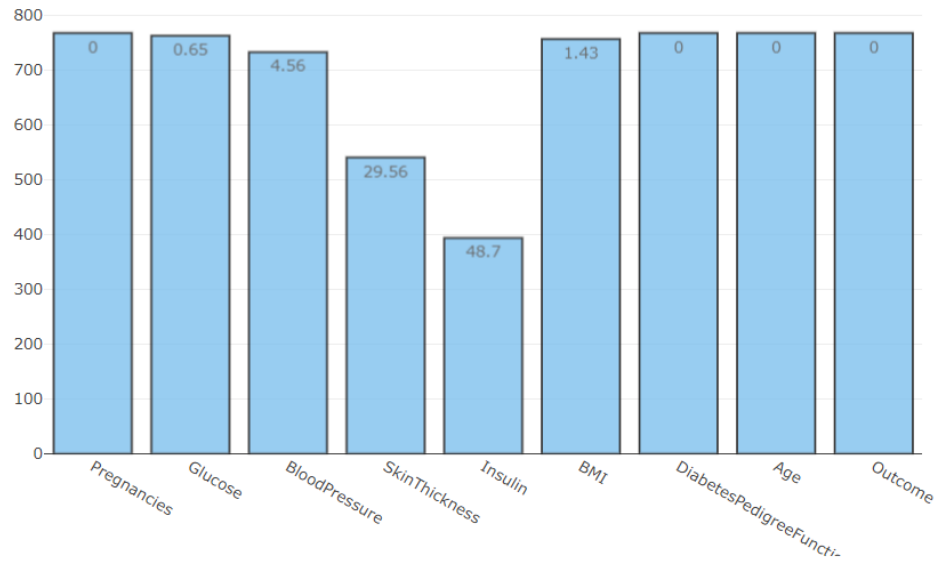

Fig. 5: NA values in dataset

The inherent association between the features that are present in the dataset can be found using the correlation matrix. The association between the nine features in the dataset is presented in Figure 6 and depicts how each of these features are related to one another on a scale between 0 and 1.

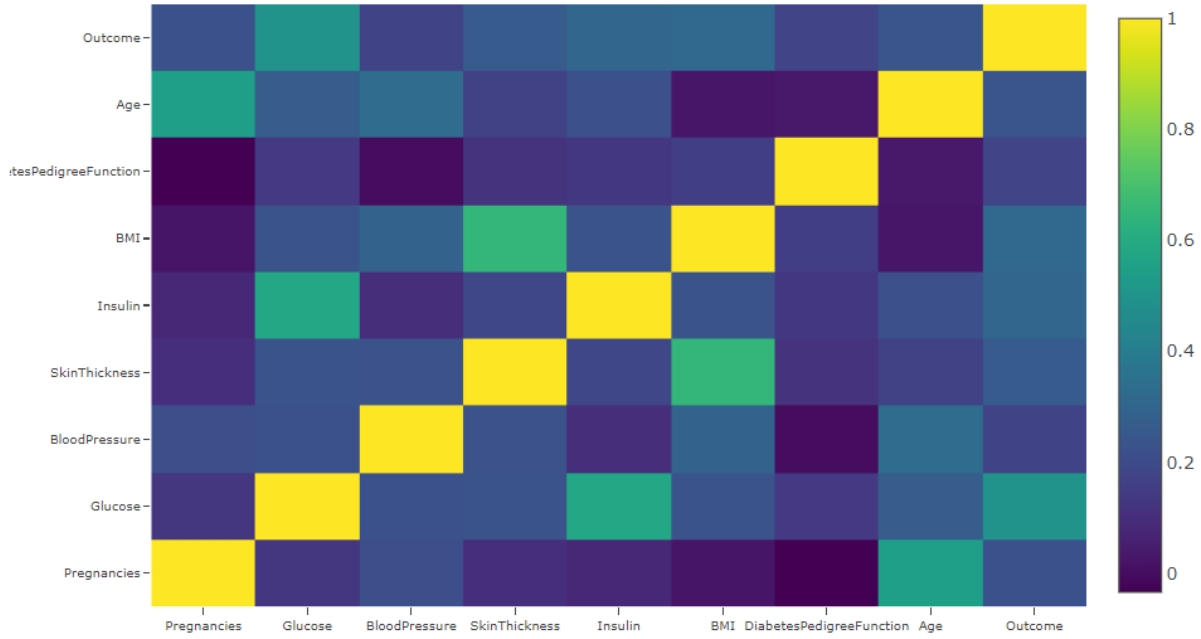

Fig. 6: Correlation matrix

B. *Experimental evaluation*

The performance of the proposed ensemble approach is assessed using various metrics such as accuracy, precision, recall and F1 score. Each of these values are computed for all the methods in the proposed approach and the results are obtained. The data mining techniques in the proposed approach are applied to the PIMA Indian Diabetes Dataset and the obtained results are presented in Table I and depicted for comparison in Figure 7.

Table. 1: Performance Evaluation of Proposed Approach

Techniques	Accuracy	Precision	Recall	F1 score
Decision Trees	90.7	89.5	89.8	90.2
Random Forests	91.8	90.5	89.9	91.3
AdaBoost	92.6	91.3	91.5	92.2
Gradient Boosting	95.7	94.5	94.7	95.1
XGBoost	98.8	97.1	97.3	98.4

It can be observed that the XGBoost method produced highest accuracy among the classifiers as 98.8%. The other methods such as Decision Trees, Random Forests, AdaBoost and Gradient Boosting produced 90.7%, 91.8%, 92.6% and 95.7% respectively.

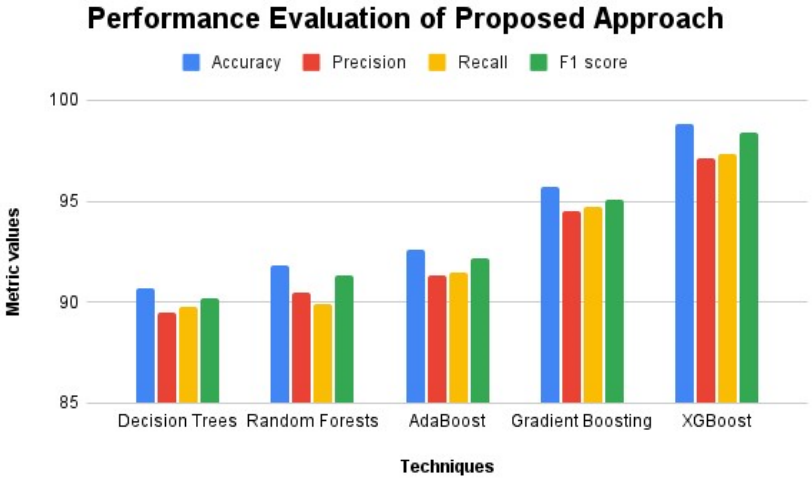

Fig.7: Performance Evaluation of Proposed Approach

Further, the proposed approach is also compared against the other conventional data mining techniques such as Logistic Regression (LR), Support Vector Machine (SVM), Naïve Bayes (NB) and k-Nearest Neighbor (kNN) to demonstrate the performance superiority. The outcomes are tabulated in Table II and presented graphically in Figure 8. It was inferred that LR methods produced an accuracy, precision, recall and F1 score values of 85.4%, 84.1%, 84.3% and 84.9% correspondingly. It is the lowest among the conventional methods compared.

Table. 2: Performance Comparison with conventional methods

Techniques	Accuracy	Precision	Recall	F1 score
Logistic Regression	85.4	84.1	84.3	84.9
Support Vector Machine	87.2	85.7	86.2	86.8
Naïve Bayes	89.6	87.5	87.9	88.6
K-Nearest Neighbor	93.5	91.8	92.2	92.8
Proposed Approach	98.8	97.1	97.3	98.4

Among the considered approaches, it is observed kNN produced 93.5% accuracy, 91.8% precision, 92.2% recall and 92.8% F1-score. Though the performance exhibited by the kNN model is better than the other conventional approaches, it is relatively low when compared with the proposed approach which produces 98.8% accurate prediction for diabetes classification.

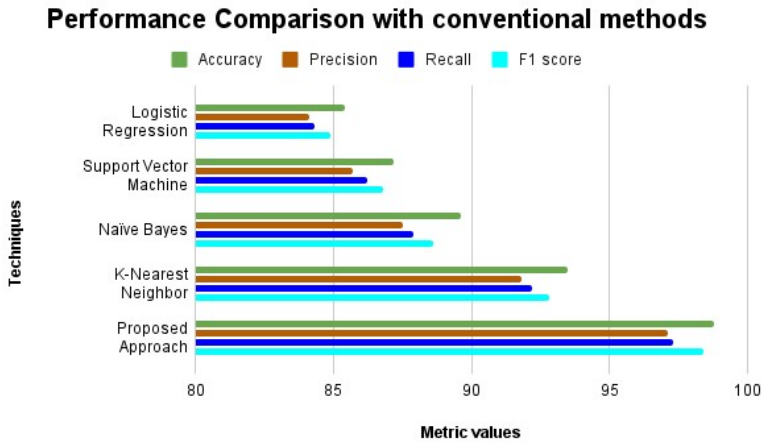

Fig. 8: Performance Comparison with conventional methods

Additionally, the proposed approach is also compared against the existing methods in the literature with respect to the ensemble methods. The authors in [22] suggested an approach which combines the results of Random Forests (RF) with Multi-layer Perceptron (MLP) and k-Nearest Neighbor algorithms (kNN).

Table. 3: Performance Comparison of Existing and Proposed methods

Ensemble Techniques	Accuracy	Precision	Recall	F1 score
RF+MLP+kNN [22]	88.4	87.1	87.3	87.9
SVM+NB+DT+RF [24]	89.5	88.2	88.4	89.1
RF+SVM+AdaBoost+GradientBoost[25]	92.6	91.4	91.7	91.9
kNN+SVM+DT+GradientBoost [27]	96.4	95.3	95.7	96.2
Proposed Approach	98.8	97.1	97.3	98.4

The research recommended in [24] is an amalgamation of SVM, NB, DT and RF methods. The investigators in [25] uses the data mining techniques such as RF, SVM, AdaBoost and GradientBoost. In another research [27], the algorithms such as kNN, SVM, DT and GradientBoost have been combined. The performance exhibited by these methods are assessed and presented in Table III and a comparative graph is shown in Figure 9. The performance evaluation results prove the supremacy of the proposed approach in predicting diabetes mellitus with improved efficiency and reliability.

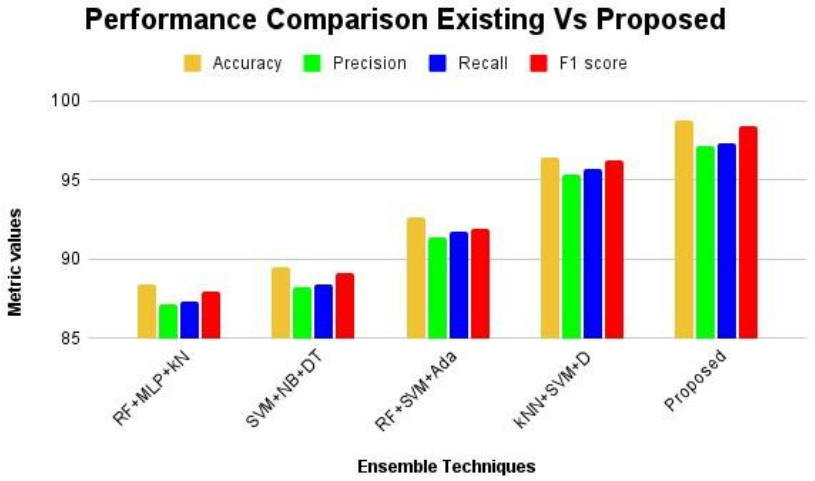

Fig. 9: Performance Comparison Existing Vs Proposed

V. CONCLUSION

Identifying diseases at their earliest stages is one of the most important medical practices. Across all age groups, individuals with diabetes levels are rising globally in recent times, and there is currently no therapy that can reverse diabetes. Diabetes mellitus is a significant global issue due to its age-neutral nature. Early identification with confirmed diabetes patients lowers medical costs, mortality, and threat to patients. In this research, investigations are performed using the Pima Indian Diabetes dataset and achieved 98.8% accuracy in categorizing individuals as diabetic or not. Five different data mining techniques are combined to make the final predictions. The optimized features are obtained using Improved Chaotic Whale Optimization algorithm. The suggested system using the ensemble approach on data mining techniques has the potential to forecast further chronic illnesses in the future. The work on automating the analysis of diabetes can be expanded upon and strengthened with the use of artificial intelligence techniques.

REFERENCES

[1] G. Pradhan, R. Pradhan, and B. Khandelwal, ‘‘A study on various machine learning algorithms used for prediction of diabetes mellitus,’’ in Soft Computing Techniques and Applications (Advances in Intelligent Systems and Computing), vol. 1248. London, U.K.: Springer, 2021, pp. 553–561, doi: 10.1007/978-981-15-7394-1_50.

[2] Gandin I, Saccani S, Coser A, Scagnetto A, Cappelletto C, Candido R, et al. Deep-learning-based prognostic modeling for incident heart failure in patients with diabetes using electronic health records: a retrospective cohort study. PLoS ONE. 2023;18(2): e0281878.

[3] B. Jain, N. Ranawat, P. Chittora, P. Chakrabarti, and S. Poddar, ''A machine learning perspective: To analyze diabetes,'' Mater. Today: Proc., pp. 1–5, Feb. 2021, doi: 10.1016/J.MATPR.2020.12.445.
[4] S. Kumari, D. Kumar, and M. Mittal, ''An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier,'' Int. J. Cogn. Comput. Eng., vol. 2, pp. 40–46, Jun. 2021, doi:10.1016/j.ijcce.2021.01.001.
[5] Joseph JJ, Deedwania P, Acharya T, Aguilar D, Bhatt DL, Chyun DA, et al. Comprehensive management of cardiovascular risk factors for adults with type 2 diabetes: a scientifc statement from the American Heart Association. Circulation. 2022;145(9):e722–59.
[6] Dinh A, Miertschin S, Young A, Mohanty SD. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Med Inform Decis Mak. 2019;19(1):1–15.
[7] S. Saru and S. Subashree. Analysis and Prediction of Diabetes Using Machine Learning. Accessed: Oct. 22, 2022. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3368308
[8] Kulkarni AR, Patel AA, Pipal KV, Jaiswal SG, Jaisinghani MT, Thulkar V, et al. Machine-learning algorithm to non-invasively detect diabetes and pre-diabetes from electrocardiogram. BMJ Innov. 2023. https://doi.org/10.1136/bmjinnov-2021-000759.
[9] Ravaut M, Harish V, Sadeghi H, Leung KK, Volkovs M, Kornas K, et al. Development and validation of a machine learning model using administrative health data to predict onset of type 2 diabetes. JAMA Netw Open. 2021;4(5): e2111315.
[10] N. Sneha, T. Gangil, Analysis of diabetes mellitus for early prediction using optimal features selection, J. Big Data 6 (1) (2019) http://dx.doi.org/10.1186/s40537-019-0175-6.
[11] V. Chang, V.R. Bhavani, A.Q. Xu, M. Hossain, an artificial intelligence model for heart disease detection using machine learning algorithms, Healthc. Anal. 2 (2021) (2022) 100016, http://dx.doi.org/10.1016/j.health.2022.100016.
[12] D. J. D. M.D., Diabetes mellitus diabetes mellitus, Ferri's Clin. Advis. 2020 512 (January) (2020) 432–441, http://dx.doi.org/10.1016/B978-0-323-67254-2.00255-2.
[13] L. Zhang, Y. Wang, M. Niu, C. Wang, Z. Wang, Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: The Henan rural cohort study, Sci. Rep. 10 (1) (2020) 1–10, http://dx.doi.org/10.1038/s41598-020-61123-x.
[14] K. Hasan, A. Alam, D. Das, E.H. Senior, Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers, VOl. X, 2020, pp. 1–19, http://dx.doi. org/10.1109/ACCESS.2020.2989857
[15] B. Davazdahemami, H.M. Zolbanin, D. Delen, An explanatory analytics framework for early detection of chronic risk factors in pandemics, Healthc. Anal. 2 (January) (2022) 100020, http://dx.doi.org/10.1016/j.health.2022.100020.
[16] M.F. Ijaz, G. Alfian, M. Syafrudin, J. Rhee, Hybrid prediction model for type 2 diabetes and hypertension using DBSCAN-based outlier detection, synthetic minority over sampling technique (SMOTE), and random forest, Appl. Sci. 8 (8) (2018) http://dx.doi.org/10.3390/app8081325.
[17] N.P. Tigga, S. Garg, Prediction of type 2 diabetes using machine learning classification methods, Procedia Comput. Sci. 167 (2019) (2020) 706–716, http://dx.doi.org/10.1016/j.procs.2020.03.336.
[18] S.M. Ganie, M.B. Malik, T. Arif, Machine learning techniques for diagnosis of type 2 diabetes using lifestyle data, in: International Conference on Innovative Computing and Communications, in: Advances in Intelligent Systems and Computing, vol. 1394, Springer, Singapore, 2022, pp. 487–497.
[19] M.K. Hasan, M.A. Alam, D. Das, E. Hossain, M. Hasan, Diabetes prediction using ensembling of different machine learning classifiers, IEEE Access 8 (2020) 76516–76531, http://dx.doi.org/10.1109/ACCESS.2020.2989857.
[20] B.A. Artha Imjr, N.K. Dharmawan, U.W. Pande, K.A. Triyana, P.A. Mahariski, J. Yuwono, V. Bhargah, I.P.Y. Prabawa, I.B.A.P. Manuaba, I.K. Rina, High level of individual lipid profile and lipid ratio as a predictive marker of poor glycemic control in type-2 diabetes mellitus, Vasc. Health Risk Manag. 5 (2019) 149–157, https://doi.org/10.2147/VHRM.S209830, doi: 10.2147/VHRM.S209830
[21] N.P. Maratni, et al., Association of apolipoprotein E gene polymorphism with lipid profile and ischemic stroke risk in type 2 diabetes mellitus patients, J. Nutr. Metabol. 20 (21) (2021) 1–16, https://doi.org/10.1155/2021/5527736.
[22] D. Sisodia, D.S. Sisodia, Prediction of diabetes using classification algorithms, Procedia Comput. Sci. 132 (2018) 1578–1585.
[23] Z.S. Ageed, S.R. Zeebaree, M.M. Sadeeq, S.F. Kak, H.S. Yahia, M.R. Mahmood, I. M. Ibrahim, Comprehensive survey of big data mining approaches in cloud systems, Qubahan Acad. J. 1 (2) (2021) 29–38.
[24] L. Ismail, H. Materwala, M. Tayefi, P. Ngo, A.P. Karduck, Type 2 diabetes with artificial intelligence machine learning: methods and evaluation, Arch. Comput. Methods Eng. 29 (1) (2022) 313–333, https://doi.org/10.1007/s11831-021-09582-x.
[25] M. Gollapalli, A. Alansari, H. Alkhorasani, M. Alsubaii, R. Sakloua, R. Alzahrani, W. Albaker, A novel stacking ensemble for detecting three types of diabetes mellitus using a Saudi Arabian dataset: pre-diabetes, T1DM, and T2DM, Comput. Biol. Med. 147 (2022), 105757.

[26] H. Ikegami, Y. Hiromine, S. Noso, Insulin-dependent diabetes mellitus in older adults: current status and future prospects, Geriatr. Gerontol. Int. 22 (8) (2022) 549–553
[27] R. Krishnamoorthi, S. Joshi, H.Z. Almarzouki, P.K. Shukla, A. Rizwan, C. Kalpana, B. Tiwari, A novel diabetes healthcare disease prediction framework using machine learning techniques, J. Healthc. Eng. 2022 (2022).
[28] H. Wu, S. Yang, Z. Huang, J. He, X. Wang, Type 2 diabetes mellitus prediction model based on data mining, Inform. Med. Unlocked 10 (2018) 100–107.
[29] F.A. Khan, K. Zeb, M. Al-Rakhami, A. Derhab, S.A.C. Bukhari, Detection and prediction of diabetes using data mining: a comprehensive review, IEEE Access 9 (2021) 43711–43735.
[30] C. Fiarni, E.M. Sipayung, S. Maemunah, Analysis and prediction of diabetes complication disease using data mining algorithm, Procedia Comput. Sci. 161 (2019) 449–457.