Chapter 10
Early Detection and Analysis of Diabetics and Non-diabetics Using Machine Learning

Vikas Somani, Awanit Kumar, and Geetanjali Amarawat

10.1 Introduction

Multiple healthcare opportunities are generated because machine learning models have advanced predictive analytics potential. Machine learning models can also predict chronic diseases such as heart infections and intestinal disorders. There will also be several future machine learning models for predicting non-communicable conditions and increasing healthcare benefits. Researchers are working on machine learning models to predict particular diseases in patients at an early stage and to produce successful methods of disease prevention. This will also minimize patient hospitalization. This transition would be of great benefit to health organizations [1]. Healthcare systems that use advanced computing methods are the most studied field of healthcare research. The allied fields are moving toward more ready-to-to-assemble systems, as seen above. Patients with diabetes cannot generate insulin, resulting in hyperglycemia, a medical measure for increased sugar in the body. In other words, the body cannot repress the hormone insulin release. This leads to abnormal carbohydrate metabolism and high blood glucose levels. Because of the above causes, early detection of diabetes is very critical. Many people worldwide have diabetes, and this is increasing daily. This condition can involve multiple essential organs so that the medical equipment can heal it early in the diagnosis. The number of diabetic patients increasingly requires unnecessarily relevant medical information. Researchers need to create a system that stores,

V. Somani (✉) · A. Kumar
Sangam University, Bhilwara, Rajasthan, India
e-mail: vikas.somani@sangamuniversity.ac.in; awanit.kumar@sangamuniversity.ac.in

G. Amarawat
Swaminarayan University, Kalol, Gujarat, India
e-mail: deanengg@swaminarayanuniversity.ac.in

F. J. J. Joseph et al. (eds.), *Computational Intelligence for Clinical Diagnosis*,
EAI/Springer Innovations in Communication and Computing,
https://doi.org/10.1007/978-3-031-23683-9_10

updates, and analyses this diabetes knowledge and also recognizes threats using today's expanding technologies. [2] Diabetes is one of the diseases that spread over the world like epidemics. It was proved that every generation, including infants, teenagers, youth and the ages, suffers. Pro-long impacts on organs, such as hepatitis, kidneys, heart, stomach, and death, can be worse. Retinopathy and neuropathy conditions are also interlinked. Diabetes mostly forms type 1 and type 2 [3].

Type-1 Diabetes: In this case, the liver produces no insulin whatsoever. Insulin is a hormone necessary for the use of blood glucose in the body. Blood sugar will increase and lead to type 1 diabetes without insulin in the bloodstream. It is common for children and adolescents. It happens primarily due to genetic disorders. It is also referred to as a youth disease. Its frequent signs are frequent urination, loss of weight, increased hunger, blurred vision, and nerve problems. It can be treated with insulin.

Type-2 Diabetes: In people over 40 years, it is usually a long-term metabolic condition. High blood sugar, resistance to insulin and high insulin are apparent. Fatness and lack of exercise are the main factors. This bad lifestyle will lead to blood glucose storage and diabetes. Only 90% of people with type 2 diabetes are affected. Metformin is administered to treat insulin resistance.

Diabetic Neuropathy: These are the nerve abnormalities acquired over time in diabetic patients. They also happen in the hands and feet. The typical symptoms are pressure, numbness, pinching, and loss of a hand, foot, arms, etc.

Diabetic Retinopathy: It is a diabetic condition which leads to continuous blindness of the eye. At first, there is no noticeable symptom, and symptoms gradually occur. The second stage is the formation of blood vessels on the back of your eyes which can lead to agile bleeding.

10.2 Background and Key Issues

An analysis of computer and artificial intelligence (AI) techniques for the early detection of diabetes is given in Table 10.1.

The diabetes deduction literature survey reveals that a single approach to diabetes detection is not very sophisticated in early stages of diabetes. A hybrid solution with classificatory as primary elements, vector support machines analysis of genetic algorithms can improve the efficiency of artificial neural networks since this technique helps to minimize data noise by extracting features and then using learning methods to recognize hidden patterns, providing more accurate output.

Table 10.1 Survey for early diabetes detection using different machine learning and artificial intelligence techniques

Author	Central idea	Pros	Cons
Mohammed Imran et al.	Diabetic retinopathy (DR) detection using extended fuzzy logic 2. The OWE calculation is based on damage to the eyes Retina [4].	It enables the identification and calculation of retinal damage	Complex method and time taken
Mani Butwall Shraddha Kumar et al.	The approach to data mining for diabetes prevention is based on the random forest classification [5]	Classifiers are a good way of handling massive datasets	As compared to hybrid, a single classification the approach is not very successful
Kamadi V.S. R.P.Varma et al.	It uses the key component analysis and the modified Gini Index SLIQ Decision Tree Algorithm [6]	With fugitive SLIQ, sharp judgment limits can be overcome	Precision can be further improved by fluid membership
Kiarash Zahirnia Mehdi et al.	This document presents and compares various cost-sensitive methods of learning for type 2 diabetes diagnosis [7]	Cost-sensitive approach to resource use is successful	Assumptions in data sets, matrices are used to produce results
Kemal polat et al.	Combining c-means and SVM fuzzy is used for the dataset diabetes prediction [8]	Fuzzy C-means better classify data set with the membership feature	Real-time data is noisy, so that it can be used for processing
Nawaz Mohamudally et al.	Diabetes is shown in this study C4.5, neural network, K-means [9]	It is a successful approach because of the use of a hybrid system	Prediction, description, visualization demands enormous effort
Mostafa Fathi Ganji et al.	ACO is used to derive a set of FADD diagnostic rules [10]	FADD is an excellent solution to diabetes detection.	A single deduction method must be associated with others

10.2.1 Machine Learning Models for Diabetes Detection

The following techniques are used to detect diabetes in the early stages using machine learning techniques and discuss the advantages and disadvantages of the methods in this section.

10.2.2 SVM

SVM can be used for regression and classification applications but is better known for classification applications. This procedure, also known as the dimension plane, plots each point inside a data object into three-dimensional space, where n is the

number of data attributes. Distinctions between classes serve as the basis for categorization; these classes can be thought of as tiers of information. Since this is a controlled learning method, data sets are prepared in advance for use [11]. It displays datasets in space as cloud points. The objective is to create a hyperplane that divides data into different categories. The hyperplane divides data collection into groups for data collection and classification. The overall margin of this hyperplane should be the other categories. However, advanced kernel configuration techniques are used if the data categories are wide-ranging.

10.2.2.1 Advantages

- SVM is used to efficiently identify diabetes data by assigning hyperplane information to various categories.
- It removes the fitness of the samples.

10.2.2.2 Disadvantages

- For massive datasets, SVM cannot be used.
- SVM is running slowly.

10.2.3 Fuzzy C-Means

It is an extension to a K-mean clustering algorithm that aims to form clusters and then discover the centers of the cluster, and that cluster with a minimum distance to its centroids is allocated the incoming dataset. However, often there is very little space for new data packages to fall for more than one cluster [12]. The fluid C-means that cluster algorithm prevented this because it uses a fluid partition that is the member variable. The findings are also more precise.

10.2.3.1 Advantages

- In this respect, participation in the fuzzy logic of the membership function helps produce better classification results.
- The learning method is unsupervised so that the effects are more real-time.

10.2.3.2 Disadvantages

- It takes time to calculate.
- It is more likely to be a misconception in the early stages.

10.2.4 PCA (Principal Component Analysis)

PCA is a statistical model for classifying datasets to have the highest correlation in data collection [13]. This aims to create an orthogonal plane to organize data along with this plane. Another plane, well known for its second relationship between the datasets, is perpendicular to that plane. It supports function extraction and measures the main component using Eigen values and Eigen vectors.

10.2.4.1 Advantages

- It helps to reduce the dimension and preserve the alteration between datasets.
- It helps to reduce noise by selecting the highest variance dataset.

10.2.4.2 Disadvantages

- Eigenvalues and covariance matrices are difficult to quantify.
- PCA alone does not provide excellent results when it comes to diabetes detection.

10.2.5 Naive Bayes Classifier

Bayes theorem is a controlled learning technique. It is the algorithm family, assuming that the value of one function is independent (native). It considers the conditional chance of determining the likelihood of an occurrence if any of the events have already occurred. It is used for diabetes diagnosis and diabetic retinopathy detection. The Generative Learning Model can also be referred to as his classificatory. The classification is based on Baye's theorem, which implies separate predictors. In short, this classifier assumes that certain features are not connected to any other function in a class. Success and failure are equally possible if there is a dependence on the characteristics of each other. This tool is handy and easy for databases with large volumes of data [14].

10.2.5.1 Advantages

- It helps reduce noise by averaging values.
- The higher likelihood value provides a better outcome.

10.2.5.2 Disadvantages

- The shape of the distribution is taken very strongly.
- Data is lost when the continuous functionality is discreet.

10.2.6 Decision Trees

Decision trees support very advanced techniques in support of decision making. A structure like a tree or a diagram is based on cost, classification, and effort. When all the requirements are in place, go from the root to the leaves (until the tasks are complete). Gini indexes are used to calculate the split node. The value of the Gini index aids in separating the nodes. Classification forests also have a random subset of decisions like these. This classification also helps us to diagnose diabetes. This algorithm constructs the regression models. These models are built in a tree-like system, which provides a tree-like structure. It also divides the set data into sub-sets and smaller subsets, creating a tree gradually. This tree includes the particulars of the decision to classify the leaf node, while the decision contains branches. The tree's top decision node will match the root node. This is the best forecast [15].

10.2.6.1 Advantages

- The best predictive model is a thorough analysis of the problem.
- Random forest classifications are ideally suited for vast amounts of data and incomplete data.

10.2.6.2 Disadvantages

- Random forests are quick to train but slow to predict once they are educated.
- Decision trees are unstable even if the input is minimal.

10.2.7 Random Forest

There are parallels between the classification system and this is the technique used for categorizing data. Different trees may use different methods for classifying data and making predictions about regression, but ultimately the goal is to construct a set of decision-making bodies at the data and class levels. This invariance in categorization is more important than gathering training data, especially for decision-making bodies [16].

10.2.8 Neural Network

This classifier's unit names reference vectors, referred to as "nodes," convert the inputs to the vectors known as "features" to their respective outputs. Each neuron enters an input, often a non-linear input, and an output function is given in the next step. The first level entry is the next level output, so the classification algorithm follows a feedback loop. This way, the previous level has no input so that signals passing through neurons and layers can be weighed, and these signals are then translated into a training process that eventually becomes a network to handle a given problem [17].

10.2.8.1 Advantages

- ANN is used for feature extraction with backpropagation for the detection of diabetes.
- When combined with fluid logic, uncertainty can be managed.

10.2.8.2 Disadvantages

- Training requires a great deal of work.
- It is hard to ensure that all inputs are prepared.

10.2.9 Nearest Neighbor

In reality, classification rules the nearest algorithm. The nearest algorithm for classification is a common term for this method. Labeled point clusters are used to shed light on how the other points were assigned their labels. When adding a new marker, it first looks for possible neighbors (those that are geographically close) to the spot being added. When a new item receives a neighbor's same rating, it becomes equivalent to the rating of other items in the vicinity, depending on the neighbor's vote. In the 'k' algorithm, this is the count of confirmed neighbors [18].

All of the necessary knowledge for implementing the categorization algorithms and procedures used to foresee the sickness has already been acquired. Following this poll, it was suggested that a combination of a hybrid classification algorithm and any kind of learning might increase the disease's predicted accuracy by more than 80%. When more than two classifiers are combined, accuracy improves. We build a model for assessing the quality of training data using a combination of the decision tree and other classifiers. In addition to the aforementioned techniques, we have also employed XGBoost for each classifier and analyzed its performance. With this combo, we can boost accuracy by over 80% [19].

10.3 Hybrid Classification Algorithm System

Specificity for the presence of diabetes in patients is maximized using the proposed strategy. In this discussion, we examine the many types of machine learning algorithms and how they can be used to produce predictions and inferences. More than one algorithm can be used to boost the reliability of predictions (Fig. 10.1).

After collecting information related to a disease, it is sent to a pre-processed device for analysis. Extractive functions will be used to glean information from unstructured, large, or otherwise uninteresting data sets. The necessary machine learning method is used for the data training and data gathering, and the results are then evaluated [20]. Then, to get the desired outcome, we employ a classifier

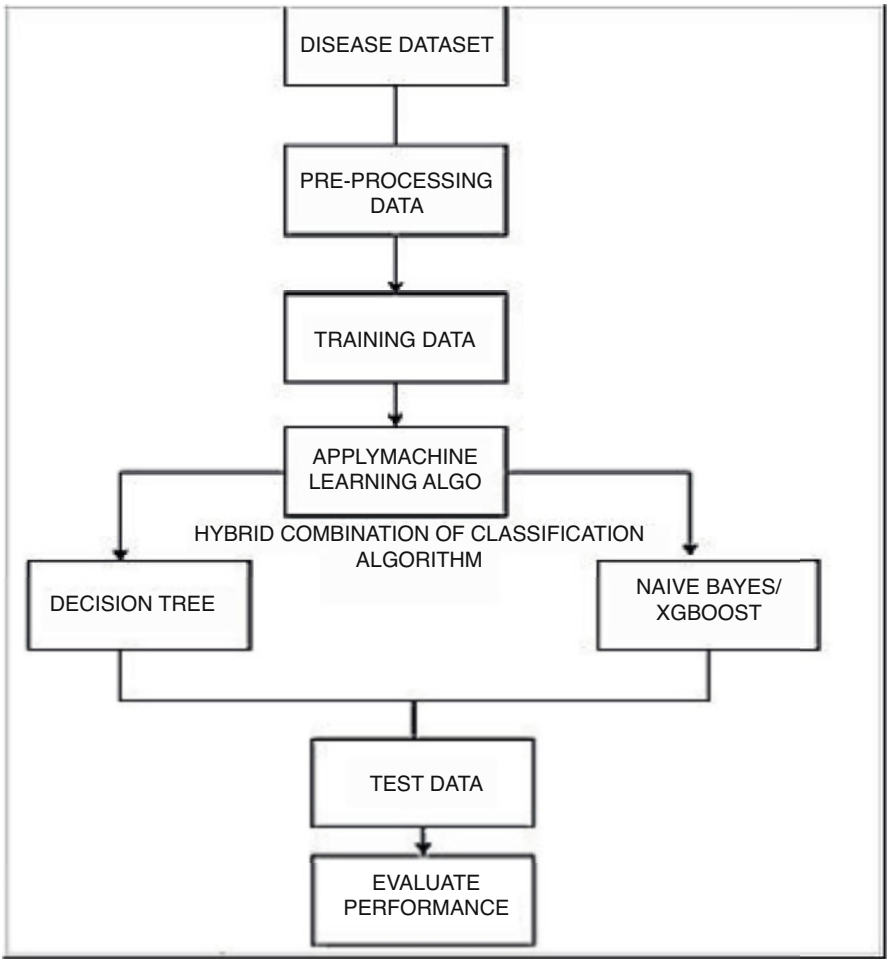

Fig. 10.1 Diagram of the proposed system architecture

combination. In this way, we recommend combining the Decision Tree with the Support Vector Machine, the Decision Tree with XGBoost and the Hybrid Results Test method. The information will then be monitored and the desired outcomes evaluated. We now study the various classifiers and explore the hybrid mix used in our scheme. There are multiple forms of classification; an algorithm is a classification system that maps data entered into a particular group.

10.3.1 System Flow

Figure 10.2 explains the device design flow map; we also made feature selection for pre-processing: advance feature selection and backward feature selection. This is to show how data can be provided using ADA Boost, XGboost, voting classifiers, and stacking classifiers to assist in the identification of people who are likely to develop diabetes. The proposed approach has two major phases in which the desired effects can be achieved together. Data are prepared in the first stage, and the second stage is classified. The machine input is then the PID dataset, and the output is a stable or diabetic class. Steps are taken to increase productivity. The data collection eliminates the noisy and inaccurate data, first of all.

The main goal is to identify whether or decrease the likelihood of diabetes. When the number of samples increases, so does the classification accuracy but with no real increase in statistical significance (Fig. 10.3).

In certain instances, but not all, the algorithm yields performance high in rhythm but low in classification. Our model's primary goal is to achieve high precision. Different methods for classifying diabetic and non-diabetic outcomes were

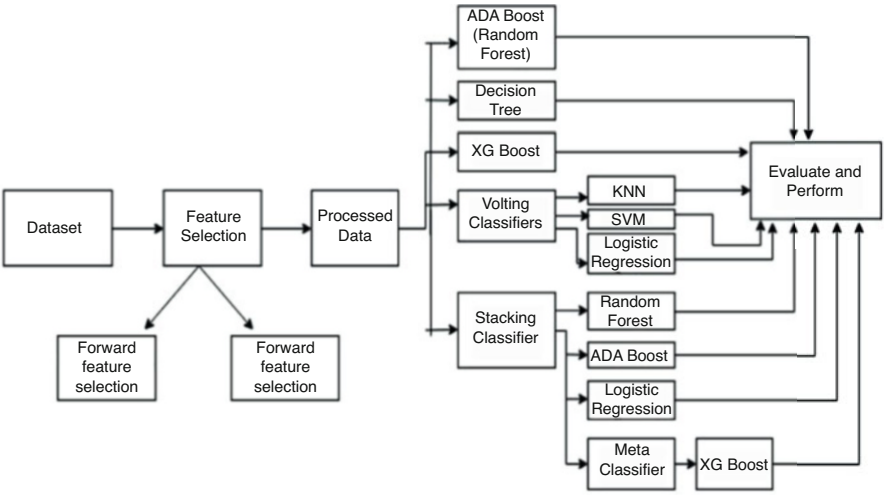

Fig. 10.2 Flow of proposed system

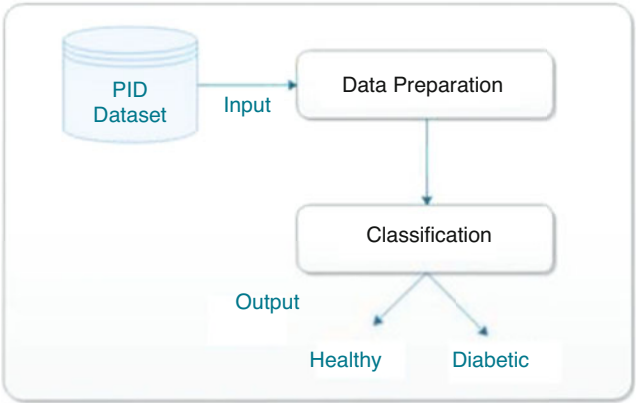

Fig. 10.3 Proposed system diagram

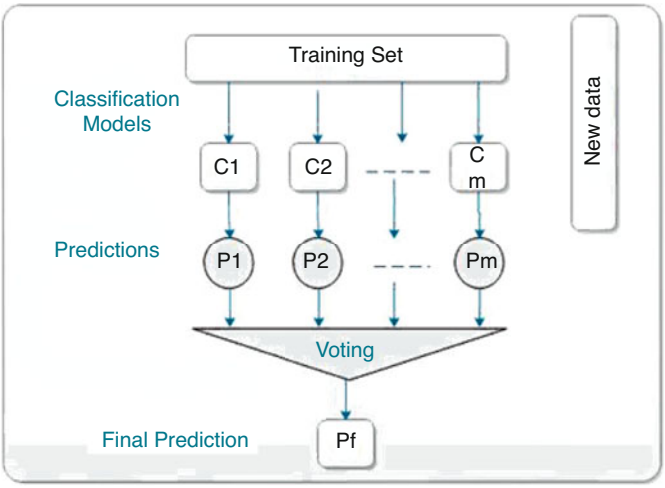

Fig. 10.4 Functional diagram for voting classifier

examined. This proposed approach uses techniques such as AdaBoost, Tree Classification Decision, XGBoost, voting classification, and Diabetes Prediction Stacking. We will go into classifiers now and then go through the classification of the stacking and voting in the following parts (Fig. 10.4).

- Stacking: Stacking is an ensemble learning strategy that integrates predictions from several simple models plus a new dataset. These new data are taken for another classifier as input data. The issue was resolved with the help of this rating. A synonym for stacking is mixing.
- Voting Classifier: "hard" and "soft" voting is carried out by the Vote Classifier Ensemble. When we vote strongly, the final class label in the classification

models is expected to be the most previewed class label. By averaging class probabilities in soft votes, we predict class labeling (advocated only if the classifiers are considered accurate).

10.3.2 Working Principle

Gives step-by-step instructions on how to use various classifiers to improve accuracy.

10.3.3 Hardware Requirements

For the device implementation, the following hardware was used:

- RAM SIZE: 4GB
- 10GB HDD
- Processor Type: Intel 1.66 GHz Pentium 4B.

10.3.4 Implementation Steps

PIMA Indian Diabetes Dataset has been selected as the diabetes data kit of choice. There are a total of 768 instances, both diabetic and non-diabetic, and four risk factors: the prevalence among pregnant women, the concentration of plasma glucose after 2 hours of oral glucose administration, diastolic blood pressure, and triceps skin thickness. Whether it is done automatically or manually, Feature Selection determines which features are most relevant to the desired prediction attribute or output. The quality of our models can suffer if our data contains extraneous information (Fig. 10.5).

- We use a PIMA Indian diabetic dataset.
- The system uses the function selection method: Advanced selection of features and backward feathering for pre-processing. We train five different graders and determine which graders are highly accurate. We used these AdaBoost, XGBoost, Voting Classifier, and Stacking Classifier classifiers.
- The Stacking Classifier is used as its basis by Random Forest, AdaBoost, Logistical Regression, and XGBoost for its Meta classifier.
- The Acoustic and Stacking Classifier is the most powerful because of its improved ability to classify power and class strength.
- In order to better comprehend the sequence of our measures and their intended results, we have included screenshots below. The ADA Boost classifier's output

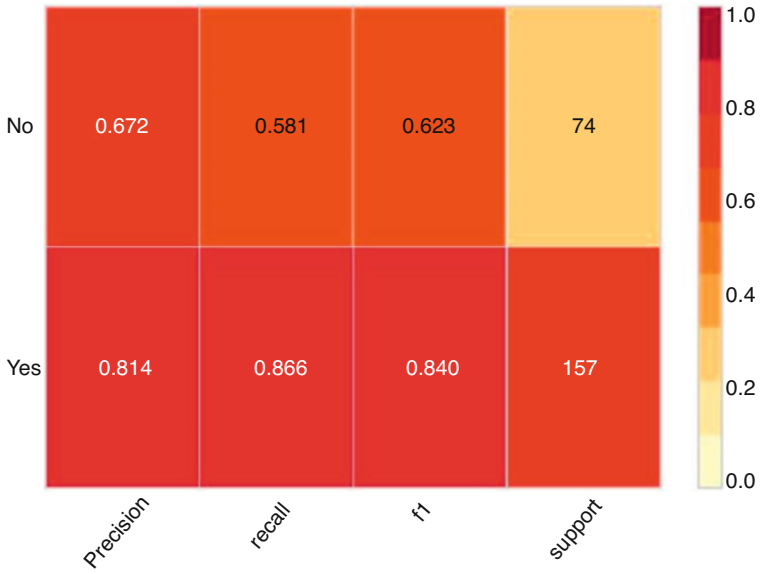

Fig. 10.5 AdaBoost classification report

will be displayed graphically. For Decision Tree, XG Boost, Vote, and Stacking, we have taken analogous measures.

We discuss the classification findings after first introducing the AdaBoost classification method. ROC curves were calculated using the same methodology applied to the decision tree, XGBoost, voting, and stacking classifiers. For more examples, check out the images below (Figs. 10.6, 10.7, 10.8, and 10.9).

We then determine the total number of diabetes sponsors. Next, a screen displays test results for a person with diabetes (Fig. 10.10).

10.4 Results

Achieving a success rate of 80% or above in predicting diabetes using five separate classification schemes is a major step forward. You may do some basic analysis using the graphs in Figs. 10.11 and 10.12. The AUC, accuracy, recall, and F1 results from various classifications are displayed in Fig. 10.11. Figure 10.12 depicts how histograms are appropriately displayed using different classifiers (Table 10.2).

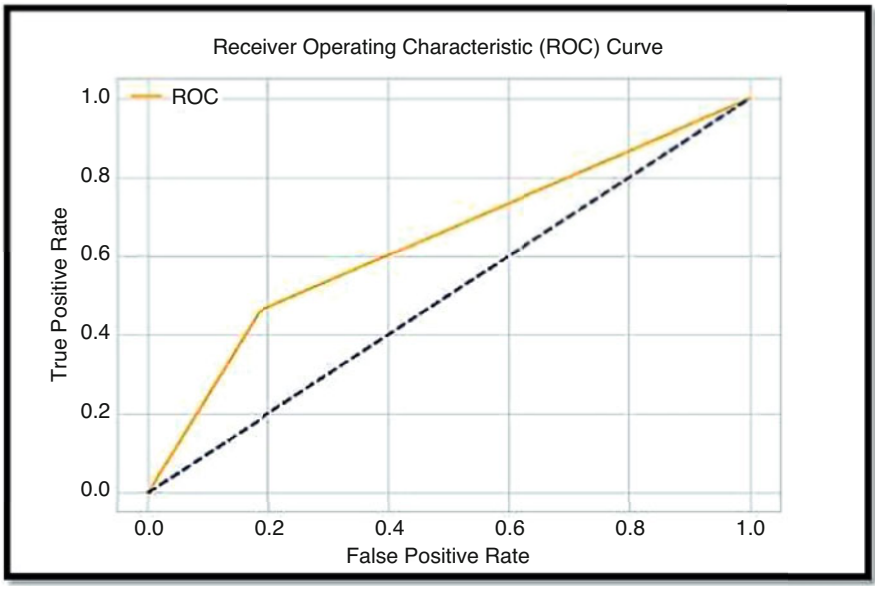

Fig. 10.6 ROC decision tree

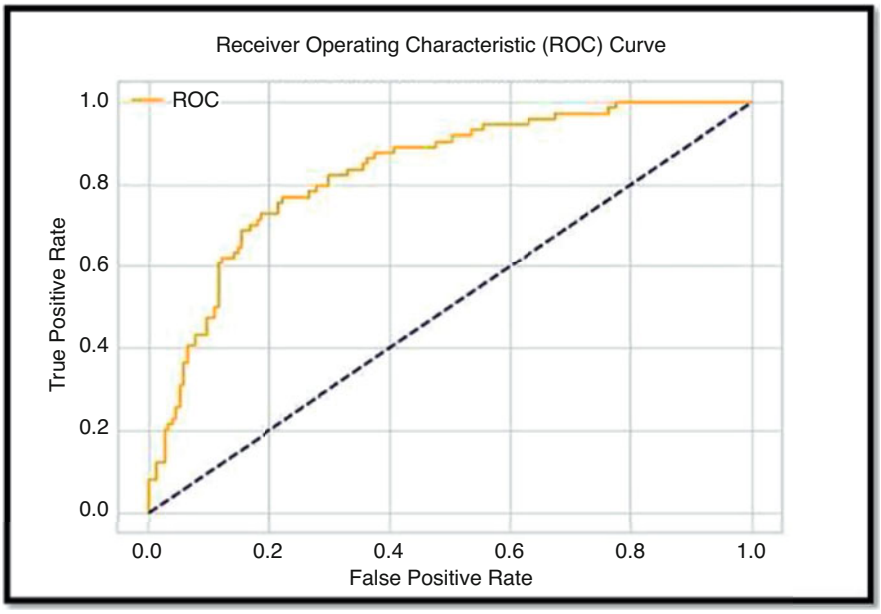

Fig. 10.7 ROC XGBoost

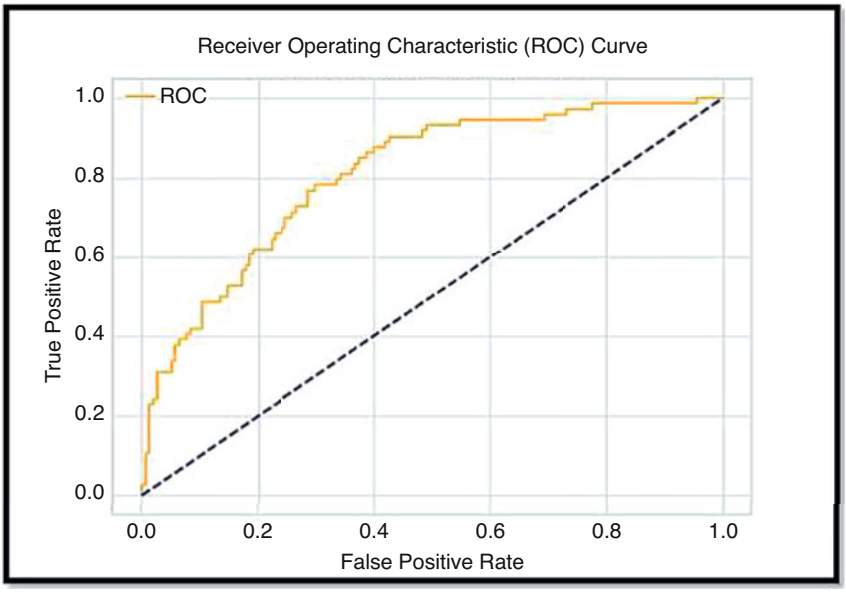

Fig. 10.8 Classifier ROC voting

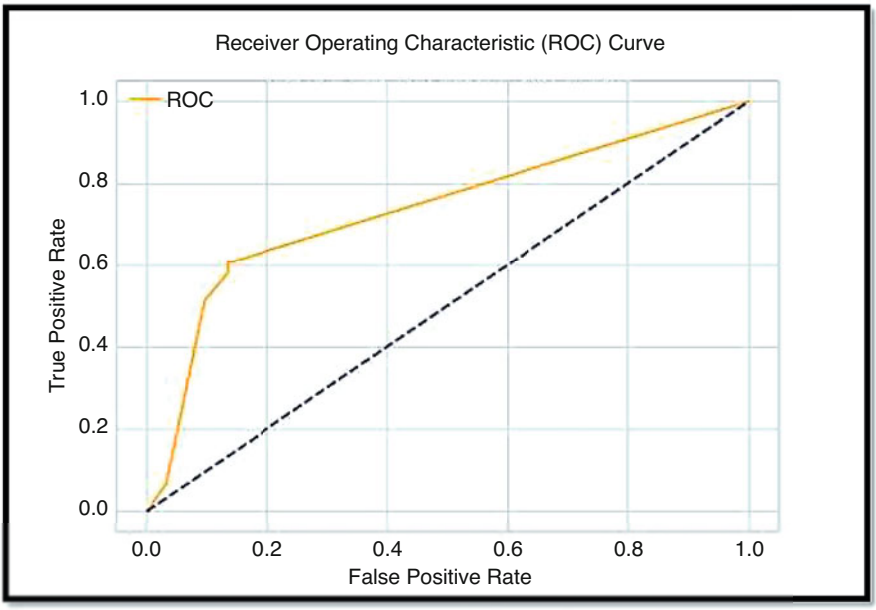

Fig. 10.9 Classifier ROC voting

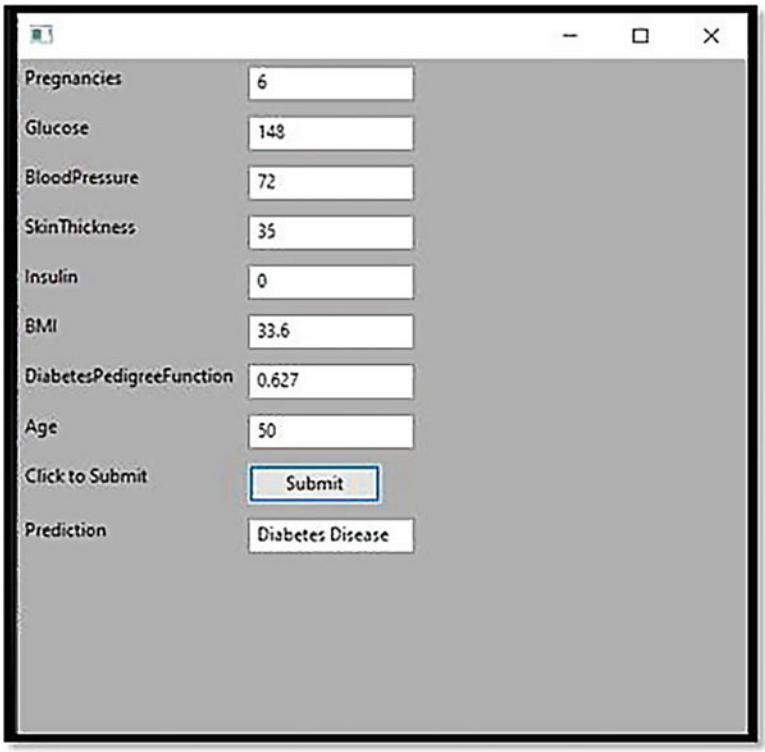

Fig. 10.10 Detected diabetes shown on the test screen

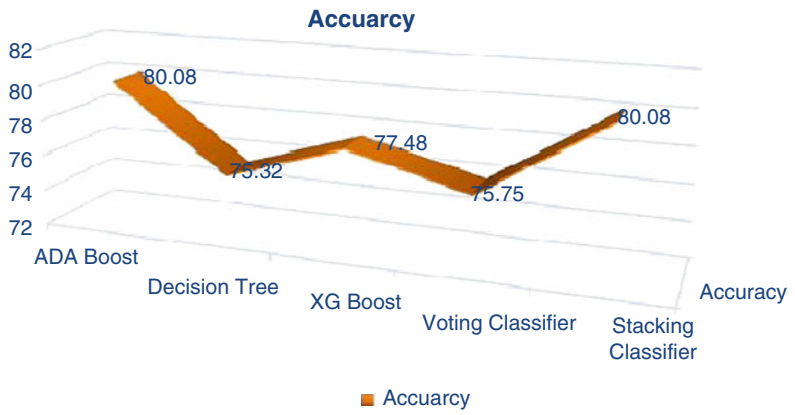

Fig. 10.11 Graph for precision, F1 score, AUC, and recall

Fig. 10.12 Comparison graph for all algorithms

Table 10.2 Observations

Classifier	Precision	Recall	AUC	F1	Accuracy
ADA BOOST	0.82	0.90	0.83	0.86	80.08
Decision Tree	0.80	0.84	0.70	0.82	75.32
XGBOOST	0.80	0.90	0.83	0.84	77.48
Voting (KNN, SVM, Logistic regression)	0.77	0.91	0.83	0.84	75.75
Stacking (Random Forest, ADAboost, Logistic regression)	0.82	0.90	0.75	0.86	80.08

10.5 Conclusion

Medical professionals can benefit from the use of machine learning techniques for the diagnosis and treatment of diabetes. Finally, we will state that enhanced categorization accuracy paves the way for enhanced machine learning model outputs. The performance analysis focuses on the accuracy of all classification techniques and also found that the present method was less than 71% accurate, so we suggest using a mixture of classifications known as the hybrid approach. The hybrid solution is based on the advantages of two or more technologies. Our scheme provides 75.33% for decision-making tree classification, 77.47% for XGBoost, and 75.76% for voting classification. We have therefore found that Stacking Classifier and AdaBoost are the strongest of all the above classifiers.

References

1. Kaur, H., & Kumari, V. (2018). Predictive modelling and analytics for diabetes using a machine learning approach. *Applied Computing and Informatics*.
2. Carter, J. A., Long, C. S., Smith, B. P., Smith, T. L., & Donati, G. L. (2019). Combining elemental analysis of toenails and machine learning techniques as a non-invasive diagnostic tool for the robust classification of type-2 diabetes. *Expert Systems with Applications, 115*, 245–255.
3. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Computational and Structural Biotechnology Journal, 15*, 104–116.
4. Mahmud, S. M., Hossin, M. A., Ahmed, M. R., Noori, S. R. H., & Sarkar, M. N. I. (2018). Machine learning based unified framework for diabetes prediction. In *Proceedings of the 2018 international conference on big data engineering and technology* (pp. 46–50). ACM.
5. Patil, R., & Tamane, S. (2018). A comparative analysis on the evaluation of classification algorithms in the prediction of diabetes. *International Journal of Electrical and Computer Engineering, 8*(5), 3966.
6. Dagliati, A., Marini, S., Sacchi, L., Cogni, G., Teliti, M., Tibollo, V., et al. (2018). Machine learning methods to predict diabetes complications. *Journal of Diabetes Science and Technology, 12*(2), 295–302.
7. Barik, R. K., Priyadarshini, R., Dubey, H., Kumar, V., & Yadav, S. (2018). Leveraging machine learning in mist computing telemonitoring system for diabetes prediction. In *Advances in data and information sciences* (pp. 95–104). Springer.
8. Choudhury, A., & Gupta, D. (2019). A survey on medical diagnosis of diabetes using machine learning techniques. In *Recent developments in machine learning and data analytics* (pp. 67–78). Springer.
9. Samant, P., & Agarwal, R. (2017). Diagnosis of diabetes using computer methods: Soft computing methods for diabetes detection using iris. *Threshold, 8*, 9.
10. Dankwa-Mullan, I., Rivo, M., Sepulveda, M., Park, Y., Snowdon, J., & Rhee, K. (2019). Transforming diabetes care through artificial intelligence: The future is here. *Population Health Management, 22*(3), 229–242.
11. Joshi, T. N., & Chawan, P. M. (2018). Logistic regression and svm based diabetes prediction system. *International Journal For Technological Research In Engineering 5*.
12. Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. *JAMA, 319*(13), 1317–1318.
13. Nnamoko, N., Hussain, A., & England, D. (2018). Predicting diabetes onset: An ensemble supervised learning approach. In *2018 IEEE congress on evolutionary computation (CEC)* (pp. 1–7). IEEE.
14. Yadav, B., Sharma, S., & Kalra, A. (2018). Supervised learning technique for prediction of diseases. In *Intelligent communication, control and devices* (pp. 357–369). Springer.
15. Joshi, R., & Alehegn, M. (2017). Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach. *International Research Journal of Engineering and Technology, 4*(10).
16. Singh, D. A. A. G., Leavline, E. J., & Baig, B. S. (2017). Diabetes prediction using medical data. *Journal of Computational Intelligence in Bioinformatics, 10*(1), 1–8.
17. Gujral, S. (2017). Early diabetes detection using machine learning: A review. *International Journal for Innovative Research in Science & Technology, 3*(10), 57–62.
18. Zia, U. A., & Khan, N. (2017). Predicting diabetes in medical datasets using machine learning techniques. *International Journal of Scientific & Engineering Research, 8*.
19. Naqvi, B., Ali, A., Hashmi, M. A., & Atif, M. (2018). Prediction techniques for diagnosis of diabetic disease: A comparative study. *International Journal of Computer Science and Network Security, 18*(8), 118–124.
20. Chen, J. C. H., Kang, H. Y., & Wang, M. C. (2018). Integrating feature ranking with ensemble learning and logistic model trees for the prediction of postprandial blood glucose elevation. *Journal of Universal Computer Science, 24*(6), 797–812.