INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

An Ensemble Learning Approach for Effective Prediction of Diabetes Mellitus Using Hard Voting Classifier

Mohammad Atif[1]*, **Faisal Anwer**[1], **Faisal Talib**[2]

1 Department of Computer Science, Aligarh Muslim University, Aligarh, U.P, India
2 Department of Mechanical Engineering, Aligarh Muslim University, Aligarh, U.P, India

OPEN ACCESS

Received: 23-07-2022

Accepted: 16-09-2022

Published: 15-10-2022

Citation: Atif M, Anwer F, Talib F (2022) An Ensemble Learning Approach for Effective Prediction of Diabetes Mellitus Using Hard Voting Classifier. Indian Journal of Science and Technology 15(39): 1978-1986. https://doi.org/10.17485/IJST/v15i39.1520

* **Corresponding author**.

atif.sidau@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: People all across the world are afflicted by the deadly ailment known as diabetes. Diabetes is a terrible condition characterized by high blood glucose levels. This chronic condition is one of the leading causes of death for people worldwide. Early identification and prediction of diabetes can be aided by machine learning techniques. The purpose of this study is to use an ensemble of machine learning algorithms to predict diabetes efficiently in order to help the patients suffering from this lethal disease. **Methods:** The existing methods use a single model to predict diabetes, which may have an impact on accuracy because no one model can fit all datasets. Therefore we propose a robust model based on ensemble learning using hard voting classifier. Both the Pima Indians Diabetes dataset and the Early Stage Diabetes Risk Prediction Dataset, which collect data on people with and without diabetes, were tested. For classification, the proposed ensemble hard voting classifier uses a combination of three machine learning algorithms namely logistic regression, decision tree, and support vector machine. **Findings:** On the PIMA diabetes dataset, the proposed ensemble approach achieves the highest accuracy, precision, recall, and F1 score value of 81.17%, while on the Early Stage Diabetes Risk Prediction Dataset, it achieves the highest accuracy, precision, recall, and F1 score value of 94.23%. **Novelty:** The proposed methodology was experimentally tested using the state-of-the-art technology and basic classifiers such as K-Nearest Neighbor, Logistic Regression, Support Vector Machine, and Random Forest. The results are validated by computing the confusion matrix and ROC for each classier type.

Keywords: Diabetes Detection; Machine Learning; Supervised Classification; Ensemble Classification; Hard Voting Classifier

1 Introduction

Diabetes mellitus is a metabolic disease that gets worse as the body loses the ability to use glucose[1]. It is a serious illnesses that affects people all over the world. This chronic illness is among the most common causes of mortality in people all over the world. It is estimated that there are approximately 415 million people throughout the world who are between the ages of 20 and 79 and suffering with some sort of diabetic disease right now[2]. In 2013, global diabetes data showed that the disease affected 382 million people worldwide . It was the fifth biggest cause of mortality for women in 2012, and the eighth for both men and women . Diabetes is more common in high-income countries. It is anticipated that there would be 963 million diabetics in the globe by 2045, with half of the population untreated. In addition, in 2017, 850 million dollars have been spent on diabetes . A report claims that the prevalence of diabetes is increasing and that there are currently 500 million cases worldwide, with rates anticipated to reach 25% and 51% in 2030 and 2045, respectively[3].

Alcohol, salicylates, and narcotics all raise A1C. Vitamin C may boost A1C levels by electrophoresis but reduce by chromatography. A higher leukocyte count during hypertension is associated with chronic inflammation, according to the majority of studies[4]. A single parameter can't diagnose diabetes and lead to wrong judgments. To accurately forecast diabetes at an early stage, different characteristics must be combined. Therefore automatic diabetes detection based on a combination of factors can assist clinicians in treating patients more effectively and efficiently. Machine learning based solutions are always promising because they learn from actual features to diagnose diabetes[5]. The enormous amount of data generated in this field presents two significant challenges for researchers and developers trying to construct diabetes predictive models. First off, the machine learning approaches utilized in earlier studies have been somewhat heterogeneous, making it difficult to choose the best one. Second, the features that were used to train the models are not transparent, which decreases their interpretability—a feature that is incredibly important to the clinician. Numerous studies on disease prediction, including diagnosis, prognosis, categorization, and therapy, have been conducted. Recent research demonstrates that a variety of machine learning methods have been utilized for disease identification and forecasting. They have resulted in significant efficiency benefits and improvements in both traditional and machine learning algorithm methodologies.

According to the study, a number of machine learning algorithms and ensemble approaches were employed to categorise diseases on the PIMA dataset and the Early Stage Diabetes Risk Prediction dataset, however none of them were able to do so with an accuracy rate of more than 76%. So as to improvise outcomes, we suggested a system framework based on vote categorization that employs an ensemble learning approach. The theoretical foundations, properties, and performances of learning models and algorithmic techniques are the primary emphasis of this research. Instead of using regression to predict disease, the classification strategy has been used. A hard voting classifier has been suggested to classify diabetes using the ensemble approach. For the logistic regression, decision tree, support vector machine, and voting classifier algorithms, the performance of the ensemble technique has demonstrated improved results when compared to base classifiers. The performance of the aforementioned technique has been evaluated using accuracy, precision, recall, F1-score, and AUC.

In short, our significant contributions are as follows:

- We employ ensemble learning approach instead of a single model that combines multiple techniques together.
- We propose a system model that integrates several ML techniques to provide a single robust result based on Voting Classifier.
- We compute accuracy, F1 score, precision, recall, Confusion matrix, ROC Curve for our predicted labels and compare them with base line algorithms.

Related Work

[5] used the KNN and Naive Bayes approach to predict diabetes. Their method is implemented as professional software, where users enter input in the form of patient data to determine if a patient has diabetes.[6] implemented voting classifier that has sigmoid SVC, AdaBoost, and Decision tree algorithms for the prediction of diabetes and heart disease.[7] utilized Hard Voting Ensemble Approach for the Detection of Type 2 Diabetes with non-glucose related features.[8] implemented random forest, XGBoost and Light GBM techniques for the early detection of type 2 diabetes.[9] used logistic regression on PIDD to predict diabetes disease. They discovered that the number of glucose level, pregnancies, and BMI are the most relavent indicators for diabetes detection among all the parameters in PIDD.[10] applied Decision Tree, Random Forest and Neural Network for the prediction of diabetes. In this study Random Forest accurately predicts diabetes with a sensitivity, specificity, and accuracy of 80%, 89%, and 85% respectively.

2 Methodology

2.1 Datasets

Data from the PIMA Indians Diabetes (PID) dataset and the Early Stage Diabetes Risk Prediction Dataset were used to train and test the machine learning models. The data for this study was made available by the National Institute of Diabetes and Digestive and Kidney Diseases (published in the UCI ML repository[11]).The primary goal of using this dataset was to utilise its diagnostic capabilities to predict if a patient has diabetes through diagnosis. There were many restrictions when choosing occurrences from larger datasets. Both datasets and tasks are monitored binary classification tasks. All features are described in detail in Table 1.

The second dataset contains 520 people's reports of diabetes related symptoms[12]. It includes information on individuals, including diabetic symptoms. This information was gathered by a direct poll of persons who had just been diagnosed with diabetes or who were not diabetic but had one or more symptoms. The information was gathered from patients at Early Stage Diabetes Risk Prediction Dataset, using a direct questionnaire. The missing values were handled using the approach of disregarding tuples with partial values during the data preprocessing. After preprocessing, a total of 500 instances remained. There are 314 positive values and 186 negative ones among them. Tables 1 and 2 provide a detailed overview of the dataset and its properties. For a patient to be diabetic or not, there are two classes: positive and negative.

2.2 System Model

The purpose of this research is to enhance diabetes diagnosis results and accuracy. We suggested an ensemble of machine learning algorithms based on a hard voting classifier for binary classification of diseases into positive and negative states. The data is initially pre-processed before it is entered into the model. The proposed ensemble approach, which employs a hard voting classifier, is depicted in Figure 1.

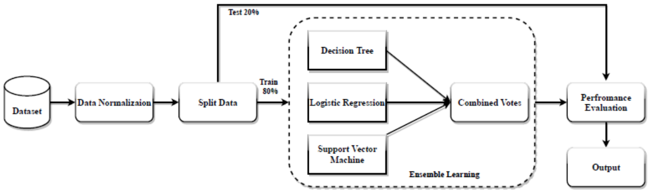


Fig 1. System Model

Table 1. PIMA Indian Diabetes Dataset

S.No	Attributes	Overview
1	Pregnancies	Number of times a woman got pregnant
2	Glucose (mg/dl)	Glucose concentration in oral glucose tolerance test for 120 min
3	Blood Pressure (mmHg)	Diastolic Blood Pressure
4	Skin Thickness (mm)	Fold Thickness of Skin
5	BMI (kg/m2)	Body Mass Index (weight/(height)^2)
6	Diabetes Pedigree Function	Diabetes pedigree Function
7	Age	Age (years)
8	Outcome	Class variable (class value 1 for positive 0 for Negative for diabetes)

Table 2. Early Stage Diabetes Risk Prediction Dataset

S.No.	Attributes	Values
1	Age	20–35, 36–45, 46–55,56–65, above 65
2	Sex	Male, Female
3	Polyuria	Yes, No.
4	Polydipsia	Yes, No.
5	Sudden weight loss	Yes,No.
6	Weakness	Yes, No.
7	Polyphagia	Yes, No.
8	Genital thrush	Yes, No.
9	Visual blurring	Yes, No.
10	Itching	Yes, No.
11	Irritability	Yes, No.
12	Delayed healing	Yes, No.
13	Partial paresis	Yes,No.

Continued on next page

Table 2 continued

14	Muscle stiffness	Yes, No.
15	Alopecia	Yes, No.
16	Obesity	Yes, No.
17	Class	Positive, Negative

2.3 Model Architecture

- **Data Prepossessing:** There may be missing values, as well as noisy and inconsistent data, in real-world data[13]. Poor data quality could lead to unfavourable results. Preprocessing the data is necessary to generate findings of a high calibre. Data is cleaned as part of preprocessing. In terms of time, cost, and quality, improving the data's suitability for data mining and analysis is critical. The model entails a number of phases as shown in Figure 1. Filling in missing values and eliminating noisy data are two aspects of data cleaning. Outliers are removed from noisy data to reconcile inconsistencies[14]. The PIMA dataset has zero (0) values for glucose, blood pressure, skin thick- ness, insulin, and BMI. As a consequence, the attribute's median value was substituted for all zero values. For better model training, we transformed 'Yes' to 1 and 'No' to 0 in the second dataset.
- **Data Normalization :** Data X is normalised in the range (0, 1) in the pre- processing step using Eq (1). Where X_{min} and X_{max} are minimum and maximum values respectively obtained from data set. The network receives the value of normalised data x as input. Following that, the network is trained and evaluated, and temperature predictions are made. Furthermore, the data is split into two categories: training and testing, with the former accounting for 80% of the total. Afterwards, the ensemble learner input is fed with the training data. We used an ensemble of machine learning methods in this proposed methodology, including Decision Tree, Logistic Regression, and SVM classifiers. Combining the aforementioned algorithms with a hard voting classifier improved their accuracy. In the following part, we'll go through these algorithms in more detail.

$$x_i = \frac{X_i - X_{min}}{X_{max} - X_{min}} \tag{1}$$

- **Logistic Regression**[15]: To predict binary outcomes (y = 0 or 1), logistic regression employs statistical approaches. As a linear learning algorithm, logistic regression is used. A logistic regression prediction is one that takes into account the probability of an event. With the LR approach, each data point is mapped using the sigmoid function. Using the typical logistic function, an S-shaped curve is created. The sigmoid function is shown in Eq. (2).

$$sigmoid = \frac{1}{1+e^{-x}} \tag{2}$$

- **Decision Tree** [16]: A decision tree is a tree-based approach in which each path coming from the root is distinguished by a data-separating sequence until the leaf node yields a Boolean outcome. Decision trees are supervised learning systems which can address classification and regression problems, but in general they are used to resolve classification problems. In this tree-structured classifier, internal nodes reflect data record properties, branching displays decision criteria, and every leaf node provides an answer. The decision tree's two nodes are the decision node and the leaf node. The leaf node depicts the outcome of these decisions and has no more branches, but the selection node is used to make the decision and has many branches.
- **Support Vector Machine**: SVM is a non-parametric algorithm that can use both linear and non-linear functions to solve regression and classification problems. These functions put vectors of features from the input into a space with n dimensions called a "feature space"[17,18]. Finding hyperplanes in N-dimensional space that can distinguish between data points is the goal of the SVM approach (N = number of attributes). The two sorts of data points can be divided using a variety of hyperplanes. The aim is to locate the plane containing the largest margin, or separation between data points from the two classes. By increasing the margin distance, more reinforcement is produced, which makes it easier to classify the subsequent data points.
- **Proposed Approach:** This classifier, a meta-classifier, combines similar or conceptually dissimilar machine learning models to create predictions via majority voting[19]. Voting classifiers use both a hard voting approach and a soft voting

approach. The final prediction is made by a majority vote in hard voting[20], where the aggregator selects a class prediction that is displayed repeatedly between the base models. The suggested model combines Logistic Regression, Decision Trees, and SVM classifiers. Hard voting is the most basic kind of majority voting. To predict the class label Z, we use the majority (plurality) voting of each classifier as shown in Eq. (3). Hard voting analyses the decision of maximal classifiers, as Algorithm 1 demonstrates. A workflow of hard voting classifier has been shown in Figure 2. Before voting, it calculates the predictions of each classifier. Finally, it calculates the set's mode and expresses the conclusion based on the majority of classifiers' decisions.

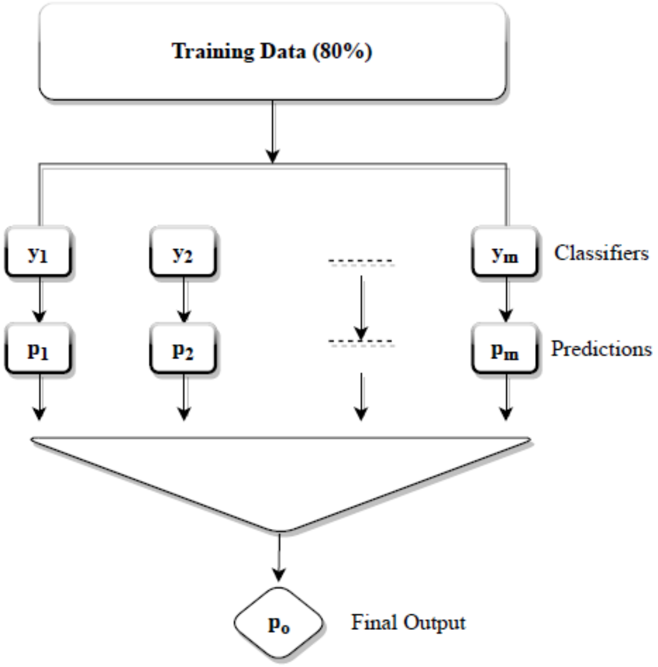
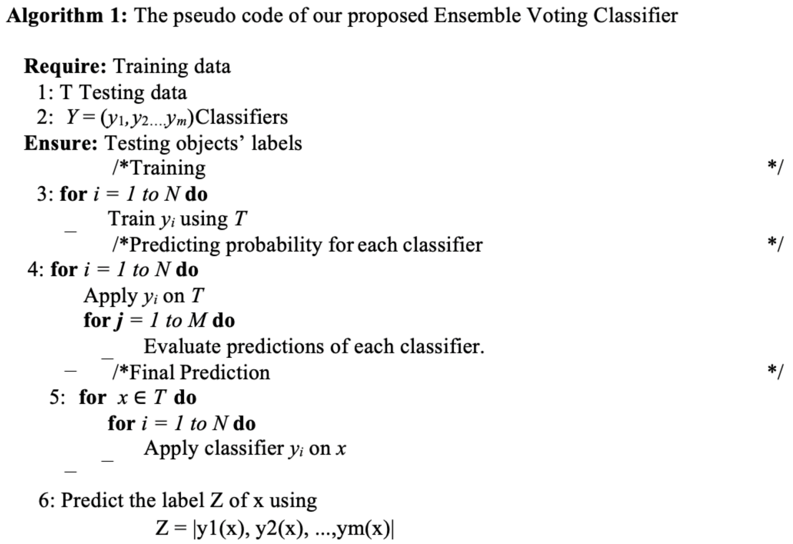


Fig 2. Workflow of Hard voting Classifier

$$z = |y1(x), y2(x), \ldots, ym(x)| \quad (3)$$



Algorithm 1: The pseudo code of our proposed Ensemble Voting Classifier

```
Require: Training data
  T Testing data
  $Y = (y_1, y_2 ... y_m)$Classifiers
Ensure: Testing objects' labels
      /*Training                                          */
  for $i$ = $1$ to $N$ do
      Train $y_i$ using $T$
      /*Predicting probability for each classifier        */
  for $i$ = $1$ to $N$ do
      Apply $y_i$ on $T$
      for $j$ = $1$ to $M$ do
          Evaluate predictions of each classifier.
      /*Final Prediction                                  */
   for $x \in T$ do
      for $i$ = $1$ to $N$ do
          Apply classifier $y_i$ on $x$

  Predict the label Z of x using
          Z = |y1(x), y2(x), ...,ym(x)|
```

Table 3. Accuracy

Dataset	KNN	LR	SVM	RF	Proposed Approach
D_1	0.817307692	0.913461538	0.615384615	0.932692308	0.942307692
D_2	0.74025974	0.811688312	0.798701299	0.792207792	0.811688312

Table 4. Precision

Dataset	KNN	LR	SVM	RF	Proposed Approach
D_1	0.894736842	0.936507937	0.615384615	0.938461538	0.942307692
D_2	0.581395349	0.725	0.735294118	0.682926829	0.811688312

Table 5. Recall

Dataset	KNN	LR	SVM	RF	Proposed Approach
D_1	0.796875	0.921875	1	0.953125	0.942307692
D_2	0.531914894	0.617021277	0.531914894	0.595744681	0.811688312

Table 6. F1 Score

Dataset	KNN	LR	SVM	RF	Proposed Approach
D1	0.842975207	0.929133858	0.761904762	0.945736434	0.942307692
D2	0.555555556	0.666666667	0.617283951	0.636363636	0.811688312

3 Results and Discussion

3.1 Evaluation Metrics

The proposed methodology combines three machine learning models: Logistic Regression, Decision Trees, and Support Vector Machines, as well as a hard voting classifier. The PIMA diabetic dataset as well as the Early Stage Prediction Dataset were used in the experiment. The dataset was separated into two parts: test and training, each having 20% and 80% of the total data. The accuracy, precision, recall, and F1 score of the algorithms are the most common assessment criteria used to examine their robustness and efficiency. True positive (Tp) denotes that the expected and actual class values are both 1. A true negative (Tn) shows that the expected and actual class values are both 0. False positives (Fp) and false negatives (Fn) appear whenever the expected class and the actual class aren't the same. Accuracy is the most crucial metric. It is derived by dividing the total number of correctly predicted observations by the total number of observations. The following formulas can be used to determine accuracy, precision, recall, and F1 score as shown in Equations (4), (5), (6), and (7) respectively:

$$Accuracy = \frac{T_p + T_n}{T_p + T_n + F_n + F_p} \tag{4}$$

$$Recall = \frac{T_p}{T_p + F_n} \tag{5}$$

$$Precision = \frac{T_p}{T_p + F_p} \tag{6}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{7}$$

T_p (y-axis) and F_p (x-axis) are plotted in ROC Curves (Receiver Operating Characteristic Curves) (x-axis). We can choose the threshold of the probabilities because our proposed model identifies the patient as having diabetes or not depending on the probability provided for every class. For instance, let's say we wish to fix a threshold of 0.4. This indicates that if the chance of the patient having diabetes is greater than 0.4, the model will categorise the data point/patient as having diabetic condition. This

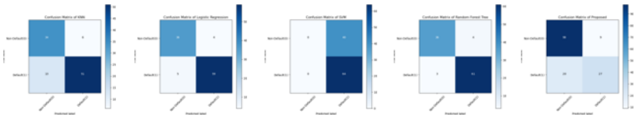
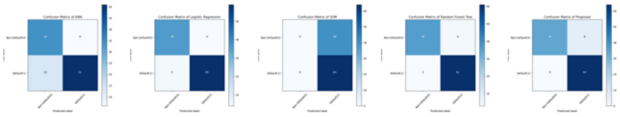


Fig 3. D1: Confusion Matrix [(a)KNN (b) LR (c) SVM (d) RF (e) Proposed]



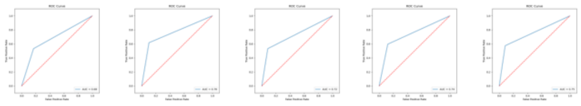
(a) KNN (b) LR (c) SVM (d) RF (e) Proposed

Fig 4. D2: Confusion Matrix [(a)KNN (b) LR (c) SVM (d) RF (e) Proposed]

will, unsurprisingly, result in a high recall value and a lower amount of False Positives. Likewise, we can use the ROC curve to see how our model performs for various threshold settings.KNN (b LR (c SVM (d RF (e Proposed

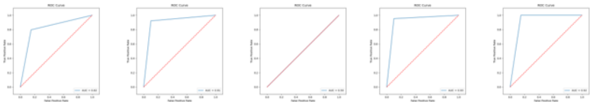
On datasets D1 and D2, our proposed approach achieved maximum accuracy of 94.24% and 81.12%, respectively, as shown in Table 3. Individual models like KNN, LR, SVM, and RF are demonstrably less accurate than the proposed method. As a result, the ensemble's performance and resilience in terms of prediction and performance should be superior to that of any one contributing model. It also improves the model's performance by reducing the spread or dispersion of the predictions. The good predictive performance is due to the fact that adding bias reduces the variance component of the prediction error. In comparison to other baseline algorithms, the suggested approach likewise achieved the highest precision, as shown in Table 4. On the datasets D_1 and D_2, our proposed approach has precision scores of 94.24% and 81.12%, respectively. As shown in the table, recall is a measure of how well a model detects true positives. It is shown in Table 5 . As a result, recall shows the exact number of people with diabetes who have been identified as having diabetes. For D_1 and D_2, recall = 94.24% and 81.12%, respectively. The recall of our model is a metric that measures how well it can recognise relevant data. The highest recall values were obtained by SVM and RF, which were 100% and 95.53%, respectively.

Although we strive for great precision and recall value, we cannot achieve both at the same time. We also calculated F1-score for our datasets. It is calculated using the Harmonic Mean of Precision and Recall as shown in Table 6. On dataset D1, our model achieved a maximum F1 score of 94.24%, as shown in Table 6. RF, with a F1 score of 94.57%, has the highest F1 score. However, on dataset D2, our model achieved a maximum F1 score of 81.12%.



(a)KNN (b) LR (c) SVM (d) RF (e) Proposed

Fig 5. D1: ROC Curve [(a) KNN (b) LR (c) SVM (d) RF (e) Proposed]



(a) KNN (b) LR (c) SVM (d) RF (e) Proposed

Fig 6. D2: ROC Curve [(a) KNN (b) LR (c) SVM (d) RF (e) Proposed]

The confusion matrix is used to convey critical predictive metrics including recall, specificity, accuracy, and precision. Figures 3 and 4 show the confusion matrix for both D_1 and D_2 datasets that the proposed ensemble hard voting classifier correctly or erroneously predicts. Furthermore, a comparison with baseline algorithms has been provided. The dataset D_1 had 98 true positives, 28 false positives, a total of 29 false positives and false negatives. As a result, the confusion matrix makes the prediction clearer. For the dataset, D_2 34 was the correct answer and 64 was the false positive, resulting in a total of 6 false positives and false positives. The ROC curve shows a compromise between sensitivity (or TPR) and specificity (1-FPR). A classifier with a curve near the upper left corner performs better than a classifier with a far curve. As a baseline, the random classifier should give diagonal points (FPR = TPR). As the curve in ROC space approaches 45 degrees diagonal, the accuracy of the test decreases. Comparison of machine learning models ROC (receiver operating characteristic) curves are shown in Figures 5 and 6. These are indicated by adjusting the true positive rate (TPR) for false positive rates created at different thresholds (FPR). According to the ROC curves for the datasets D_1 and D_2, the proposed model has a higher percentage of coverage of 75% (D_1) and 92% (D_2). Although RF covered 93% of the area of the dataset D_2.

The proposed approach has been shown to achieve better results when compared with the state-of-the-art methods in terms of accuracy, precision, recall, F1 score, and area under the curve (AUC). This was determined by comparing the proposed method with the methods that are currently considered to be state-of-the-art. We are now in a position to assert that the model that we have proposed, which is essentially an ensemble method, has the potential to be applied in an effective way for the purpose of forecasting diabetes at an early stage.

4 Conclusion and Future Works

In disease diagnosis, machine learning approaches are useful. The capacity to detect diabetes early plays a critical role in determining treatment options for patients. In this research, the accuracy of a few existing ways to classify diabetic patients for medical diagnosis is addressed. The expressions of accuracy have been found to have a classification difficulty. The existing systems rely on a single classifier to predict diabetes classifications, which could lead to inaccuracies. As a result, we proposed a hard voting classifier model built on a combination of the LR, DT, and SVM machine learning algorithms. The accuracy, precision, recall, F1-score, and AUC of this model were assessed.

The results from the experiment indicated that all of the models performed well. Proposed model has the highest accuracy and precision of 94.24% and 81.12% on datasets D1 and D2 respectively. Individual models like KNN, LR, SVM, and RF are demonstrably less accurate than the proposed method. When it came to Recall, the highest recall values were obtained by SVM and RF, which were 100% and 95.53% for dataset D1, and the proposed model achieved the highest recall value of 81.12%, for dataset D2. In case of F1 score proposed model achieved a maximum F1 score of 94.24% and 81.12% for dataset D1 and D2 respectively. Because our dataset is an example of an unbalanced dataset, the F1 score offers a better understanding of our models' performance. The F1 score achieves a good mix of precision and Recall. According to the ROC curves for the datasets D1 and D2, the proposed model has a higher percentage of coverage of AUC of 75% (D1) and 92% (D2). AUC of this magnitude implies that proposed model is a trustworthy model. Finally, we may conclude that, among all the classifiers tested in this work, proposed model is the best for predicting diabetes.

The performance of the suggested model is yet to be investigated on various types of datasets because this work only assesses the performance of models on two datasets, namely the Early Stage Diabetes Risk Prediction Dataset and the PIMA Indian Diabetes Dataset. Future research will concentrate on using the suggested model on different datasets or combining different datasets with novel feature selection-based techniques. In order to further enhance the performance, we will also try to build prediction models based on other multi-criteria decision-making techniques, such as the Analytic Hierarchy Process (AHP).

5 Data Availability Statement

https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

https://www.kaggle.com/datasets/ishandutta/early-stage-diabetes-risk-prediction-dataset

References

1) Arjun P, Verma J. Methods for detection of Diabetes Mellitus using Machine Learning Techniques. *Journal of Multidisciplinary Engineering Science and Technology (JMEST)*. 2020;7(11):12948–12956.

2) Azbeg K, Boudhane M, Ouchetto O, Andaloussi SJ. Diabetes emergency cases identification based on a statistical predictive model. *Journal of Big Data*. 2022;9(1):31–31. Available from: https://doi.org/10.1186/s40537-022-00582-7.

3) Saeedi P, Petersohn I, Salpea P, Malanda B, Karuranga S, Unwin N, et al. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition. *Diabetes Research and Clinical Practice*. 2019;157:107843–107843. Available from: https://doi.org/10.1016/j.diabres.2019.107843.

4) Merad-Boudia HN, Dali-Sahi M, Kachekouche Y, Dennouni-Medjati N. Hematologic disorders during essential hypertension. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*. 2019;13(2):1575–1579. Available from: https://doi.org/10.1016/j.dsx.2019.03.011.

5) Mushtaq Z, Ramzan MF, Ali S, Baseer S, Samad A, Husnain M. Voting Classification-Based Diabetes Mellitus Prediction Using Hypertuned Machine-Learning Techniques. *Mobile Information Systems*. 2022;2022:1–16. Available from: https://doi.org/10.1155/2022/6521532.

6) Dhande B, Bamble K, Chavan S, Maktum T. Diabetes & Heart Disease Prediction Using Machine Learning. ITM Web of Conferences - ICACC-2022. 2022;44. Available from: https://doi.org/10.1051/itmconf/20224403057.

7) Morgan-Benita JA, Galván-Tejada CE, Cruz M, Galván-Tejada JI, Gamboa-Rosales H, Arceo-Olague JG, et al. Hard Voting Ensemble Approach for the Detection of Type 2 Diabetes in Mexican Population with Non-Glucose Related Features. *Healthcare*. 2022;10(8):1362–1362. Available from: https://doi.org/10.3390/healthcare10081362.

8) Kopitar L, Kocbek P, Cilar L, Sheikh A, Stiglic G. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Scientific Reports*. 2020;10(1):11981–11981. Available from: https://doi.org/10.1038/s41598-020-68771-z.

9) Tigga NP, Garg S. Predicting Type 2 Diabetes Using Logistic Regression. In: Lecture Notes in Electrical Engineering. Springer Singapore. 2021;p. 491–500. Available from: https://doi.org/10.1007/978-981-15-5546-6_42.

10) Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting Diabetes Mellitus With Machine Learning Techniques. *Frontiers in Genetics*. 2018;9:515–515. Available from: https://doi.org/10.3389/fgene.2018.00515.

11) Dua D, Graff C, Uci. UCI Machine Learning Repository (2019). Irvine, CA: University of California, School of Information and Computer Science. 2019. Available from: http://archive.ics.uci.edu/ml.

12) Islam MMF, Ferdousi R, Rahman S, Bushra HY. Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques. *Computer Vision and Machine Intelligence in Medical Image Analysis*. 2020;p. 113–125. Available from: https://doi.org/10.1007/978-981-13-8798-2_12.

13) Alam TM, Iqbal MA, Ali Y, Wahab A, Ijaz S, Baig TI, et al. A model for early prediction of diabetes. *Informatics in Medicine Unlocked*. 2019;16:100204–100204. Available from: https://doi.org/10.1016/j.imu.2019.100204.

14) Abidin NZ, Ritahani A, A N. Performance Analysis of Machine Learning Algorithms for Missing Value Imputation. *International Journal of Advanced Computer Science and Applications*. 2018;9(6):442–447. Available from: https://doi.org/10.14569/IJACSA.2018.090660.

15) Neamah M, Wahhab. Utilizing the Logistic Regression Model in Analyzing the Categorical Data of Economic Effects. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*. 2021;12:638–646. Available from: https://doi.org/10.17762/turcomat.v12i4.547.

16) Charbuty B, Abdulazeez A. Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends*. 2021;2(01):20–28. Available from: https://doi.org/10.38094/jastt20165.

17) Fregoso-Aparicio L, Noguez J, Montesinos L, García-García JA. Machine learning and deep learning predictive models for type 2 diabetes: a systematic review. *Diabetology & Metabolic Syndrome*. 2021;13(1):148–148. Available from: https://doi.org/10.1186/s13098-021-00767-9.

18) Muhammad LJ, Algehyne EA, Usman SS. Predictive Supervised Machine Learning Models for Diabetes Mellitus. *SN Computer Science*. 2020;1(5):1–10. Available from: https://doi.org/10.1007/s42979-020-00250-8.

19) Kumari S, Kumar D, Mittal M. An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. *International Journal of Cognitive Computing in Engineering*. 2021;2:40–46. Available from: https://doi.org/10.1016/j.ijcce.2021.01.001.

20) Fauzi MA, Bours P. Ensemble Method for Sexual Predators Identification in Online Chats. *2020 8th International Workshop on Biometrics and Forensics (IWBF)*. 2020. Available from: https://doi.org/10.1109/IWBF49977.2020.9107945.