PREPRINT

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

An Improved Ensemble Machine Learning Approach for Diabetes Diagnosis

Mohanad Mohammed Rashid[1], Omar Mahmood Yaseen[2], Rana Riyadh Saeed[1] and Maher Talal Alasaady[3]*

[1]*Department of Radiology Techniques, Northern Technical University, Mosul 41001, Iraq*
[2]*Administrative and Financial Department, Ministry of Higher Education and Scientific Research, Baghdad 10001, Iraq*
[3]*Computer Systems Department, Northern Technical University, Mosul 41001, Iraq*

ABSTRACT

Diabetes is recognized as one of the most detrimental diseases worldwide, characterized by elevated levels of blood glucose stemming from either insulin deficiency or decreased insulin efficacy. Early diagnosis of diabetes enables patients to initiate treatment promptly, thereby minimizing or eliminating the risk of severe complications. Although years of research in computational diagnosis have demonstrated that machine learning offers a robust methodology for predicting diabetes, existing models leave considerable room for improvement in terms of accuracy. This paper proposes an improved ensemble machine learning approach using multiple classifiers for diabetes diagnosis based on the Pima Indians Diabetes Dataset (PIDD). The proposed ensemble voting classifier amalgamates five machine learning algorithms: Decision Tree (DT), Logistic Regression (LR), K-Nearest Neighbor (KNN), Random Forests (RF), and XGBoost. We obtained the individual model accuracies and used the ensemble method to improve accuracy. The proposed approach uses a pre-processing stage of standardization and imputation and applies the Local Outlier Factor (LOF) to remove data anomalies. The model was evaluated using sensitivity, specificity, and accuracy criteria. With a reported accuracy of 81%, the proposed approach shows promise compared to prior classification techniques.

Keywords: Diabetes diagnosis, ensemble learning, machine learning, PIDD, soft voting

ARTICLE INFO

Article history:
Received: 15 August 2023
Accepted: 20 November 2023
Published: 04 April 2024

DOI: https://doi.org/10.47836/pjst.32.3.19

E-mail addresses:
mohanad.rashid@ntu.edu.iq (Mohanad Mohammed Rashid)
omar.yaseen1987777@gmail.com (Omar Mahmood Yaseen)
ranasaeed_1987@ntu.edu.iq (Rana Riyadh Saeed)
maher.alasaady@ntu.edu.iq (Maher Talal Alasaady)
* Corresponding author

ISSN: 0128-7680
e-ISSN: 2231-8526

INTRODUCTION

Diabetes is a chronic illness that affects millions of individuals worldwide and may lead to major health complications such as cardiovascular disease, stroke, and kidney failure (Chen et al., 2016). Patients with diabetes are divided into two distinct groups: Types 1 and 2. Patients with Type 1 diabetes are dependent on insulin for disease management. Patients with Type 2 diabetes do not require insulin to control the disease. According to research by the World Health Organization (WHO), more than ninety percent of people with this illness have Type 2 diabetes (WHO, 2014). Diabetes is associated with several negative effects, such as an increased chance of blindness, hypertension, kidney damage, and cardiovascular disease (Centers for Disease Control and Prevention, 2011). However, quick treatment may be started by people who get an early diagnosis of diabetes, lowering or even eliminating the possibility of negative outcomes.

Since diabetes has become one of the most common causes of severe illnesses, an expert system should be established and used to identify this condition. Machine Learning (ML) methods for developing autonomous diagnostic systems for various health disorders have been deemed beneficial (Saeed et al., 2022). Even though several ways have been presented to detect diabetes, the accuracy of different machine-learning algorithms is not exceptionally high (Barik et al., 2021). Previous efforts to enhance the predictive accuracy of these systems have frequently encountered challenges (Mirzajani & Salimi, 2018). In addition, the algorithms used to diagnose diabetes often come across data that is imprecise, missing, erroneous, or inconsistent (Swapna et al., 2018). The success of the model is dependent on the correctness of the diabetes data; thus, the researcher must offer precise data to the classifier to guarantee accurate illness prediction (Alasaady et al., 2022). Data pre-processing allows the construction of a highly accurate, robust classification model (Alasaady et al., 2019).

Ensemble learning is a method for ML in which many models are learned and integrated to enhance the system's overall performance and predictive ability (Khairan et al., 2023; Kunwar & Timalsina, 2021). Individual models, called base learners, are trained on separate subsets of data or with different techniques. Their outputs are merged using a predetermined way to give a final prediction. Ensemble learning may make ML models more accurate, stable, and generalizable. Voting, weighted averaging, and arithmetic mean are common strategies for mixing the outputs of the base learners (Kumari et al., 2021).

This research aims to detect diabetes using an ensemble approach to classify diabetes using a soft voting classifier. The techniques include Decision Tree (DT), Logistic Regression (LR), K-nearest neighbor (KNN), Randon Forest (RF), and eXtreme Gradient Boosting (XGBoost). In addition, the Pima Indians Diabetes Dataset (PIDD) undergoes many preparatory procedures to improve classification accuracy. The assessment processes use measures for sensitivity, specificity, and accuracy.

LITERATURE REVIEW

This literature review explores previous research on diabetes diagnosis using machine-learning approaches, specifically focusing on an ensemble machine-learning approach based on the PIDD dataset.

Kavakiotis et al. (2017) conducted a systematic review of 104 studies. They identified decision trees, neural networks, and support vector machines as the most used ML algorithms in diabetes research. Ensemble methods, such as bagging and boosting, have also been applied to a diabetes diagnosis. Bagging combines multiple models to reduce variance and improve accuracy (Breiman, 1996). Conversely, boosting involves iteratively training weak models and integrating them into robust ones (Fernández-Delgado et al., 2014).

Qin (2022) presents a diabetes prediction model utilizing ensemble learning techniques. The model incorporates LR KNN, DT, Gaussian Naïve Bayes (GNB), and Support Vector Machines (SVM). The initial four algorithms with low correlation are designed as fundamental learners, which are subsequently incorporated into a meta-learner SVM to establish an integrated learning model. The experiment was conducted on the PIDD. The level of accuracy achieved was 81%.

Atif et al. (2022) proposed a hard voting classifier-based ensemble learning approach. Both the Early-Stage Diabetes Risk Prediction Dataset and the PIDD dataset were put to the test. LR, DT, and SVM are the three ML techniques combined in the proposed ensemble hard voting classifier. The suggested ensemble technique achieves 81% accuracy on the PIDD diabetes dataset.

In Noor et al. (2021), various machine learning techniques were employed for the diagnosis of diabetes mellitus, including individual algorithms as well as ensemble approaches. Methods such as adaptive boosting via AdaboostM1, bagging, and hybrid classifiers that combine Random Forest with other base classifiers were investigated, along with the standalone Random Forest algorithm. The study selected an optimal diabetes classification model based on its accuracy and performance metrics. To improve the quality of the data inputted into the supervised learning models, data pre-processing methods such as Synthetic Minority Over-sampling Technique (SMOTE) were implemented to counteract data imbalance and eliminate missing values. The study concluded that the most effective diabetes classification model utilized a hybrid classifier combining Random Forest and Bayes Net, achieving an accuracy rate of 83%.

Kumari et al. (2021) employed an ensemble approach, which involved the combination of three ML algorithms: RF, LR, and Naïve Bayes (NB). The experimentation involved the utilization of two datasets: the PIDD dataset and the breast cancer dataset. A comparative analysis of the proposed methodology and conventional ML algorithms was conducted using both datasets. The ensemble approach that has been proposed demonstrates the highest level of accuracy, achieving a value of 79%, when applied to the PIDD dataset.

Kunwar and Timalsina (2021) constructed an ensemble model for the classification of diabetes. The model incorporates various ML algorithms: LR, SVM, NB, and DT. The proposed ensemble method combines the base classifiers using the probability assigned to each classifier. It is done to determine the final result by calculating the statistical mode of the output. The empirical findings support that hybrid approaches exhibit greater implicitness than individual classifiers' isolated utilization. The level of accuracy achieved was 81%.

Agrawal et al. (2021) used an ensemble model by employing the voting classifier, which incorporated the RF, AdaBoost, and DT as contributing models. The PIDD dataset is utilized in the experiment. The ensemble models demonstrated superior accuracy compared to the individual models and reduced the occurrence of False Negatives. The level of accuracy achieved was 77%.

Singh and Singh (2020) propose developing a stacking-based evolutionary ensemble learning system to predict diabetes. The PIDD dataset is employed. In selecting a base learner, a multi-objective optimization algorithm is employed to effectively balance the objectives of maximizing classification accuracy and minimizing ensemble complexity to achieve this objective. The level of accuracy achieved was 83%.

Soni and Varma (2020) employed ensemble techniques and ML algorithms to predict diabetes, which includes GB, RF, DT, SVM, KNN, and LR. The results demonstrate that RF outperformed other ML methods in terms of accuracy. The precision was 79%.

Akyol and Şen (2018) used an ensemble method to diagnose diabetes. There are two main stages to this investigation. The feature selection or weighting approaches are examined in the first phase to determine the best qualities for this condition. The performance of the ensemble learning techniques AdaBoost, Gradient Boosted Trees (GBT), and RF are assessed in the following stage. According to test results, the Stability Selection method and AdaBoost learning algorithm's prediction accuracy is somewhat higher than that of other algorithms, which is 73%.

Li (2014) proposes a methodology that integrates three distinct classifiers, namely SVM, Artificial Neural Network (ANN), and NB, to diagnose diabetes. He has proposed a voting classifier technique called the weight-adjusted voting technique. The proposed methodology entails the modification of the weight assigned to each classifier, considering their performance and past track record in accurately predicting outcomes. After being implemented on the PIDD dataset, this method demonstrates a prediction accuracy of 77%.

In summary, the studies that have been identified suggest that individual ML and basic ensemble techniques may not attain satisfactory accuracy when diagnosing diabetes using the PIDD dataset. Hence, developing an enhanced ensemble machine learning methodology is imperative to augment the precision of diabetes diagnosis. Table 1 presents a comparative analysis of previous research endeavors, highlighting the respective levels of accuracy attained.

Table 1
Comparison of accuracies of related work

Author	Year	Method	Accuracy
Qin	2022	LR, KNN, DT, GNB, and SVM	81%
Atif et al.	2022	LR, DT, and SVM	81%
Kumari et al.	2021	RF, LR, and NB	79%
Noor et al.	2021	RF, LR, MLP, NB, AdaboostM1	83%
Kunwar and Timalsina	2021	LR, SVM, NB, and DT	81%
Agrawal et al.	2021	RF, AdaBoost, and DT	77%
Singh and Singh	2020	SVM, RF, and KNN	83%
Soni and Varma	2020	GB, RF, DT, SVM, KNN, and LR.	79%
Akyol & Şen	2018	AdaBoost, GBT, and RF	73%
Lin Li et al.	2014	SVM, ANN, and NB	77%

METHOD

This paper proposes an improved ensemble approach for predicting diabetes to get accurate classifications of patients with Type 2 diabetes based on PIDD. DT, LR, KNN, RF, and XGBoost algorithms have been ensembled. The model was tested using PIDD and implemented using Python 3.10.9. Standardization, imputation, and anomaly detection using the LOF technique are carried out at the pre-processing stage. Figure 1 illustrates the architecture and the activities carried out by each architectural component throughout the diagnosis of diabetes. The specifics of this diagram are as follows:

- The PIDD dataset has been used to analyze and test the proposed approach.
- The data pre-processing stage has been done to transform raw data into a format that can be understood. Standardization, imputation, and anomaly detection have been used in this stage.

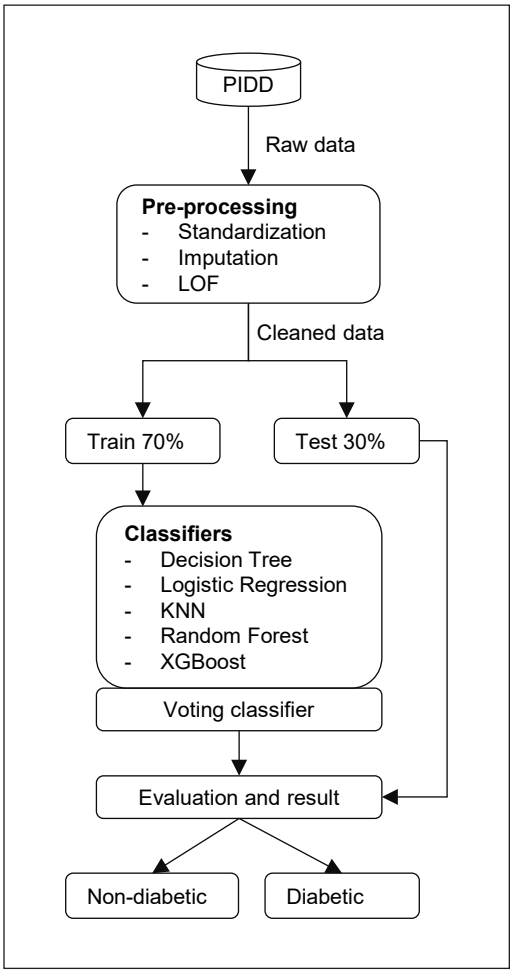


Figure 1. The main architecture

- The cleaned data has been split into train and test sets.
- Five machine learning models have been applied to the train set.
- Ensemble learning creates a hybrid model using the soft voting classifier. Finally, the trained algorithms and ensemble are applied to the test set and the evaluation.

PIDD Dataset

The dataset frequently employed for evaluating the effectiveness of diabetes diagnostic algorithms is the Pima Indians Diabetes Dataset (PIDD). The Pima Indians, a group of Native Americans who reside in the Arizona region of the USA, have the world's highest prevalence of Type 2 diabetes. All the patients in this dataset are women over the age of 21. There was a total of 768 occurrences in the data set. The dataset is separated into two categories, diabetes and health, designated by 1 and 0, respectively. There are 268 examples in class 1, whereas there are 500 instances in class 0. Eight attributes are present: Number of pregnancies, Levels of plasma glucose, Heart rate (mm Hg), The triceps' skinfold thickness (mm), The amount of serum insulin (mu U/ml), Body Mass Index, (BMI), Pre-degree function for diabetes, and Age. The features of PIDD are shown in Table 2. PIDD is frequently used to test new machine learning models, particularly in binary classification. The task at hand is to predict the onset of diabetes based on several medical predictor variables.

Although the dataset has several limitations regarding its representativeness for a diverse global population, it offers numerous benefits for academic applications. The dataset provides a straightforward way to evaluate the performance of algorithms due to its relatively clean and complete nature, making it suitable for academic and introductory applications (Ganesh & Sripriya, 2020). PIDD, originally collected to study the high prevalence of diabetes in the Pima Indian community in Arizona, USA, has some biases that we should mention. It is ethnically specific to Pima Indian women, excludes men, focuses on individuals over the age of 21, and is geographically restricted to Arizona. In healthcare, the data used to train models is one of the privacy concerns. Systems may collect data in a

Table 2
Pima Indians Diabetes Dataset

Attribute	Mean	SD	Min/max	Missing Value
Pregnant	3.8	3.4	1/17	0
Glucose	120.9	32	56/197	5
DBP	69.1	19.4	24/110	35
TSFT	20.5	16	7/52	227
INS	79.8	115.2	15/846	374
BMI	32	7.9	18.2/57.3	11
DPF	0.5	0.3	0.0850/2.3290	0
Age	33.2	11.8	21/81	0

way that violates privacy, such as scraping personal information or gathering information without consent. PIDD ethical consideration was mentioned in detail in (https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database).

Data Pre-processing

Preparing raw data for analysis by cleaning and converting it into a suitable format is known as data pre-processing (Han et al., 2022). It is an essential step in the data analysis pipeline as it can significantly impact the accuracy of the results obtained from data analysis. The pre-processing techniques for the proposed approach are standardization, imputation, and anomaly detection.

Standardization. Data standardization is essential to data preparation, which entails putting the data into a consistent and uniform format (Shanker et al., 1996). Standardization is especially crucial when dealing with data from many sources, which may employ different units of measurement, scales, and conventions (Berner & Judge, 2019). In data standardization, the data are rescaled with a mean of zero and a standard deviation of one. This procedure facilitates data comparison and analysis and increases the accuracy and dependability of the results received through data analysis. In our proposed approach, we utilized standard scalar as the standardization technique, as shown in Equation 1, where μ is the mean, and σ is the standard deviation.

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

Imputation. Data imputation is a method used to replace missing data values in a dataset with estimated values based on existing data (Gelman & Hill, 2006). Imputation may be especially valuable in cases where missing data is prevalent since it permits the use of accurate data for analysis, which can increase the accuracy and dependability of the findings. There are numerous methods for data imputation, such as mean imputation, regression imputation, and multiple imputation (Buuren, 2012). Mean imputation entails substituting missing values with the mean of the existing data.

Local Outlier Factor. In data analysis, the Local Outlier Factor (LOF) is a kind of unsupervised technique used to find outliers (Breunig et al., 2000). The LOF method evaluates each data point's local density and compares it to the local densities of its neighbors. Potential outliers are points that have a substantially lower density than their neighbors. The relative density of a data point X with k neighbors is expressed by Equation 2, where N = Average density of all data points in the neighborhood. The average distance between the k nearest data points and the X density have a proportional connection.

$$X = \frac{Density\ of\ X}{N} \quad (2)$$

Classification Models

In ensemble machine learning, the selection of models to combine is governed by several key criteria. Choosing base models that are diverse, competent, and ideally independent is crucial to ensure that the ensemble captures a wide range of features and characteristics of the data. The overarching objective is to form an ensemble that enhances generalization and robustness by effectively leveraging the strengths and mitigating the weaknesses of its constituent models (Caruana et al., 2004). The proposed approach has included several ML methods, including decision trees, logistic regression, KNN, random forests, and XGBoost classifiers. Combining the techniques above with a soft voting classifier increases accuracy.

Decision Tree. A supervised learning approach for classification and regression analysis is called a Decision Tree (DT) (Breiman et al., 2017). Each leaf node represents a class name or a numerical value, and each inside node reflects a choice based on a particular trait. The decision tree method is beneficial for studying complex relationships between variables and identifying the essential properties of a dataset (Breiman, 2001). Decision trees have been used in various sectors, including finance, health, and environmental science, as well as ensemble approaches like Random Forest and Boosting.

Logistic Regression. Logistic Regression (LR) is a common statistical model for binary classification issues with a categorical response variable (Hosmer et al., 2013). Based on one or more predictor variables, the logistic regression model evaluates the likelihood of a binary result. The model's output is a logistic function of the input variables, which maps the input space to the probability space of the binary development. Numerous fields extensively use logistic regression, including medical diagnosis, social sciences, and finance (Agresti, 2015).

K-Nearest Neighbors. In machine learning and data mining, K-Nearest Neighbors (KNN) is a well-known non-parametric classification and regression technique (Cover & Hart, 1967). The KNN algorithm determines a prediction based on the class or regression value of most of the k nearest neighbors of a particular test instance in the training dataset. KNN is a simple and versatile technique applicable to various applications, although it may be computationally costly for big datasets and high-dimensional feature spaces. KNN has been implemented in several domains, including bioinformatics, image processing, and recommendation systems (El Houby et al., 2017).

Random Forest. Random Forest (RF) is a well-known ensemble machine learning technique that blends a variety of decision trees to boost prediction accuracy and model dependability (Breiman, 2001). It produces several trees and applies the bootstrap technique to each tree in the training data set. Every tree in the forest receives method input during classification, and each tree casts a unique vote for that class. The RF chooses the class with the most significant votes (Mansour & Schain, 2001). Various strategies can be employed to avoid overfitting in ensemble learning, including using RF, which generates random subsets of data and average predictions to negate individual model overfits.

XGBoost. XGBoost (eXtreme Gradient Boosting) is a gradient boosting method that has attracted much interest recently because of its excellent accuracy and scalability in extensive machine learning applications (Chen & Guestrin, 2016). With the help of a tree-based model, XGBoost is an improved version of gradient boosting that iteratively adds weak learners to the ensemble to reduce a particular loss function. The approach uses various regularization techniques to avoid overfitting and boost generalization performance, including shrinkage, subsampling, and pruning. XGBoost is used in many industries, including banking, natural language processing, and computer vision; XGBoost has won several machine-learning contests because of its outstanding performance (Ke et al., 2017).

Proposed Ensemble Voting Classifier

The ensemble is a strategy whose meta-algorithms integrate many machine learning approaches into a single optimal predictive model to reduce variance, bias or improve predictions. This strategy improves the prediction performance over a single model. Ensembling techniques include bagging, boosting, stacking, and voting (Prema et al., 2019). On the PIDD dataset, we have applied a voting-based ensembling technique. The vote-based ensembling approach mixes comparable or conceptually distinct machine learning classifiers for classification using majority or plurality voting. Utilize a voting mechanism to determine the best option among several alternatives. As a result, multiple classifiers can choose from a variety of options. There are two types of voting, hard and soft voting (Mahabub, 2019):

- **Hard Voting:** Hard voting is the most straightforward majority voting. Here, the results of all classifiers are treated equally. Votes are only computed using their median value. Each classifier C_j votes with the majority to select the class label Y, hard voting representation in Equation 3.

$$Y = mode\{C1(x), C2(x), \ldots, Cm(x) \quad [i = 1,2,\ldots,m] \tag{3}$$

- **Soft Voting:** Soft voting predicts the class using the classifier's projected probability (p.). Soft voting representation in Equation 4. Where W_j is the maximum load that the j^{th} classifier can handle.

$$Y = argmax_i \sum_{j=1}^{m} W_j P_{ij} \quad (4)$$

The proposed model has ensembled decision trees, logistic regression, KNN, random forest, and XGBoost classifiers. A classifier based on soft voting has been employed. Using a voting aggregator and a soft voting strategy, each model generates its forecasts, and the majority vote results determine the final prediction. Figure 2 depicts the algorithm behind the suggested technique.

```
Procedure PreProcess (PIDD)
PIDD.StandardScalar()
PIDD.Imputation()
PIDD.LOC()
Return PIDD

Procedure Split_data(PIDD)
   Train_set, Test_set=split(PIDD, parameters)
   Return Train_set, Test_set
CI=Decision_tree(Train_set, Label, Test_set)
C2=Logistic_regression(Train_set, Lable,
Test_set)
C3=KNN(Train_set, Label, Test_set)
C4=Randon_forest(Train_set, Lable, Test_set)
C5=XGBoost(Train_set, Label, Test_set)

Procedure Voting(C1, C2, C3, C4, C5,
voting=soft)
   Voting.fit(Train_set, Label)
   Voting.predict(Test_set)
```

Figure 2. The proposed approach algorithm

Evaluation

Model evaluation is an essential process that uses some metrics to evaluate the model's performance. The proposed ensemble approach for diabetes prediction has been tested in two experimental settings: (1) an assessment of the proposed approach's data preparation practices and (2) an evaluation of the effectiveness of machine learning classifiers. Sensitivity, specificity, and accuracy evaluate each situation's efficiency.

Sensitivity. The term sensitivity is used in biostatistics. It can distinguish between positive and negative patients in a test. This statistic represents the percentage of diabetic patients appropriately recognized as such. Equation 5 calculates sensitivity, where (TP) represents the true positives and (FN) the false negatives.

$$Sensitivity = \frac{TP}{(TP+FN)} \quad (5)$$

Specificity. Specificity is the process of differentiating between actual sturdy and general sturdy. It is the proportion of people classed as non-diabetic who do not have diabetes, the negative instances. It is the ratio of true negatives to the total number of true negatives and false positives (Equation 6).

$$Specificity = \frac{TN}{(TN+FP)} \quad (6)$$

Accuracy. The test findings will likely be accurate when the correct sensitivity and specificity are combined in a single measure. The accuracy is calculated by dividing the number of correct predictions by the total number of predictions. Equation 7 is used to calculate accuracy.

$$Accuracy = \frac{(TN + TP)}{(TN + TP + FN + FP)} \quad (7)$$

RESULTS AND DISCUSSION

The values in the dataset were first scaled, meaning that each value fits within a range (0 and 1). The StandardScalar method is used to standardize Equation 1. This modification helps offset the negative consequences of the prevalence of specific traits, especially undesired ones with more comprehensive value ranges. As numerous features have a value of zero, data imputation is then performed (for example, a blood pressure reading of 0 seems improbable, yet it is the lowest possible result).

As a direct consequence of this, incorrect information is provided. We may substitute these data with the median, as we cannot disregard these data. We have opted to impute since our dataset is small. The fact that pregnancy has a zero value rather than a missing value makes it an example of a feature that should not be imputed. Lastly, anomaly data is detected by utilizing the LOF technique (Equation 2).

The ensemble approach was implemented using Python 3.10.9. The PIDD dataset was used in the experiments. There are 768 rows and eight features in the dataset. The dataset was randomly split into two sets; the training set consists of 537 records (about 70% of the dataset), while the test set has 230 records (or 30 percent). The model is "trained" using the training data and then "tested" using the data to ensure accuracy and effectiveness.

The proposed approach employs five ML models: DT, LR, KNN, RF, and XGBoost. A comparison of traditional ML techniques for classifying diabetes as positive or negative has been conducted. It has been performed to compare and analyze the accuracy of conventional algorithms. Positive and negative classes comprise the PIDD dataset, which has been used for testing. Table 3 compares the outputs of several ML models using the PIDD dataset. Compared to previous machine learning methods. The proposed ensemble approach had the greatest accuracy, sensitivity, and specificity values of 64%, 74%, and 81%, respectively (Table 3). The results for the individual models in Table 3 also indicate the data quality after performing the pre-processing process.

It is clear from referring to Table 3 that none of these machine learning methods do very well on the dataset that has been presented since they both have an accuracy of less

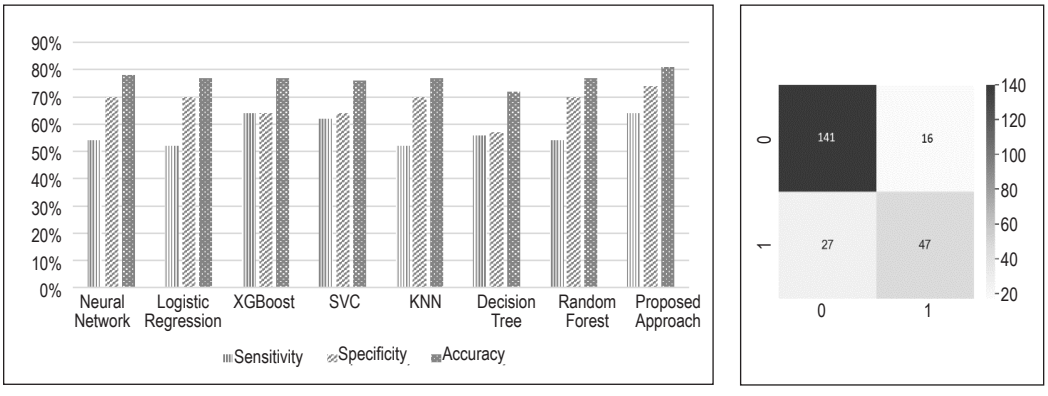
than 80. Considering the performance of the suggested method reveals that it exceeds the other machine learning algorithms in terms of accuracy by an 81% margin. In contrast, in terms of sensitivity, it was equal to XGBoost and close to SVC by 64%, and in specificity, it also overcame the rest by 74%. Figure 3(a) shows the comparative analysis graph of the different algorithms on the PIDD datasets. The dataset had 49 false positives and false negatives, in addition to 141 genuine positives and 43 false positives. Consequently, the confusion matrix has helped us comprehend our forecasts better.

Figure 3(b) depicts the confusion matrix for diabetic patients, demonstrating whether the recommended ensemble approach made accurate or inaccurate predictions. The dataset had 141 true positives, 43 false positives, and 49 false positives and false negatives. Therefore, the confusion matrix has improved our understanding of our predictions.

In contrast to earlier research, our findings indicate enhanced accuracy, with a few exceptions. Notably, Nour's study (Noor et al., 2021) employed specific data pre-processing techniques, including the Synthetic Minority Over-sampling Technique (SMOTE), to

Table 3
Comparison of accuracies with conventional machine learning algorithms

Classifier	Sensitivity	Specificity	Accuracy
Neural Network	54%	70%	78%
Logistic Regression	52%	70%	77%
XGBoost	64%	64%	77%
SVC	62%	64%	76%
KNN	52%	70%	77%
Decision Tree	56%	57%	72%
Random Forest	54%	70%	77%
Proposed Approach	**64%**	**74%**	**81%**



(a) (b)

Figure 3. (a) Graphical comparison of accuracies; and (b) confusion matrix of diabetes patients

address data imbalance and remove missing data points. Given the limited data, we chose not to adopt these methods in our study. Eliminating data points could have constrained the model's learning capacity, potentially leading to overfitting.

The performance of the proposed ensemble method is attributed to several key factors. The diverse base algorithms collectively capture different data aspects, enhancing robustness and generalizability. The quality and relevance of the PIDD dataset, coupled with optimized hyperparameters, contribute to the model's accuracy. While the ensemble approach improves performance metrics like accuracy, specificity, and sensitivity, it balances computational efficiency without significant overfitting, although it may compromise interpretability.

CONCLUSION

This research aimed to develop a robust and precise algorithmic framework for predicting diabetes in patients. The study deployed an experimental design hinged on five prominent machine learning algorithms To realize this goal: Decision Trees, Logistic Regression, Random Forest, K-Nearest Neighbors (KNN), and Extreme Gradient Boosting (XGBoost). The Pima Indians Diabetes Dataset (PIDD) was utilized for the investigation and subjected to rigorous pre-processing procedures, including standardization, imputation of missing values, and anomaly detection via the Local Outlier Factor (LOF) methodology. These pre-processing techniques were pivotal in optimizing the dataset for machine learning applications, eliminating erroneous outcomes and enhancing the model's interpretability.

The ensemble approach, which employed soft voting classifiers, achieved an accuracy rate of 81%. While this level of accuracy is noteworthy and enriches the existing literature, there remains room for further improvement. Several directions for future research emerge from the findings and limitations of this study. These include the potential application of deep learning models to further improve prediction accuracy and the exploration of hyperparameter tuning techniques about the proposed model. Future research may also benefit from using real-world data sets for further validation. In subsequent phases, the model will be fine-tuned using empirical data. Should it exhibit robust performance across multiple evaluations, consideration may be given to its integration into clinical diagnostic procedures.

ACKNOWLEDGEMENT

The authors sincerely thank the administration leaders at Northern Technical University (NTU), Iraq. The authors would also like to thank all the professors who validated the data and results for their comments.

REFERENCES

Agrawal, K., Bhargav, G., & Spandana, E. (2021). Diabetes diagnosis prediction using ensemble approach. In V. Nath & J. K. Mandal (Eds.), *Proceedings of the Fourth International Conference on Microelectronics, Computing and Communication Systems: Lecture Notes in Electrical Engineering, vol 673* (pp. 799-813). Springer. https://doi.org/10.1007/978-981-15-5546-6_66

Agresti, A. (2015). *Foundations of linear and generalized linear models*. John Wiley & Sons

Akyol, K., & Şen, B. (2018). Diabetes mellitus data classification by cascading of feature selection methods and ensemble learning algorithms. *International Journal of Modern Education & Computer Science, 10*(6), 10-16. https://doi.org/10.5815/ijmecs.2018.06.02

Alasaady, M. T., Aris, T. N. M., Sharef, N. M., & Hamdan, H. (2022). A proposed approach for diabetes diagnosis using neuro-fuzzy technique. *Bulletin of Electrical Engineering and Informatics*, *11*(6), 3590–3597. https://doi.org/10.11591/eei.v11i6.4269

Alasaady, M. T., Saeed, M. G., & Faraj, K. H. (2019, February 13-14). *Evaluation and comparison framework for data modeling languages*. [Paper presentation]. 2nd International Conference on Electrical, Communication, Computer, Power and Control Engineering (ICECCPCE), Mosul, Iraq. https://doi.org/10.1109/ICECCPCE46549.2019.203750

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

Barik, S., Mohanty, S., Mohanty, S., & Singh, D. (2021). Analysis of prediction accuracy of diabetes using classifier and hybrid machine learning techniques. In D. Mishra, R. Buyya, P. Mohapatra & S. Patnaik (Eds.), *Intelligent and Cloud Computing* (pp. 399–409). Springer. https://doi.org/10.1007/978-981-15-6202-0_41

Berner, R., & Judge, K. (2019). *The Data Standardization Challenge* (Working Paper No. 438/2019). CIGI Press. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3323719

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24,* 123-140. https://doi.org/10.1007/BF00058655.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). *Classification and Regression Trees*. Routledge. https://doi.org/10.1201/9781315139470

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000, May 15-18). *LOF: Identifying density-based local outliers*. [Paper presentation] SIGMOD '00: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, Texas, USA. https://doi.org/10.1145/342009.335388

Buuren, S. V. (2012). *Flexible imputation of missing data*. CRC Press. https://doi.org/10.1201/b11826

Caruana, R., Niculescu-Mizil, A., Crew, G., & Ksikes, A. (2004, July 4-8). *Ensemble selection from libraries of model*s. [Paper presentation]. ICML '04: Proceedings of the Twenty-first International Conference on Machine Learning, New York, USA. https://doi.org/10.1145/1015330.1015432

Centers for Disease Control and Prevention (2011). National diabetes fact sheet: National estimates and general information on diabetes and prediabetes in the United States. *Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention*, *201*(1), 2568–2569.

Chen, R., Ovbiagele, B., & Feng, W. (2016). Diabetes and stroke: Epidemiology, pathophysiology, pharmaceuticals and outcomes. *American Journal of the Medical Sciences*, *351*(4), 380–386. https://doi.org/10.1016/j.amjms.2016.01.011

Chen, T., & Guestrin, C. (2016, August 13-17). *XGBoost: A scalable tree boosting system*. [Paper presentation]. KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, California, USA. https://doi.org/10.1145/2939672.2939785

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964

El Houby, E. M. F., Yassin, N. I. R., & Omran, S. (2017). A hybrid approach from ant colony optimization and K-nearest neighbor for classifying datasets using selected features. *Informatica*, *41,* 495–506.

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real-world classification problems? *The Journal of Machine Learning Research*, *15*(1), 3133–3181.

Ganesh, P. V. S., & Sripriya, P. (2020). A comparative review of prediction methods for pima indians diabetes dataset. In S. Smys, J. M. R. S. Tavares, V. E. Balas & A. M. Iliyasu (Eds.), *Computational Vision and Bio-Inspired Computing* (pp. 735–750). Springer. https://doi.org/10.1007/978-3-030-37218-7_83

Gelman, A., & Hill, J. (2006). *Data Analysis using Regression and Multilevel/Hierarchical Models*. Cambridge University Press. https://doi.org/10.1017/CBO9780511790942

Han, J., Pei, J., & Tong, H. (2022). *Data mining: Concepts and techniques*. Morgan Kaufmann.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley. https://doi.org/10.1002/9781118548387

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. https://doi.org/10.1016/j.csbj.2016.12.005

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (pp. 30). Curran Associates, Inc.

Khairan, H. E., Zubaidi, S. L., Muhsen, Y. R., & Al-Ansari, N. (2023). Parameter optimisation-based hybrid reference evapotranspiration prediction models: A systematic review of current implementations and future research directions. *Atmosphere, 14*(1), Article 77. https://doi.org/10.3390/atmos14010077

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

Kunwar, R., & Timalsina, A. K. (2021). An ensemble approach for the diagnosis of diabetes mellitus using multiple classifiers. *Proceedings of 9th IOE Graduate Conference*, *9*, 202-207.

Li, L. (2014, November 10-12). *Diagnosis of diabetes using a weight-adjusted voting approach*. [Paper presentation]. IEEE International Conference on Bioinformatics and Bioengineering, Florida, USA. https://doi.org/10.1109/BIBE.2014.27

Mahabub, A. (2019). A robust voting approach for diabetes prediction using traditional machine learning techniques. *SN Applied Sciences*, *1*(12), Article 1667. https://doi.org/10.1007/s42452-019-1759-7

Mansour, Y., & Schain, M. (2001). Learning with maximum-entropy distributions. *Machine Learning*, *45*(2), 123–145. https://doi.org/10.1023/A:1010950718922

Mirzajani, S. S., & Salimi, S. (2018). Prediction and diagnosis of diabetes by using data mining techniques. *Avicenna Journal of Medical Biochemistry*, *6*(1), 3–7. https://doi.org/10.15171/ajmb.2018.02

Noor, N. A. B. S., Elamvazuthi, I., & Yahya, N. (2021, July 13-15). *Classification of diabetes mellitus using ensemble algorithms*. [Paper presentation]. 8th International Conference on Intelligent and Advanced Systems (ICIAS), Kuching, Sarawak. https://doi.org/10.1109/ICIAS49414.2021.9642508

Prema, N. S., Varshith, V., & Yogeswar, J. (2019). Prediction of diabetes using ensemble techniques. *International Journal of Recent Technology and Engineering, 7*(6), 203-205.

Qin, L. (2022, September 23-25). *A prediction model of diabetes based on ensemble learning*. [Paper presentation] AIPR '22: Proceedings of the 2022 5th International Conference on Artificial Intelligence and Pattern Recognition, Xiamen China. https://doi.org/10.1145/3573942.3573949

Saeed, R. R., Yaseen, O. M., Rashid, M. M., & Ahmed, M. R. (2022, June 9-11). *Applications of machine learning in battling against novel COVID-19*. [Paper presentation]. International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Ankara, Turkey. https://doi.org/10.1109/HORA55278.2022.9799969

Shanker, M., Hu, M. Y., & Hung, M. S. (1996). Effect of data standardization on neural network training. *Omega*, *24*(4), 385–397. https://doi.org/10.1016/0305-0483(96)00010-2

Singh, N., & Singh, P. (2020). Stacking-based multi-objective evolutionary ensemble framework for prediction of diabetes mellitus. *Biocybernetics and Biomedical Engineering*, *40*(1), 1–22. https://doi.org/10.1016/j.bbe.2019.10.001

Soni, M., & Varma, S. (2020). Diabetes prediction using machine learning techniques. *International Journal of Engineering Research & Technology*, *9*(9), 921-925.

Swapna, G., Soman, K. P., & Vinayakumar, R. (2018). Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals. *Procedia Computer Science*, *132*, 1253–1262. https://doi.org/10.1016/j.procs.2018.05.041

WHO. (2014). *World diabetes statistics*. World Health Organization. http://www.who.int/diabetes/en/index.html