International Journal of Electrical and Computer Engineering (IJECE)
Vol. 15, No. 6, December 2025, pp. 5347~5359
ISSN: 2088-8708, DOI: 10.11591/ijece.v15i6.pp5347-5359

An ensemble machine learning based model for prediction and diagnosis of diabetes mellitus

Moataz Mohamed El Sherbiny[1], Asmaa Hamdy Rabie[2], Mohamed Gamal Abdel Fattah[1], Ali Elsherbiny Taki Eldin[3], Hossam El-Din Mostafa[1]
[1]Department of Electronics and Communication Engineering, Faculty of Engineering, Mansoura University, Mansoura, Egypt
[2]Department of Computer and Control Systems Engineering Science, Faculty of Engineering, Mansoura University, Mansoura, Egypt
[3]Department of Cyber Security, Faculty of Artificial Intelligence, Delta University, Gamasa, Egypt

Article Info

Article history:

Received Feb 22, 2025
Revised Jul 17, 2025
Accepted Sep 14, 2025

Keywords:

Classification
Diabetes mellitus
Ensemble
Machine learning
Performance measures

ABSTRACT

Diabetes mellitus (DM) is a chronic metabolic disorder that poses significant health risks and global economic burdens. Early prediction and accurate diagnosis are crucial for effective management and treatment. This study presents an ensemble machine learning-based model designed to predict and diagnose Diabetes Mellitus using clinical and demographic data. The proposed approach integrates multiple machine learning algorithms, including random forest (RF), extreme gradient boosting (XGB), and logistic regression (LR), to leverage their individual strengths and enhance the entire performance. The ensemble model was trained and validated on multiple comprehensive datasets. Performance measures demonstrate the robustness of proposed model and its reliability in distinguishing diabetic cases from non-diabetic cases after applying several preprocessing steps. This work ensures the capability of machine learning in advancing healthcare by providing efficient, data-driven tools for diabetes management, aiding clinicians in early diagnosis, and contributing to personalized treatment strategies. Comparative analysis against standalone models highlights the superior predictive capabilities of the ensemble approach. Results had shown that ensemble model achieved an accuracy of 96.88% and precision of 89.85% outperforming individual classifiers.

Corresponding Author:

Moataz Mohamed El Sherbiny
Department of Electronics and Communication Engineering, Faculty of Engineering, Mansoura University
Mansoura, DK 35516, Egypt
Email: moatazelsherbiny@mans.edu.eg

1. INTRODUCTION

Diabetes mellitus is one of the most prevalent chronic diseases and has become a major public health challenge [1]. Diabetes mellitus (DM) is a chronic metabolic disease characterized by elevated blood glucose levels, resulting from either insufficient insulin production or ineffective insulin utilization [2]. The prevalence of diabetes has reached alarming levels with projections indicating further growth in the coming decades. According to the International Diabetes Federation (IDF), 537 million adults aged 20–79 years were living with diabetes globally in 2021 [3]. This number is projected to rise to 643 million by 2030 and 783 million by 2045, reflecting a significant upward trend. In 2021, diabetes was estimated to be the cause of 6.7 million fatalities, indicating that one person dies every five seconds due to diabetes-related complications. Diabetes is considered a silent killer where the number of undiagnosed cases globally is nearly 240 million individuals accounting for 1 in 2 adults with diabetes [4]. Egypt ranked ninth globally in the number of diabetes cases, with 10.9 million adults living with the disease. Around 50% of diabetes cases in Egypt remain undiagnosed.

The disease is associated with severe complications [5], including cardiovascular disorders, renal failure, neuropathy, and retinopathy, which significantly impact the quality of life and increase mortality rates. Early diagnosis and effective management are, therefore, critical in mitigating the development of diabetes and its accompanying complications. Traditional diagnostic methods for diagnosis of diabetes [6], such as fasting plasma glucose (FPG), oral glucose tolerance tests (OGTT), and hemoglobin A1c (HbA1c) levels, are reliable but may be limited by cost, accessibility, and the need for laboratory infrastructure. Moreover, these methods often fail to predict the onset of diabetes in prediabetic individuals, emphasizing the need for innovative approaches to enhance early detection. In recent years, advances in machine learning (ML) have demonstrated significant potential in healthcare, offering data-driven solutions for disease prediction, diagnosis, and personalized treatment. Telemedicine has become a game-changing solution [7], where it improves healthcare accessibility by eliminating the need for in-person hospital visits through the utilization of digital communication technology that enables distant consultations.

Most of previous studies employed the Pima Indians Diabetes Dataset (PIDD). It is considered as one of the most well-known datasets in binary classification of diabetes using machine learning. Febrian *et al.* [8] applied two supervised machine learning algorithms on the PIDD. Train and test split were performed without cross validation. The results of K-nearest neighbor (KNN) were outperformed by naïve Bayes (NB) in both experiments. Authors compared results in terms of accuracy, recall as well as precision. NB achieved the highest accuracy of 78.52 %. Kangra and Singh [9] split data into training and testing using 10-fold cross-validation for preprocessing stage. Authors compared six supervised machine learning algorithms using three evaluation metrics which are accuracy, precision and recall. They NB, KNN, support vector machine (SVM), decision tree (DT), random forest (RF) and logistic regression (LR) on the PIDD indicating that SVM achieved highest accuracy score of 74.3% followed by LR which achieved 74%. Chang *et al.* [10] conducted three experiments on the PIDD. The first experiment showed that RF outperformed both DT and NB by achieving 79.57% and 89.4% in terms of accuracy and precision respectively. Authors applied feature selection of 3-factor of the entire dataset in the second experiment. NB reached accuracy of 79.13%, and F1-score of 84.71%. In their final experiment, authors utilized feature selection of 5-factor. However, accuracy went down to 77.83% by NB. Mushtaq *et al.* [11] employed a two-stage model selection methodology. LR, SVM, KNN, GB, NB and RF applied to determine the efficiency of prediction models. RF was found to be the best with accuracy of 80.7% after applying smote. The ensemble of the best 3 models yielded accuracy of 82% on original dataset and 81.7% on balanced dataset. Rawat *et al.* [12] assures the usefulness of data mining techniques to evaluate the unknown patterns on the PIDD. Authors proposed multiple techniques such as AdaBoost and Naïve Bayes for the analysis and prediction of DM patients. The results computed are found to be 79.69% classification accuracy by AdaBoost method. Barik *et al.* [13] used two machine learning algorithms on PIDD. In the case of RF, the prediction value was 71.9% but XGBoost yielded higher accuracy of 74.1%. Palimkar *et al.* [14] utilized multiple machine learning models on a questionnaire dataset such as LR, SVM, naïve Bayes and adaptive boosting (AdaBoost). Results were compared using 70%-30% training and testing accuracy respectively in addition to mean square error (MSE). LR achieved 93.59%, SVM yielded 94.23%, Gaussian NB 91.02% and AdaBoost 94.87 in terms of testing accuracy. Vocal biomarker prediction of disease has been employed in a variety of diseases, including COVID-19 detection, Parkinson's disease, pulmonary function, and coronary artery disease. Fagherazzi *et al.* [15] implemented their study on Colive study voice dataset. authors utilized three classifier algorithms such as logistic regression, support vector machine and multi-layer perceptron classifiers (MLP). Results indicated that MLP yielded the highest accuracy of 67% on female group with 66%,67% specificity and sensitivity respectively. Furthermore, MLP achieved 71%, 70% and 73% in terms of accuracy, specificity and sensitivity respectively. Kaufman *et al.* [16] investigated the prospect of speech analysis as a prescreening or tracking tool for type 2 diabetes mellitus (T2DM) through contrasting the voice recordings between non-diabetic and T2DM individuals. Total 267 participants were diagnosed as non-diabetic or diabetic based on American Diabetes Association (ADA) guidelines. Samples recruited in India using a smartphone application recording a fixed phrase in addition to demographic features such as age and body mass index. Authors implemented two supervised machine learning models which are logistic regression and naïve Bayes. LR achieved testing accuracy of 70% on women voice dataset. Accuracy went up to 82% when all features were implemented. On Men voice dataset, LR scored a testing accuracy of 69%. Moreover, when all features were considered, accuracy went up to 86%.

This study aims to design an accurate machine learning model that trains each class independent of dataset original distribution and size through the employment of a proper preprocessing approach to handle the non-existing values, data rescaling and class imbalance. Furthermore, Improving the performance of proposed model through balancing the dataset and ensemble techniques. The proposed model was applied on diverse datasets to ensure the generalizability of the results.

The remainder of this manuscript is organized as follows: Section 2 reviews methods used in the application of machine learning for diabetes diagnosis, following the description of the datasets employed in this study. Section 3 presents experimental results and discussion. Finally, a conclusion and possible future work in section 4.

2. METHOD

This section implements the stated architecture of this paper. It delves into five main parts. Firstly, dataset description, then data preprocessing, train-test split and cross validation, ML algorithms and finally performance evaluation metrics as shown in Figure 1.

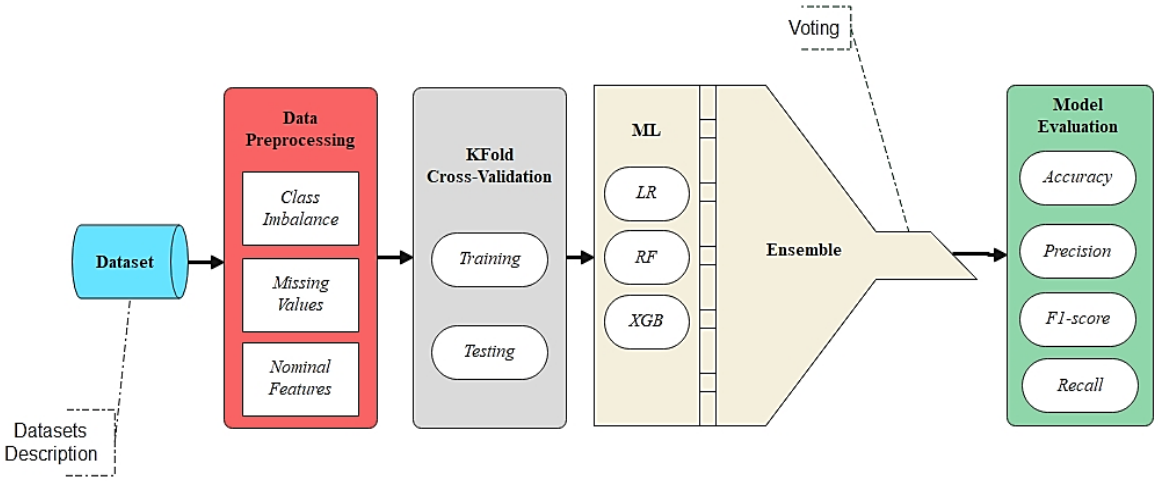


Figure 1. Proposed framework block diagram

2.1. Dataset

Labeled data is a vital input to supervised machine learning and deep learning classification problems [17]. A relevant collection of data aids to better machine learning classification. There are four datasets implemented in this study that differ in number of samples as well as the number and the type of their attributes. They were gathered from public hosts and by agreements with medical centers and doctors. They are publicly available online hosted by UCI Machine Learning. Detailed description of features in entire datasets is illustrated through Tables 1 to 4.

Table 1. Descriptive features of PIDD

Attribute	Description	Null values count	Range
Pregnancies	Number of times a patient has been pregnant	-	0-17
Glucose	Concentration of plasma glucose at two hours in an oral glucose tolerance test (GTIT)	180	0-199
BP	Diastolic blood pressure (mm Hg)	221	0-122
ST	Skin fold thickness in Triceps (mm)	292	0-99
Insulin	Serum Insulin for two hours (μ/ml)	498	0-846
BMI	Body mass index (kg/m)	80	0-67.1
DPF	Diabetes pedigree function	-	0.078 – 2.42
Age	Age in years	-	21 -81
Outcome	Binary target indicating diabetic or not	-	0 - 1

The first dataset namely Pima Indian Diabetes Dataset (PIDD). The PIDD is a widely used medical data records in machine learning. It was gathered by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). It focuses on predicting whether a patient has diabetes based on diagnostic measurements and personal data. The dataset is part of the UCI Machine Learning Repository available online on Kaggle [18]. It consists of 768 instances with 9 attributes including the target variable. All features are numeric and described in Table 2. The second dataset is submitted using a questionnaire for diabetes prediction case study [19]. It contains 520 samples and 17 predictive features. The third dataset [20] consists

of 100,000 samples with 7 features. It is considered the largest dataset in this study among the four implemented ones. The fourth and last dataset named Voice-and-diabetes-VOCADIAB [21] is available on GitHub repository. It is a part of the Colive Voice study, that focuses on using voice analysis to screen for type 2 diabetes (T2DM) in the adult population of the United States. The goal of the study is to analyze acoustic recordings which are in the form of voice embeddings. Participants likely provided standardized voice recordings, such as: Sustained vowels *(/aa/ or /oo/*). Additionally, it involves the associated participant meta data to develop a machine learning-based screening tool for type 2 diabetes.

Table 2. Descriptive features of questionnaire dataset

Attribute	Description	Range (Distribution)
Age	Age of person in years	16-90
Gender	Sex of patient	Male (63%) or Female (37%)
Polyuria	Excess urination	Yes (50% or No (50%)
Polydipsia	Excess thirst	True (45%) or False (55%)
Sudden weight loss	Unintentional and rapid weight loss	True (42%) or False (58%)
Weakness	Reduced energy	True (59%) or False (41%)
Polyphagia	Excessive hunger or increased appetite	True (46%) or False (54%)
Genital thrush	Fungal infection	True (22%) or False (78%)
Visual blurring	Difficulty in seeing clearly	True (45%) or False (55%)
Itching	Persistent skin pruritus	True (49%) or False (51%)
Irritability	Emotional sensitivity or mood swings	True (24%) or False (76%)
Delayed healing	Slow wound healing	True (46%) or False (54%)
Partial paresis	Weakness or paralysis of muscle group	True (43%) or False (57%)
Muscle stiffness	Reduced flexibility in muscles	True (38%) or False (62%)
Alopecia	Hair loss and hormone imbalance	True (34%) or False (66%)
Obesity	Excess body fat	True (17%) or False (83%)
Class	Binary target indicating diabetic or not	Positive (62%) or Negative (38%)

Table 3. Descriptive features of third dataset

Attribute	Description	Range
Gender	Sex of patient	Male (1) – Female (0)
Age	Age in years	0-80
Ht	Hypertension	Yes (1) – No (0)
Hd	Heart Disease	Yes (1) – No (0)
Smoking	Participant history of smoking	Current, never, former, no_info
BMI	Body mass index (kg/m)	10.01 – 95.69
HbA1c	Glycated Hemoglobin: a blood test that measures the average blood sugar (glucose) levels over the past 2-3 months	3.5-9
Bgl	Blood glucose level	80 - 300
diabetes	Binary target indicating diabetic or not	Diabetic (1) –non-Diabetic (0)

Table 4. Descriptive features of VOCADIAB

Attribute	Description	Range
Byols_embeddings	Numerical representations of key acoustic and speech characteristics extracted from participants' voice recordings	-
Gender	Sex of participant	Male (1) – Female (0)
Age	Age in years	18 - 81
BMI	Body mass index (kg/m)	15.82 to 66.93
Ethnicity	Race of participant	Latino – white – black – mixed – other – Asian - unknown
ADA_score	Diabetes-related score based on the American Diabetes Association's classification.	Integer from 0 to 7
Diabetes	Binary target indicating diabetic or not	Diabetic (1) –non-Diabetic (0)

2.2. Data preprocessing

Preprocessing is an important building block in the process of development of the proposed model. Where the efficiency of the prediction model is altered by the inconsistent data. There are some serious observations in these datasets such as non-existing values or zero values, nominal features and target class unequal distributions. The preprocessing of data is implemented in three different stages which are class imbalance, handling missing values, and encoding.

2.2.1. Class imbalance

Class imbalance occurs when there are comparatively more samples in one class of a dataset than in other class. Machine learning models might have difficulties in prediction stage because of this imbalance. As they tend to favor the majority class resulting in a biased model prediction and misleading performance, particularly for the minority class [16]. Class imbalance issues can be addressed in several ways such as under sampling the dominant class or oversampling the minority class. Additionally, data augmentation, which is commonly employed in image datasets.

All datasets implemented in this study suffer from class imbalance as shown in Figure 2 except for VOCADIAB where the target class distribution is equal. PIDD includes 500 diabetic patients while the number of non-diabetic individuals is 268. The third dataset contains 91,500 non-diabetic individuals and only 8,500 diabetic patients. The questionnaire dataset contains 320 diabetics and 200 non-diabetics.

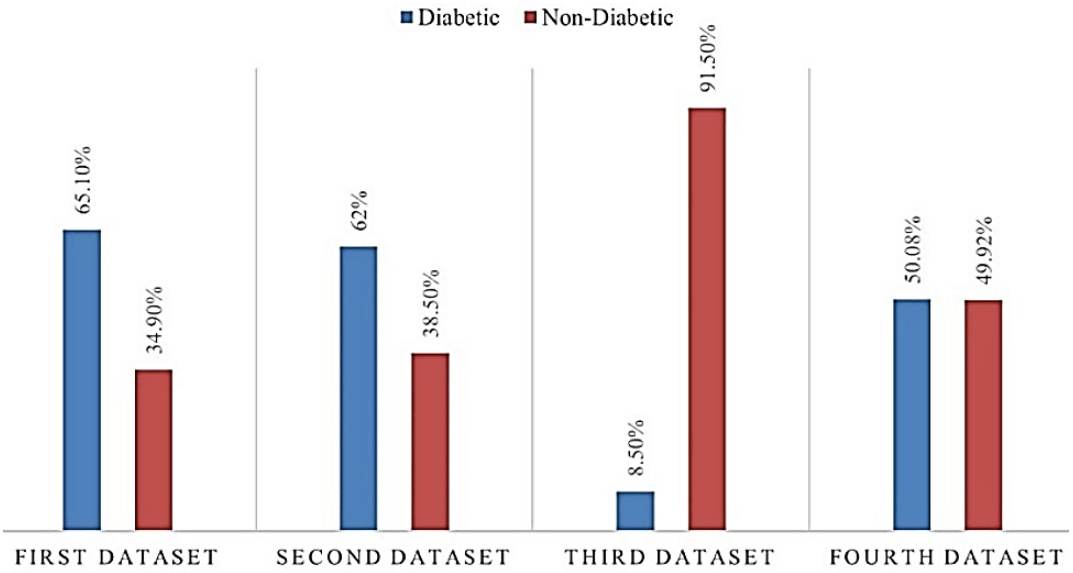


Figure 2. Dataset target class distribution

2.2.2. Missing values

There are some serious observations in these datasets such as null or zero values. Some features, such as Glucose, Blood Pressure, Skin Thickness, Insulin, and BMI, have zero values, which are unlikely in a real-world medical setting. For example, a zero BMI or glucose is biologically impossible. For a variety of reasons, patients often neglect multiple required tests. Therefore, non-existing values will appear in data, requiring the employment of suitable imputation techniques.

There are multiple approaches to handle non-existing attributes like exchanging them with a constant, mean, median and most frequent. Dealing with incomplete medical records can be performed through different methods [22]. Swapping out missing features with a constant "zero" has no effect on the prediction biasing of model. On the other hand, this assumption is biologically impossible. Neglecting incomplete record by simply removing them can affect small-scale datasets. Other mathematical approaches such as replacing non-existing values with a constant, mean, median or most frequent.

2.2.3 Nominal features

The machine learning algorithm needs to transform nominal values into numerical values so that it can comprehend the data it receives to enable further processing. Categorical variables were encoded using one hot encoder. It transforms each unique value in the nominal attribute into a binary vector. Every unique value is represented by a vector with a single "1" indicating the presence of that category while the remaining categories are represented by "0". Encoding is crucial because machine learning models works with numerical data, not categorical labels.

2.3. K-Fold cross validation

Cross validation and train-test split are techniques used in machine learning to evaluate model performance. Since they estimate how well a model will generalize to unseen data. In the train-test split

method, the dataset is divided into two parts. The train set which is used to train the model, and the test Set that is used to evaluate the performance on unknown data. It is considered the simplest approach where data is divided into a 70%-80% for training phase and a 20%-30% for testing phase. It is simple, quick, easy to implement and computationally efficient. It works well for large datasets. On the other hand, there is a high variance where performance depends on how the data is split. The results might vary with different random seeds. It is less reliable for small datasets where the single split might not capture the variability in the data.

The dataset is divided into a number "K" of subsets that are approximately equal size in cross validation. The model is trained and tested K times, with each fold used as the test set exactly once and the remaining folds as the training set as shown in Figure 3. All data points are used for both training and so it is considered computationally expensive and can be time-consuming, especially for large datasets or complex models.

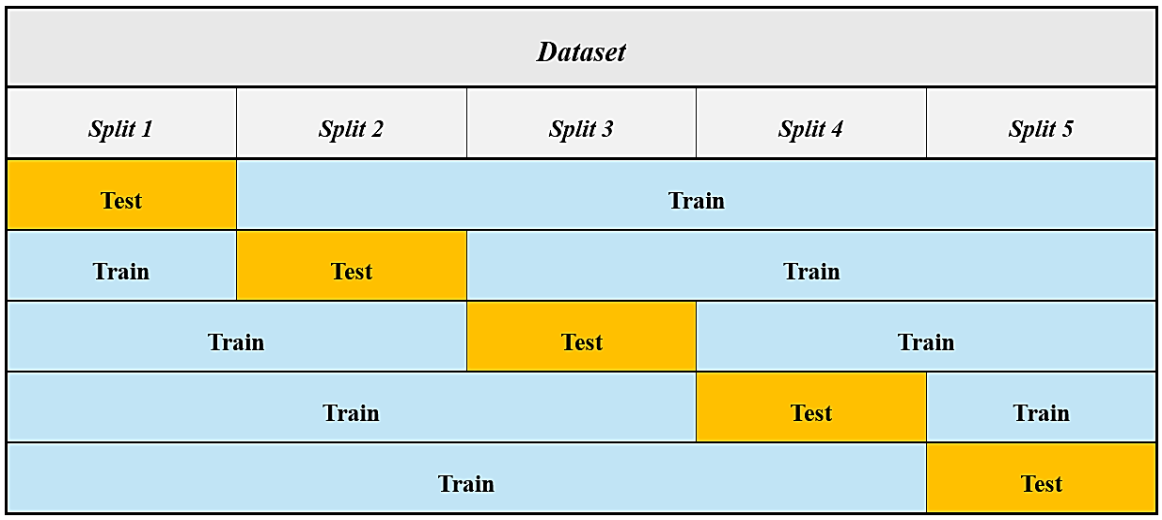


Figure 3. Five-fold cross validation [23]

2.4. Machine learning

Machine learning (ML) make use of mathematical and statistical algorithms in order to identify patterns in data so that it can perform an accurate and precise predictions [24]. ML enhance their performance over time through being exposed to more data. Supervised learning trains on labeled data used in classification as in our case. This study involves an ensemble of machine learning classifiers, such as random forest (RF), extreme gradient boosting (XGB) and logistic regression for the purpose of predicting diabetes mellitus.

2.4.1. Logistic regression

logistic regression (LR) is one of the popular supervised learning algorithms in healthcare systems. It is known for its simplicity and ease of implementation, making it one of the most suitable algorithms for binary classification problems. The LR uses a collection of independent features to predict the likelihood of the class output [25]. The threshold used to identify which data belongs to a particular class is known as the decision boundary [26]. The logistic sigmoid function is used to get this categorization probability. The coefficients of LR provide clear insights into the relationship between each feature and the outcome class.

2.4.2. Random forest

Random forest (RF) creates a number of decision trees and gives the output class of each tree in the training phase [27]. RF can handle a large number of features even if they include missing data, making it suitable for real-world datasets. Moreover, it provides insights to feature importance that determine which variables contribute the most to the prediction. This model offers a straightforward modification that utilizes a correlated tree in the bagging process, this. A certain amount of attributes are ignored across all columns during bootstrapping [28]. This technique aids in the process of reducing variance. On the other hand, it raises the probability of biasing.

2.4.3. Extreme gradient boosting

An extreme gradient boosting (XGB) is a tree-based sequential DT algorithm applied to relatively small or medium size tabular datasets [29]. It is considered to be among the most effective techniques for classification and prediction. It is known for its speed and performance due to optimized gradient boosting algorithms. By combining comparatively weaker and simpler models. Scalability is considered the most important feature in XGB [30], where it implement learning through distributed computing and memory usage is well structured. The use of Lasso and Ridge regularization aids in preventing overfitting. XGB can work with different types of data making it versatile for many medical applications.

2.4.4. Ensemble modeling

Since different models have different strengths and weaknesses. Ensemble methods utilize the collective decision of multiple base models which are more robust and accurate than any individual model. Errors due to biasing, variance, or even noise in the data can be minimized through combining multiple models [31]. Predictions from multiple models are averaged for regression or combined due to majority voting for classification as in our case. Ensemble methods can better identify and utilize important features. Although individual models may be computationally efficient, ensembles can still be efficient through parallel processing or optimized algorithms making it compared to training a single complex model.

2.5. Evaluation measures

Key performance metrics in classification tasks include accuracy, precision, recall, and F1-score. To calculate these metrics, relying on four key components: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). These components are typically represented in a confusion matrix, as illustrated in Figure 4.

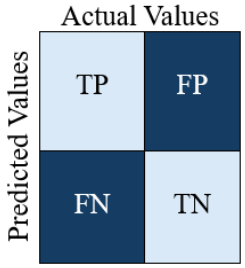


Figure 4. Binary class confusion matrix

Accuracy measures the proportion of correctly classified instances out of the total cases.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

Precision indicates the ratio of correctly identified positive instances to the total predicted positives.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Recall reflects the proportion of actual positive cases that were correctly predicted.

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

F1-Score is the weighted average of precision and recall.

$$F1 - score = \frac{2*Precsion*Recall}{Precision+Recall} \quad (4)$$

2.6. Experimental setup

The model was conducted on the Kaggle platform on an intel i7-10th generation processor. The code was written in Python programming language. The script includes the following key elements:

a. Environmental setup: Importing tools and libraries which were implemented in this study as shown in Figure 5 and illustrated in Table 5.
b. Kaggle configuration: the utilized memory for model training was 1.2 GiB while the disk space was 2.3 GiB. Furthermore, the runtime of entire code was 1,275 seconds without accelerator.
c. Hyperparameter tuning for implemented machine learning models as shown in Figure 6. Values are discussed in Table 6.

```
# Input Required Libraries and Tools
import pandas as pd
import numpy as np
import time as t
import plotly.express as px
import tensorflow as tf
from sklearn.ensemble import RandomForestClassifier, VotingClassifier
from sklearn.linear_model import LogisticRegression
from xgboost import XGBClassifier
from sklearn.preprocessing import OneHotEncoder , OrdinalEncoder , LabelEncoder
from sklearn.utils import shuffle
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import cross_validate
from sklearn.model_selection import train_test_split
from imblearn.over_sampling import SMOTE
from sklearn.metrics import ConfusionMatrixDisplay
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score
```

Figure 5. Screenshot of the input libraries and tools

Table 5. Imported libraries

Library	Implementation purpose
Pandas	Data manipulation and analysis of data frames, handling structed data, cleaning data and transformation.
NumPy	Numerical computing for multidimensional arrays and mathematical operators.
Scikit-learn	Machine learning algorithm for classification and model evaluation.
Plotly	Graphing libraries for quality visualization.
Time	Measuring and managing time-related tasks in programs.
TensorFlow	Open-source library for building and training both machine and deep learning models.

```
# Parameter Tuning
model_pipeline = []
model_pipeline.append(LogisticRegression(solver='liblinear'))
model_pipeline.append(RandomForestClassifier(n_estimators=100, max_depth=None))
model_pipeline.append(xgb.XGBClassifier(n_estimators=100, max_depth=None, learning_rate=0.01))

# Ensemble Model - Voting Classifier
log_clf = LogisticRegression()
rf_clf = RandomForestClassifier()
xgb_clf=xgb.XGBClassifier()
voting_clf = VotingClassifier(estimators=[('lr', log_clf), ('rf', rf_clf), ('xgb',xgb_clf)],
                              voting='hard', n_jobs=-1)
```

Figure 6. Screenshot of the Kaggle ML models and parameter tuning

Table 6. Experiment parameters

ML model	Parameter	Value	Description
LR	solver	"liblinear"	Optimization algorithm.
RF	n_estimators	100	Number of trees in forest.
	max_depth	None	Nodes spread till every leaf is pure to ensure each leaf node represents a distinct class without any ambiguity.
XGB	n_estimators	100	Number of boosting iterations.
	max_depth	None	Max depth of tree.
	learning_rate	0.1	Shrinkage parameter that controls the contribution of each tree to the final model decision.
Ensemble	estimators	LR, RF, XGB	List of tuples where each estimator is a classifier.
	voting	"Hard"	Majority voting class.
	n_jobs	-1	Running a number of jobs for fitting and prediction in parallel where -1 means all processors are being used.

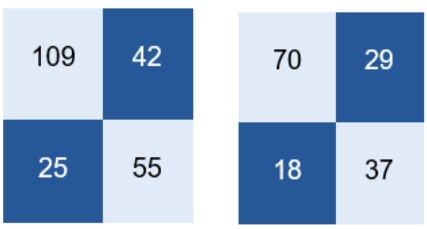
3. RESULTS AND DISCUSSION

A considerable preprocessing is taken into account in three main steps after importing libraries and datasets. First step, dealing with class imbalance issues through applying synthetic minority oversampling technique (SMOT). Second step is replacing non-existing value with mean value instead of simply removing the sample row in order to preserve dataset size. Lastly, converting non-numeric features into numeric one by applying one-hot encoder. Datasets are divided into training and testing partitions with 70%-30% or 80%-20% then 5-folds cross validation is applied. An ensemble model for three base classifiers which are logistic regression, random forest and extreme gradient boosting. Majority voting was chosen in the proposed prediction model. Results are compared using accuracy, precision, recall and F1-score as shown in Table 7. Additionally, confusion matrices are illustrated in Figures 7 to 10, (a) 70% train and 30% test and (b) 80% train and 30% test. Comparison between the proposed model and previous related work is summarized in Table 8.

The ensemble technique combines the predictions of multiple individual models to create a more robust and accurate prediction method. This improvement arises because ensemble methods leverage the strengths of integrated models while minimizing their weak points. Combining models with complementary capabilities allows capturing both linear and non-linear relationships and patterns, where LR captures simple patterns while RF handles non-linear interactions. Additionally, XGB focuses on misclassified data. Standalone RF has some biases depending on the depth of trees. Linear LR models have high bias when problems are non-linear such as in our case. However, overfitting might be a serious issue.

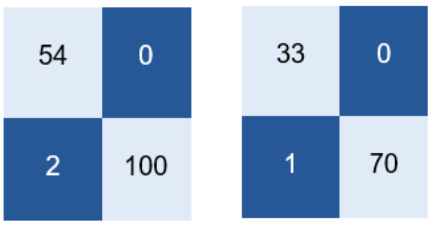
Table 7. Ensemble model results on all datasets

Train-Test Split	Dataset	Accuracy	Precision	Recall	F1-score
80 -20 %	PIDD	81%	80%	83%	81%
	Questionnaire	95%	97%	94%	95%
	Third dataset	96.88 %	89.85%	79.66%	71.55%
	VOCADIB	90.98%	89.47%	90.27	91.07%
70 – 30 %	PIDD	82%	81%	84%	82%
	Questionnaire	96%	97%	95%	96%
	Third dataset	96.83%	89.67%	79.17%	70.87%
	VOCADIB	92.35%	95.18%	91.86%	88.76%



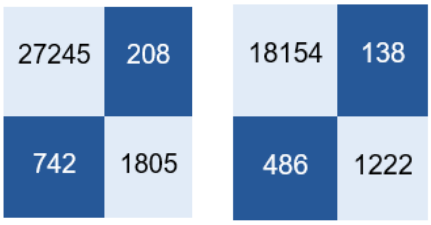
(a) (b)

Figure 7. Confusion matrix for the ensemble model on first dataset (a) 70% train and 30% test and (b) 80% train and 30% test



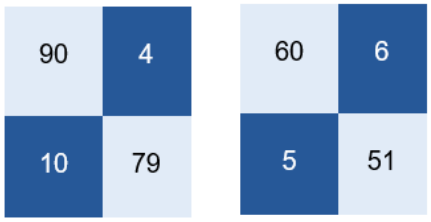
(a) (b)

Figure 8. Confusion matrix for the ensemble model on second dataset (a) 70% train and 30% test and (b) 80% train and 30% test



(a) (b)

Figure 9. Confusion matrix for the ensemble model on third dataset (a) 70% train and 30% test and (b) 80% train and 30% test



(a) (b)

Figure 10. Confusion matrix for the ensemble model on fourth dataset (a) 70% train and 30% test and (b) 80% train and 30% test

Table 8. Proposed model results versus related work in literature review

Authors	Dataset	Technique	Accuracy
Febrian *et al.* [8]	Pima Indian Diabetes dataset	KNN	77.92%
		NB	78.52%
Kangra and Singh [9]	(PIDD)	NB	72.6%
		KNN	66.1%
		SVM	74.3%
		DT	71.8%
		RF	64.9%
		LR	74%
Chang *et al.* [10]		RF	79.57%
		NB and feature selection (3-Factor)	79.13%
		NB and feature selection (5-Factor)	77.83%
Mushtaq *et al.* [11]		Standalone RF	80.7%
		Ensemble (balanced dataset)	81.7%
Rawat *et al.* [12]		AdaBoost	79.69%
Barik *et al.* [13]		RF	71.9%
		XGB	74.1%
Palimkar *et al.* [14]	Case study dataset	LR	93.59%
		SVM	94.23%
		NB	91.02%
		AdaBoost	94.87%
Fagherazzi *et al.* [15]	VOCADIAB	Female group - LR	67%
		Female group - MLP	63%
		Female group - SVM	57%
		Male group - LR	69%
		Male group - SVM	70%
		Male group - MLP	71%
Kaufman *et al.* [16]	Voice records dataset	LR (women – voice features)	70%
		LR (women – all features)	82%
		NB (men – voice)	69%
		NB (men – all features)	86%
Proposed model	PIDD	Ensemble of LR, RF and XGB	82%
	Case study dataset		96%
	Third dataset		96.83%
	VOCADIAB		92.35%

4. CONCLUSION

The prediction of diabetes mellitus is considered a challenging medical research topic. This research involved the development of a machine learning-based pipeline for the process of predicting diabetes mellitus depending on four different datasets. These datasets have serious observations such as class imbalance, missing values in addition to categorical features. Training and testing were performed by applying 5-fold cross validation. Consequently, our goal was met by applying LR, RF and XGB in an ensemble model. The proposed model yielded results which are superior to those of other studies in literature review reaching 82% 81%, 84%,82% in terms of accuracy, precision, recall and F1-score respectively on the PIDD. The results were 92.35%, 95.18%, 91.86% and 88.76 for accuracy, precision, recall and F1-score respectively when applying performance metrics vocal dataset. The highest results are 96.88%, 89.85%, 79.66%, and 71.55% on the third dataset. Results were 96%, 97%, 95% and 96% for accuracy, precision, recall and F1-score respectively on questionnaire dataset. In future work, it is suggested to apply different non-existing value imputation techniques close to real-life situations in addition to various class imbalance techniques. Furthermore, more machine learning and deep learning techniques will be applied on hybrid datasets.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Moataz Mohamed El Sherbiny	✓	✓	✓	✓		✓	✓		✓					
Asmaa Hamdy Rabie			✓		✓		✓	✓		✓				
Mohamed Gamal Abdel Fattah		✓		✓		✓				✓		✓		
Ali Elsherbiny Taki Eldin			✓			✓	✓			✓	✓	✓		
Hossam El-Din Mostafa	✓			✓	✓					✓	✓	✓	✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

REFERENCES

[1] S. Pinyopodjanard, P. Suppakitjanusant, P. Lomprew, N. Kasemkosin, L. Chailurkit, and B. Ongphiphadhanakul, "Instrumental acoustic voice characteristics in adults with type 2 diabetes," *Journal of Voice*, vol. 35, no. 1, pp. 116–121, Jan. 2021, doi: 10.1016/j.jvoice.2019.07.003.

[2] P. Prabhu and S. Selvabharathi, "Deep belief neural network model for prediction of diabetes mellitus," in *2019 3rd International Conference on Imaging, Signal Processing and Communication, ICISPC 2019*, Jul. 2019, pp. 138–142, doi: 10.1109/ICISPC.2019.8935838.

[3] C. Bommer *et al.*, "Global economic burden of diabetes in adults: Projections from 2015 to 2030," *Diabetes Care*, vol. 41, no. 5, pp. 963–970, Feb. 2018, doi: 10.2337/dc17-1962.

[4] R. L. Siegel, A. N. Giaquinto, and A. Jemal, "Cancer statistics, 2024," *CA: A Cancer Journal for Clinicians*, vol. 74, no. 1, pp. 12–49, Jan. 2024, doi: 10.3322/caac.21820.

[5] L. Guo and X. Xiao, "Guideline for the management of diabetes mellitus in the elderly in China (2024 edition)," *Aging Medicine*, vol. 7, no. 1, pp. 5–51, Feb. 2024, doi: 10.1002/agm2.12294.

[6] A. D. A. P. P. Committee, "2. Classification and diagnosis of diabetes: Standards of medical care in diabetes—2022," *Diabetes Care*, vol. 45, pp. 17–38, Dec. 2022, doi: 10.2337/dc22-S002.

[7] O. J. Mbanugo, "AI-enhanced telemedicine: A common-sense approach to chronic disease management and a tool to bridging the gap in healthcare disparities," *International Journal of Research Publication and Reviews*, vol. 6, no. 2, pp. 3223–3241, Feb. 2025, doi: 10.55248/gengpi.6.0225.0952.

[8] M. E. Febrian, F. X. Ferdinan, G. P. Sendani, K. M. Suryanigrum, and R. Yunanda, "Diabetes prediction using supervised machine learning," *Procedia Computer Science*, vol. 216, pp. 21–30, 2022, doi: 10.1016/j.procs.2022.12.107.

[9] K. Kangra and J. Singh, "Comparative analysis of predictive machine learning algorithms for diabetes mellitus," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 3, pp. 1728–1737, Jun. 2023, doi: 10.11591/eei.v12i3.4412.

[10] V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms," *Neural Computing and Applications*, vol. 35, no. 22, pp. 16157–16173, mar, 2023, doi: 10.1007/s00521-022-07049-z.

[11] Z. Mushtaq, M. F. Ramzan, S. Ali, S. Baseer, A. Samad, and M. Husnain, "Voting classification-based diabetes mellitus prediction using hypertuned machine-learning techniques," *Mobile Information Systems*, vol. 2022, pp. 1–16, Mar. 2022, doi: 10.1155/2022/6521532.

[12] V. Rawat, S. Joshi, S. Gupta, D. P. Singh, and N. Singh, "Machine learning algorithms for early diagnosis of diabetes mellitus: A comparative study," *Materials Today: Proceedings*, vol. 56, pp. 502–506, 2022, doi: 10.1016/j.matpr.2022.02.172.

[13] S. Barik, S. Mohanty, S. Mohanty, and D. Singh, "Analysis of prediction accuracy of diabetes using classifier and hybrid machine learning techniques," in *Smart Innovation, Systems and Technologies*, vol. 153, Springer Singapore, 2021, pp. 399–409, doi: 10.1007/978-981-15-6202-0_41.
[14] P. Palimkar, R. N. Shaw, and A. Ghosh, "Machine learning technique to prognosis diabetes disease: Random Forest classifier approach," in *Lecture Notes in Networks and Systems*, vol. 218, Springer Singapore, 2022, pp. 219–244, doi: 10.1007/978-981-16-2164-2_19.
[15] A. Elbéji *et al.*, "A voice-based algorithm can predict type 2 diabetes status in USA adults: Findings from the Colive Voice study," *PLOS Digital Health*, vol. 3, no. 12, p. e0000679, Dec. 2024, doi: 10.1371/journal.pdig.0000679.
[16] J. M. Kaufman, A. Thommandram, and Y. Fossat, "Acoustic analysis and prediction of type 2 diabetes mellitus using smartphone-recorded voice segments," *Mayo Clinic Proceedings: Digital Health*, vol. 1, no. 4, pp. 534–544, Dec. 2023, doi: 10.1016/j.mcpdig.2023.08.005.
[17] R. Krishnamoorthi *et al.*, "A novel diabetes healthcare disease prediction framework using machine learning techniques," *Journal of Healthcare Engineering*, vol. 2022, pp. 1–10, Jan. 2022, doi: 10.1155/2022/1684017.
[18] C. Bartley, "Replication data for: Pima Indians diabetes," Harvard Dataverse, 2016.
[19] C.-Y. Guo and Y.-J. Lin, "diabetes_data_upload." IEEE Dataport, 2023.
[20] M. Mustafa, "Diabetes prediction dataset," *Kaggle*, 2023. https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset (accessed Jul. 06, 2025).
[21] A. Elbeji, "Voice-and-diabetes-VOCADIAB," *GitHub*. https://github.com/LIHVOICE/Voice-and-diabetes-VOCADIAB (accessed Jul. 06, 2025).
[22] M. Saar-Tsechansky and F. Provost, "Handling missing values when applying classification models," *Journal of Machine Learning Research*, vol. 8, pp. 1625–1657, 2007.
[23] R. Rousyati, A. N. Rais, N. Hasan, R. F. Amir, and W. Warjiyono, "Comparison of Adaboost and bagging with naive Bayes on bank direct marketing dataset," (in Bahasa), *Bianglala Informatika*, vol. 9, no. 1, pp. 12–16, Mar. 2021, doi: 10.31294/bi.v9i1.9890.
[24] A. F. A. H. Alnuaimi and T. H. K. Albaldawi, "An overview of machine learning classification techniques," *BIO Web of Conferences*, vol. 97, p. 133, 2024, doi: 10.1051/bioconf/20249700133.
[25] S. Nusinovici *et al.*, "Logistic regression was as good as machine learning for predicting major chronic diseases," *Journal of Clinical Epidemiology*, vol. 122, pp. 56–69, Jun. 2020, doi: 10.1016/j.jclinepi.2020.03.002.
[26] A. M. Rakhimovich, K. K. Kadirbergenovich, Z. M. Ishkobilovich, and K. J. Kadirbergenovich, "Logistic regression with multi-connected weights," *Journal of Computer Science*, vol. 20, no. 9, pp. 1051–1058, Sep. 2024, doi: 10.3844/JCSSP.2024.1051.1058.
[27] S. Wang, "Diabetes prediction using random forest in healthcare," *Highlights in Science, Engineering and Technology*, vol. 92, pp. 210–217, Apr. 2024, doi: 10.54097/5ndh9a05.
[28] Y. Manzali and M. Elfar, "Random forest pruning techniques: A recent review," *Operations Research Forum*, vol. 4, no. 2, May 2023, doi: 10.1007/s43069-023-00223-6.
[29] R. Sikander, A. Ghulam, and F. Ali, "XGB-DrugPred: computational prediction of druggable proteins using eXtreme gradient boosting and optimized features set," *Scientific Reports*, vol. 12, no. 1, Apr. 2022, doi: 10.1038/s41598-022-09484-3.
[30] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, vol. 13-17-August-2016, pp. 785–794, doi: 10.1145/2939672.2939785.
[31] A. Batool and Y. C. Byun, "Toward improving breast cancer classification using an adaptive voting ensemble learning algorithm," *IEEE Access*, vol. 12, pp. 12869–12882, 2024, doi: 10.1109/ACCESS.2024.3356602.

BIOGRAPHIES OF AUTHORS

Moataz Mohamed El Sherbiny received the B.Sc. degree in communication and information engineering from Mansoura University, in 2017, and the M.Sc. degree in electronics and communication engineering from Mansoura University, in 2023. He is currently an assistant lecturer at the Electronics and Communication Department, Faculty of Engineering, Mansoura University. His research interests are in the area of artificial intelligence and biomedical engineering. He can be contacted at email: moatazelsherbiny@mans.edu.eg.

Asmaa Hamdy Rabie received a B.Sc. in computers and systems engineering with general grade excellent with class honor in 2013. She received the M.Sc. degree in the area of load forecasting using data mining techniques in 2016 at the Computer and Control Systems Department, Mansoura University, Egypt. She received the Ph.D. degree in the area of load forecasting using data mining techniques in 2020 at the Computer and Control Systems Department, Mansoura University, Egypt. Her interests include programming languages, classification, big data, data mining, healthcare system, and internet of things. She is currently a lecturer in the Faculty of Engineering, Mansoura University, Egypt. She can be contacted at email: asmaahamdy@mans.edu.eg.

Mohamed Gamal Abdel Fattah is an assistant professor of electronics and communication engineering at Mansoura University, Egypt. A cybersecurity researcher since 2012, his work spans optical image encryption, IoT security, deep learning-based cryptanalysis, and digital watermarking. He earned his Ph.D. in 2021 with a focus on advancing optical encryption techniques. He has published extensively in top-tier journals on topics including lightweight IoT protocols, data hiding, and content-based image retrieval (CBIR). His interdisciplinary research extends to biomedical applications, where he designs enhanced antenna systems for secure communication and diagnostic imaging. He can be contacted at email: eng.mo.gamal@mans.edu.eg.

Ali Elsherbiny Taki Eldin received his B.Sc. degree in electronics and electrical engineering from Air Defense College, Alexandria University, followed by both his M.Sc. (Eng.) and Ph.D. degrees from Alexandria University, Alexandria, Egypt. He is an IEEE senior member and currently serves as an associate professor and the head of the Cyber Security Department at the Faculty of Artificial Intelligence, Delta University for Science and Technology, Mansoura, Egypt. Dr. Takieldeen has authored over 50 research publications in prestigious journals. His research interests include cryptography and number theory, digital signatures, cybersecurity, and artificial intelligence-based applications. He can be contacted at email: a_takieldeen@deltauniv.edu.eg.

Hossam El-Din Mostafa is a professor at the Department of Electronics and Communications Engineering. He is the founder and former executive manager of the Biomedical Engineering Program (BME) at the Faculty of Engineering, Mansoura University. He is an IEEE senior member. His research interests include biomedical imaging, image processing applications, and bioinformatics. He can be contacted at email: hossam-moustafa@mans.edu.eg.