Novel Machine Learning Model for Fast and Accurate Diabetes Prediction

Prosanjeet Sarkar (sarkarprosanjeet08@gmail.com)
Santosh Pawar

Research Article

Keywords: Novel machine learning model, Diabetes, Accuracy, Robustness, Prediction

Posted Date: November 2nd, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3500371/v1

Additional Declarations: No competing interests reported.

Novel Machine Learning Model for Fast and Accurate Diabetes Prediction

ABSTRACT

Diabetes mellitus, one of the fastest spreading chronic diseases in the world and is related to anomalous, undeniably escalated levels of glucose in the blood. People with diabetes are at an increased risk of developing complications like strokes, diabetic retinopathy, renal failure, and neuropathy, among others, which raise the rates of morbidity and mortality. Making things even worse when medical science confirms that there is no cure for this disease yet! Even though there is no cure for diabetes. In developing nations, diabetes has emerged as the second leading cause of mortality. The primary step towards controlling and minimizing the risk factor of diabetes is the early detection of diabetes. The advancement of public healthcare facilities has prompted the gathering of sensitive and important healthcare data. The adoption of machine learning algorithms in conjunction with the analysis of acquired data facilitates the early detection and prognostication of diseases. The aim of this study is to propose a reliable and efficient novel machine learning model for predicting diabetes mellitus with better accuracy. For diabetes prediction, eight different classifiers were employed, i.e., Naïve Bayes (NB), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbour (KNN), XGBoost (XGB), Novel machine learning model. For experimental evaluation, the Pima India diabetes dataset (PIDD) and Diabetes dataset are used. During the analysis, it was observed that the novel machine learning model outperformed other classifiers with 99.6% accuracy on the PIDD and 99.8 % accuracy on the Diabetes dataset. The findings reveal the resilience of a newly developed machine learning model in identifying diabetes mellitus at an early stage.

Keywords: Novel machine learning model, Diabetes, Accuracy, Robustness, Prediction

Introduction

Diabetes is one of the chronic diseases that is currently causing problems throughout the world and is getting more severe and morbid in both industrialized and underdeveloped nations [1]. The World Health Organization survey in 1990 and 2014 found that diabetes cases are doubled from 4.7% to 8.5%, and 22 million deaths reported due to complications caused by diabetes [2]. The severe condition is that 43% of these 3.7 million die before the age of 70 years. The government has fundamental concerns about protecting and preventing from this chronic disorder, and due to the fact that their GDP is significantly spent by them on the development of advanced hospitals, medicine, and awareness programs have prolonged the life expectancy of people [3]. However, during the past several decades, there has been a shift in people's lifestyles, resulting in a significant decrease in the prioritization of health maintenance and preservation [4]. This change in lifestyle is one of the causes of diabetes. Diabetes is a contributor to other fatal diseases, which makes it a source of life-threatening illnesses i.e., strokes, kidney failure, vision loss, hypertension, nerve damage, and even premature death [5] [6].

Diabetes mellitus (DM) is a medical disorder characterized by insufficient or absent insulin production by the pancreas, or impaired responsiveness of the body to insulin [7]. It is generally divided into Type-I, Type-II, and Gestation diabetes. Type-I diabetes is also known as insulin-dependent DM [8]. It occurs when the body threatens the pancreas with antibodies. In this condition,

the pancreas is weak and does not produce insulin. It is found that body gens cause this sort of diabetes. Insulin doses, regular health check-ups, and a good diet plan can control it. Type-II diabetes is also known as adult–onset or non-insulin-dependent DM [9]. It is observed that 90% of patients have type-II diabetes. In this case, the body releases some insulin, but it is not enough, or the body does not utilize it as it would. It is typically observed that those who are overweight, more than 20% above their ideal weight due to their height, have a significantly escalated chance of type-II diabetes. Regular medical check-ups, exercise, and a healthy diet can control it. Gestational diabetes is known to be a condition that is initiated during pregnancy in women. During the gestational period, the placenta exhibits resistance to the uptake of insulin by the body's cells, resulting in elevated blood glucose levels. This disorder is transmitted to the baby through the placenta. It can cause severe complications to the baby during postnatal and fatal life [10]. Regular medical check-ups and a healthy diet can control it.

Several ML-based research article has been published for the prediction of diabetes mellitus. The developed ML model for diabetes prognosis is still in the improvised phase because of the paucity of efficiency and robustness. The early detection of diabetes manages to control the spreading of health issues and medical expenses. This study presents a novel machine learning model that is utilized for the early detection of diabetes mellitus. It is a binary classification model and is used to improve the result. Novel machine learning model consist of two levels. First, three soft voting classifier are employed, which gives binary classification output and utilizes an ensemble of three distinct sets of machine learning (ML) classifiers: NB + LR+ DT, RF + SVM + KNN, and XGB + LR + RF. Second, the output of soft voting classifiers is the input of hard voting classifier where all the inputs have equal weight, follow the majority vote concept, and give binary classification output. The performance of the proposed novel machine learning model has been evaluated by taking precision, accuracy, specificity, F1-score, MCC, and AUC. This novel machine learning model reduces the burden on medical practitioners and the likelihood of human error.

This research paper covers the following essential contribution

- Introduced novel machine learning model used for the binary classification.
- To evaluate the robustness of the novel machine learning model using accuracy, precision, specificity, F1 score, MCC and AUC.
- The experimental results demonstrate the resilience and efficacy of the proposed model. The PIDD and Diabetes dataset are used.

The paper's most notable aspects are structured in the following manner: Section 2: Discusses the literature review. Section 3: Present proposed methodology used in building the novel machine learning model. Section 4: present experimental result and analysis. Section 5: Discuss the conclusion and future work.

Related Work

Health care is one of the crucial parts of society that should developed through science and technology. Machine learning models have great strength to deal with large datasets for predictive analysis and knowledge extraction so that they can come up with ways to create applications. A large amount of research work has been done for the reorganization of diabetes patients using machine learning and deep learning models in recent years; some methods are discussed below.

In 2022, Md. Mehed et al. utilized the Pima India diabetes dataset for training and testing the supervised ML model: Multilayer perceptron (MLP), RF, DT, SVM, and KNN. In this pre-processing,

XGB based finds the feature importance dataset and the K-mean technique used for clustering the dataset. The clustering dataset model achieved an accuracy of 99. 57% for RF, 98.70% for MLP, 96.10% for KNN, 99.13% for DT and 97.40% for SVM. The analysis of all the results was found that RF achieved the most accuracy of 99.57% [11].

In 2022, Usama Ahmed et al. proposed a fused ML model for the prediction of diabetes. The fused technique combines two ML models, i.e., support vector machine and artificial neural network. The experiment used the diabetes dataset, which is collected from the UCI repository, which has 520 instances and 17 attributes. The accuracy of the proposed fused ML model was 94.87% [12].

In 2021, Pratya Nuankaew et al. proposed a novel ML model, which is an average weight objective distance-based model for the prediction of type-2 diabetes. The weighted objective distance in this model was adjusted by including knowledge gain, allowing for the identification of major and inconsequential individual factors with varying priority, as indicated by different weights. The experiment was performed on two datasets, i.e., the PIDD and the Mendeley Data for diabetes dataset. The proposed method achieved an accuracy of 93.22% on PIDD and 98.95% on the Mendeley diabetes dataset [13].

In 2023, Muhammet et al. proposed an innovative algorithm that utilizes deep learning technique for the early classification of diabetes. This algorithm uses three-step diabetes classification strategies. First, the complete image dataset feeds to the ResNet 18 and ResNet 50 CNN models. In the second, the output result of the first is fused and classified with the support vector machine, and finally, selected fusion features are classified by the support vector machine. The experimental work used the image dataset of the PIDD. The maximum achieved accuracy of the mode was 92.19% [14].

In 2022, Rashi Rastogi et al. presented a data mining technique for the prediction of diabetes. This paper used LR, SVM, Naïve Byes, and RF as machine learning classifiers. In their literature, the Pima India diabetes dataset, which comprises 768 female patients and 8 feature columns, has been utilized. In this work, data pre-processing consists of data cleaning, data integration, and data reduction. The logistic regression model achieved accuracy of 82.48% [15].

In 2021, Umair Muneer Butt et al. proposed two models for the classification of diabetes and the prediction of blood glucose levels. In this paper, for the classification of diabetes patients, multilayer perceptron, logistic regression, and random forest model were used. For the prediction of blood glucose level, long short-term memory, moving average and linear regression model. In the experimental evaluation, the PIDD was used, and it was observed that multilayer perceptron outperformed the other classifier with an accuracy of 86.08%, and long short-term memory achieved blood glucose level prediction accuracy of 87.26% [16].

In 2021, Saloni Kumari et al. presented a soft voting ensemble machine learning model for the classification and prediction of diabetes mellitus. This ensemble of three machine learning classifiers, i.e., Naïve Bayes, logistic regression, and random forest, gives binary output. The PIDD was used for the experimentation work. Accuracy, precision, recall, and F1-score are used to evaluate the performance of the model. The model achieved a classification accuracy of 79.04% [17].

In 2022, Aishwariya Dutta et al. presented a weighted ensemble machine learning model, which is the ensemble of DT, RF, SGB, and LightGBM. This experiment used a newly labeled dataset, which was collected from the Bangladesh Demographic and Health Survey (BDHS). In 2011, the survey found 4751 diabetic and 2814 non-diabetic cases, and in 2017, the survey found a total of 3492 diabetic and 4073 non-diabetic cases. In this survey no record of pre-diabetic cases. This dataset

only contains diabetic (1) and non-diabetic (0), which is a binary classification dataset. To get the higher accuracy and correct result of the model, need to do pre-processing of the dataset using missing values and feature selection method. The weighted ensemble model (DT+RF+XGB+LGB) underwent a statistical analysis of variance test, resulting in an accuracy of 73.5% and an area under the curve (AUC) of 83.20% [18].

In 2022, Rassol Jader et al. proposed an artificial neural network (ANN) model for fast and accurate diabetes reorganization. This study makes two hidden layer ann and three hidden layers ann model and compares the performance. The dataset employed for assessing the efficacy of the suggested mode came from a variety of different labs in the Iraqi Kurdistan Region and a total of 1012 rows and 7 features in this dataset. The three hidden layer ann model outperformed of the two hidden layer ann model with an accuracy of 91.43 % in 100 epochs [19].

3. Methodology

This research work is used to determine the diabetes status of the patient, and the overall workflow of the proposed methods is shown in Figure 1. The pre-processing and novel machine learning model is the two parts of the methods. The proposed methods are described in the following section.

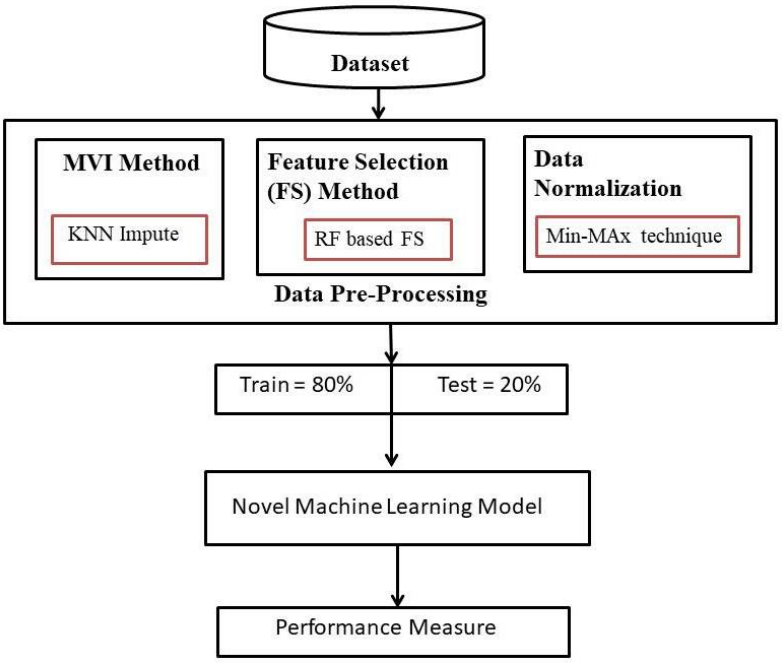


Figure 1. Block diagram of the workflow of the proposed model

3.1 Dataset

This research work used two datasets, the Pima India diabetes dataset (PIDD) and the Mendeley data for diabetes dataset. The PIDD was collected from the UCI repository. It was initially collected by the National Institute of Diabetes and Digestive and Kidney Diseases from a cohort comprising only of female patients aged 21 years or older, who identified as having Pima Indian heritage. The dataset

contains 768 records, of which 268 patients have diabetes, and other 500 are non-diabetic patients. The dataset has 8 feature columns and one binary output column; it specifies the person as diabetic (1) or non-diabetic (o). The eight features, i.e., pregnancy month, glucose level, blood pressure, fold thickness of triceps skin, quantity of insulin, body mass index, Pedigree function, and output, show the diabetic or non-diabetic person. Table 1 shows the abbreviation PIDD dataset.

Table 1. Abbreviation of PIDD dataset

Abbreviation	Feature	Description	Feature Values
Pr	Pregnancies	Number of times pregnant	Numeric Values
Gl	Glucose	Plasma glucose concentration	Numeric Values
Bp	Blood Pressure	Diastolic blood pressure	Numeric Values in (mm Hg)
St	Skin Thickness	Triceps skinfold thickness	Numeric Values in (mm Hg)
In	Insulin	2-Hours serum insulin	Numeric Values in mm
Bm	BMI	Body mass Index	Numeric Values in (mu U/ml)
PDF	Diabetes Pedigree Function	Diabetes Pedigree Function	Numeric Values in (weight in Kg/ (height in m)^2)
Ag	Age	Patient Age	Numeric Values

The diabetes dataset was gathered from the Mendeley repository and collected initially from the Iraqi society, which was acquired from the laboratory of Medical City Hospital and the Specializes Centre for Endocrinology and Diabetes-AI-Kidney Teaching Hospital. The dataset contains 1001 records, of which 898 patients have diabetes, and the other 103 are non-diabetic patients. The dataset has 10 features, i.e., Gender, Age, Urea, Creatinine ratio, hemoglobin A1c, Cholesterol, Triglycerides, High-density lipoprotein, Low-density lipoprotein, Very low-density lipoprotein and Body mass index, and output class; it specifies the person has diabetic (1) or non-diabetic (o). Table 2 shows the abbreviation diabetes dataset.

Table 2. Abbreviation of PIDD dataset

Abbreviation	Feature	Description	Feature Values
Ag	Age	Patient Age	Numeric Values
Ur	Urea	chief nitrogenous part	Numeric Values (mg/dL)
Cr	Creatinine Ratio	Parameter to access Kidney	Numeric Values in (µmol/L)
Hb	HBA1C	Sugar level	Numeric Values in (mmol/mol)
Ch	Cholesterol	Fatty substance produced by the liver	Numeric Values in (mg/dL)
Tg	Triglycerides	Type of fat in the blood	Numeric Values in (mmol/L)
Hd	HDL	High density lipoprotein, good cholesterol used for body	Numeric Values in (mmol/L)
Ld	LDL	Low density lipoprotein, bad cholesterol	Numeric Values in (mmol/L)
Vl	VLDL	Very low density lipoprotein cholesterol is produced in the liver	Numeric Values in (mmol/L)
Bm	BMI	Body mass index	Numeric Values in (mu U/ml)

Table 3 and Table 4 present examples of PIDD and Diabetes datasets, respectively, focusing on characteristics related to the identification of diabetes.

Table 3. Some examples of PIDD dataset

Pr [0-17]	**Gl** [0-199]	**Bp** [0-122]	**St** [0-99]	**In** [0-846]	**Bm** [0-67.1]	**PDF** [0.078-2.42]	**Ag** [21-81]	**Outcome** [0-1]
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1

Table 4. Some examples of PIDD dataset

Ag [20-79]	**Ur** [0.5-38.9]	**Cr** [6-800]	**Hb** [0.9-16]	**Cn** [0-10.3]	**Tg** [0.3-13.8]	**Hd** [0.2-9.9]	**Ld** [0.3-9.9]	**Vl** [0.1-35]	**Bm** [19-47.75]	**Class** [0-1]
50	4.7	46	4.9	4.2	0.9	2.4	1.4	0.5	24	0
26	4.5	62	4.9	3.7	1.4	1.1	2.1	0.6	23	0
34	3.9	81	6	6.2	3.9	0.8	1.9	1.8	23	1
31	3.4	55	5.7	4.9	1.6	1	3.2	0.7	24	1
43	2.1	55	5.7	4.7	5.3	0.9	1.7	2.4	25	1

3.2 Data Pre-processing

Data pre-processing is a crucial stage in which raw data is converted into a format that is both useful and efficient for input into a machine learning model. The method of pre-processing is succinctly described in the subsequent subsection.

3.2.1 Missing value Imputation (MVI)

The effectiveness of a trainable automated decision model is contingent upon the quality of the dataset. However, practically all dataset commonly includes nonstandard properties of missing values, typically represented as null, NaN, blank, or similar placeholder. Therefore, it is necessary to eliminate the missing values in the dataset or imputer with robust and sensible data generated by the machine learning model. In this paper, the KNN-based imputation technique is used to replace the missing value in the dataset. The basic idea for searching for missing values is the minimum distance between present class features and the existing class's feature value. This paper employed the Euclidean distance equation for calculating the minimum distance between features of two classes, as shown in equation (1).

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(yi - xi)\text{^}2} \quad (1)$$

3.2.2 Feature Selection (FS) Method

FS is the fundamental approach employed to determine the most suitable feature for a given machine learning model. It is commonly implemented for model simplification to reduce the computation time, reduce memory usage, reduce dimensions, enhance predictive accuracy by choosing imperative features, and avoid irrelevant data that overfit the model. This paper employed an

RF-based features election approach. RF commonly assigned ranks to the feature based on the impurity level of the node, minimizing all tree's impurities. The nodes that exhibit the highest impurity reduction are found at the beginning of the trees, whereas a gradual decrease in impurity is observed towards the end of the tree. Consequently, a subset of the pertinent characteristics can be acquired by the process of pruning the trees beneath a specific node.

3.2.3 Data Normalization

A dataset containing numerous features that have different ranges of values to set the features in the common range or comparable range is known as normalization, and it improves the performance of the model. The most prevalent technique for normalization is feature scaling, and in this study, we employ the min-max technique for feature normalization of the dataset. In this approach, the new sample value ($x^{|}$) is determined according to the present sample (x), maximum (x_{max}), and minimum (x_{min}) values of the features. As a result, all feature values are distributed between the same range of 0-1, and the formula for the min-max normalization method is shown in equation (2).

$$x^{|} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{2}$$

3.3 Model architecture

In the novel machine learning model, we have employed cascade of soft voting ensemble and hard voting ensemble machine learning classifier algorithms. The developments of novel machine learning algorithm need classical machine learning algorithm i.e. NB, LR, KNN, DT, RF, SVM, and XGB. These algorithms are briefly explained below.

3.3.1 Naïve Bayes (NB)

It is used to solve binary classification problems, and it is a simple and effective algorithm. It gives better results in higher dimensional training datasets. It uses a probabilistic classifier as its guideline principle , as shown in equation (3). It classifies the instances based on the likelihood that a feature will exist.

$$P(A/B) = \frac{P\left(\frac{B}{A}\right)P(A)}{P(B)} \tag{3}$$

3.3.2 Logistic Regression (LR)

It is suitable to use when the dependent variable is binary, as we have classified the person as diabetic (1) or non-diabetic (0). Moreover, this technique is employed for conducting predictive analysis and establishing the correlation between a dependent variable and one or more independent variables, as depicted in equation (4). The prediction is made by utilising the sigmoid function to estimate the likelihood of an event taking place for each individual data instance. The sigmoid function gives an s-shaped cure for the classification of diabetic or non-diabetic patients; the sigmoid function is shown in equation (5).

$$y = \theta_0 + \theta_0 x_0 + \theta_1 x_1 + \text{----------} \mp \theta_n x_n \tag{4}$$

$$\sigma(y) = \frac{1}{1+e^{-y}} \tag{5}$$

3.3.3 K- Nearest Neighbor

This algorithm is used for both regression and classification problems. The underlying concept in addressing binary classification involves effectively differentiating K instances within a training dataset that are similar and proximate to another instance, and classifying this new instance based on the majority of its nearest neighbours, assigning it to the most prevalent class among the K closest neighbours. In this experiment, we used Euclidean distance, as shown in equation (1), which is utilized to process the distance between two instances.

3.3.4 Decision Tree

It is a supervised machine learning algorithm commonly employ for both classification and regression task. In the structure, the root node indicates the category of labels, and the decision node signifies the combination of features that result in the assignment of those specific category labels. The approach utilises a divide and conquer strategy by employing a greedy search algorithm to discover the optimal split locations inside a given tree structure. The aforementioned procedure is executed in a hierarchical and iterative fashion, continuing until all predominant data entries have been categorised according to distinct labels. The classification error rate can be defined as the ratio of instances in the training set that are not assigned to the predominant class, shown in equation (6).

$$\text{Entropy (s)} = \sum_{i=1}^{n} -Pi\ Log(Pi) \quad (6)$$

Where, Pi represents the proportion of the training set from the ith class in the region.

3.4.5 Random Forest

It is, as its name suggests, an ensemble of decision tree models that work together. The idea behind the random RF is the majority wisdom of the crowd each model makes a prediction and the majority ultimately prevails. RF can be employed for a variety of scientific investigations, including the detection of diabetes. Using the bootstrap randomized resampling technique, it extracts many copies of the sample set from the initial training dataset. In order to ensure the proper functioning of an algorithm, it is necessary to select a suitable value that accurately represents the number of trees present in the forest. This can be achieved by employing the bootstrap sampling technique in conjunction with the Gini criterion, which aids in identifying the most effective splitting approach. The Gini index formula for classification, as shown in equation (7) and equation (8) represent the RF based prediction of class label.

$$\text{Gini Index} = 1 - \sum_{i=1}^{n}(Pi)^2 \quad (7)$$

Where, Pi represents the proportion of the training set from the ith class in the region.

$$C_P = \arg max_c \left(\frac{1}{n} \sum_{i=1}^{n} I\left(\frac{m_{fi,\ c}}{m_{fi}}\right)\right) \quad (8)$$

Where,

- C_P represent the prediction of class label
- c represent the class
- m_{fi} represent the score of decision tree in the fi decision tree.
- n represent the number of decision tree in RF

- $I\left(\frac{m_{fi,\ c}}{m_{fi}}\right)$ represent the indicator function for $m_{fi,\ c}$.

3.4.6 Support Vector Machine

It is a popular supervised computational machine learning model used in both regression and classification problems. In order to make it simple to classify the new data sample in the appropriate category in the feature, the SVM technique is used to generate the optimal decision boundary that can divide n-dimensional space. The selection of an ideal hyperplane in a dimensional space is widely recognised as a highly tough issue. The optimal hyperplane is defined as the one that exhibits the greatest margin between two distinct classes. The points which are close to the hyper-plane are called support vectors. Based on the hyper-plan that corresponds to one of the classes along the hyper-plan, the unknown sample point is categorized. The margin between two support vectors (π^+, π^-) of binary classification is shown in equation (9).

$$\text{Margin} = \text{dist}\ (\pi^+, \pi^-) \tag{9}$$

3.4.7 XGBoost

XGBoost, which stands for extreme gradient boosting, is a supervised robust machine learning algorithm used in regression as well as classification purposes. It uses an ensemble approach; the ensemble uses a bagging and boosting approach. XGB used a gradient boosting algorithm; it has three main functions: the loss function, weak learner, and additive module. Data over-fitting is a robust algorithm's main issue, yet gradient boosting is a greedy technique that can over-fit a large dataset. The performance of the algorithm is increased by lowering over-fitting with regularization techniques.

3.4.8 Novel Machine Learning Model

The proposed architecture of the novel machine learning model, as shown in Figure 2, is a highly accurate binary classification model. The proposed model serves as a meta-classifier that combines machine learning models, whether similar or dissimilar, to make predictions using a majority voting approach. The novel machine learning model consists of two layers. Layer 1, the soft voting classifier, has classical machine learning models and uses the predict probability method. In this, each soft voting classifier (SVC), NB + LR+ DT, RF + SVM + KNN, and XGB + LR + RF classifiers have been ensemble. A soft voting ensemble has been utilized, which incorporates the predicted probability method of each selected feature variable. Then, the training data and data points undergoes a shuffling process, and these data points are passed to the classical machine learning model. Each classical machine learning model computes prediction with voting aggregate and soft voting technique, and the majority voting is computed, which yields the output of the each soft voting classifier. In layer 2, the hard voting ensemble classifier (HVC), all the soft voting classifier output in layer 1 has been ensemble. This majority vote technique, where all the input has equal weights and output result is based on the mode of all the predictions made by the input of the hard voting classifier. The algorithm for the proposed novel machine learning model has been illustrated in figure 3.

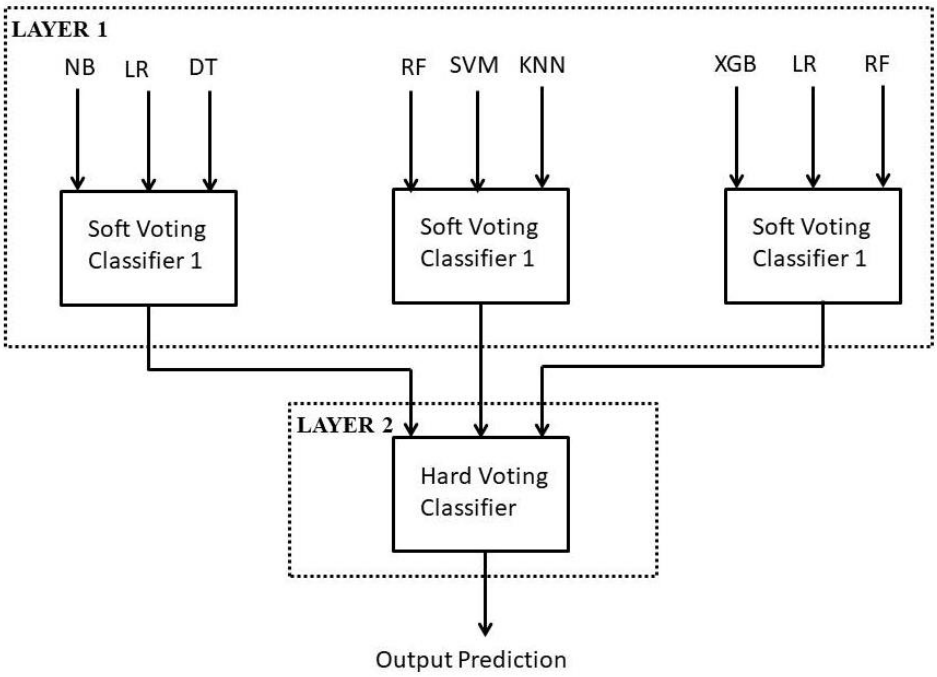


Figure 2: Architecture of Novel machine learning model for detection of diabetes

1. START
2. IMPORT : dataset = PIDD.csv/Diabetes.csv
3. REPLACE: dataset missing value -> from KNN imputation method used in dataset
4. FEATURE EXTRACT : features -> RF method used in dataset
5. NORMALIZATION : feature normalization - > min-max technique on features
6. SPLIT DATASET : train_data, test data -> split dataset
7. CLASSICAL MODEL : M1 = Naïve_Bayes(train_data, test_data)
   M2 = Logistic_Regression (train_data, test_data)
   M3 = Decision_Tree (train_data, test_data)
   M4 = Random_Forest (train_data, test_data)
   M5 = K_Nearest_Neighbor (train_data, test_data)
   M6 = Support_Vector_Machine(train_data, test_data)
   M7 = XGBoost(train_data, test_data)
8. SOFT VOTING CLASSIFIER : soft1 = concatenate (M1, M2, M3)
   soft1.fit (train_data)
   soft2 = concatenate (M4, M5, M6)
   soft2.fit (train_data)
   soft3 = concatenate (M2, M4, M7)
   soft3.fit (train_data)
9. HARD VOTING CLASSIFIER : hard = concatenate (soft1, soft2, soft3)
10. PREDICTION : hard_voting_classifier.predict (test_data)
11. END

Figure 3: Algorithm for Novel Machine Learning Model

3. Performance measure

To evaluate the performance of various classifier models, some critical metrics are computed on the basis of confusion matrix structure as shown in Figure 4.

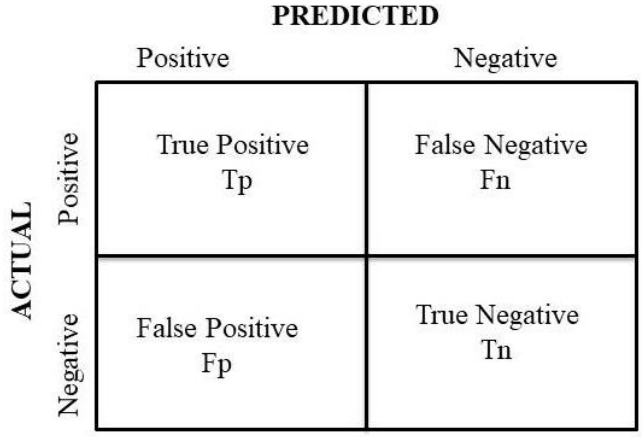


Figure 4. Structure of confusion matrices

The performance of the system is computed with the Tp, Tn, Fp, and Fn values in the matrix. The performance metrics accuracy, sensitivity, precision, specificity, F1-score, and MCC are calculated with the help of a formula, as shown in Table 5.

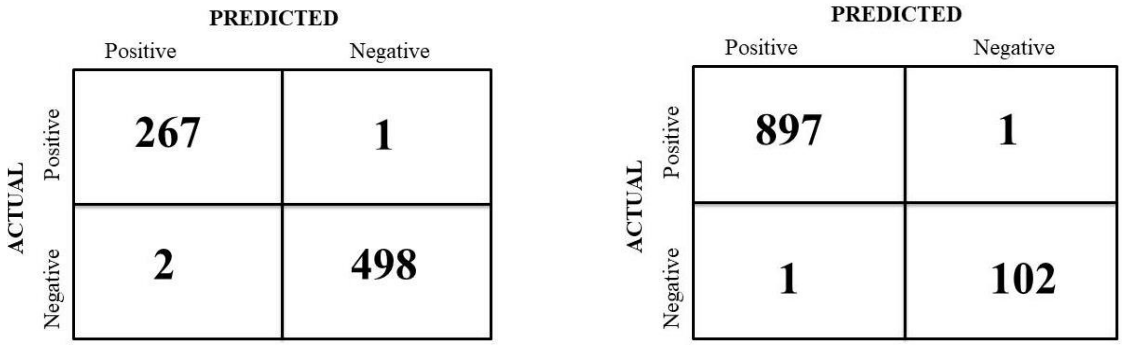
Table 5. Performance metrics

Accuracy	$\frac{Tp + Tn}{Tp + Tn + Fn + Fp}$
Sensitivity	$\frac{Tp}{Tp + Fn}$
Precision	$\frac{Tp}{Tp + Fp}$
Specificity	$\frac{Tn}{Tn + Fp}$
F1-score	$\frac{2 * Tp}{2 * Tp + Fp + Fn}$
MCC	$\frac{(Tp * Tn) - (Fn * Fp)}{\sqrt{(Tp + Fn) * (Tn + Fp) * (Tp + Fp) * (Tn + Fn)}}$

4. Result and Discussion

This section presents a discussion of the outcomes derived from the novel machine learning model. The prediction of diabetes mellitus using machine learning models was conducted on a DELL laptop with an Intel core i3-3.9 GHz Processor and 4 GB RAM. The model was created and trained on Anaconda Studio, Jupyter Notebook, using the Sckit learn library. It is an open-source studio that is used to build, train, and test the machine learning model. The proposed novel machine learning model employed default parameters of the machine learning algorithm. The hyper-parameter tuning needs to be done as per dataset, and this tuning changes for every dataset. Hyper-parameter tuning makes the model overfit. This novel machine learning model only used the default setting of parameters, and this is efficient enough to produce a superior result in our investigation. To analyze the novel machine learning model, we use the PIDD and Diabetes dataset, which have two classes of positive and

negative datasets. Figure 5 shows the confusion matrix for both the PIDD and diabetes dataset of the patient that the novel machine learning model correctly or incorrectly classifies.



(a) (b)

Figure 5: Confusion matrix. (a) Pima India diabetes dataset (b) Diabetes Dataset

Table 6 demonstrates the performance for the prediction of diabetes mellitus using a novel machine learning model. It shows that the prediction accuracy revealed that the proposed novel machine learning model provides high accuracy, with 99.6% for PIDD and 99.8% for the diabetes dataset. The evaluation of the prediction performance with accuracy, sensitivity, precision, specificity, F1-score, MCC, and AUC confirmed that the proposed novel machine learning model has a high potential to predict diabetic or non-diabetic persons.

Table 6. Performance of proposed model on PIDD and Diabetes dataset.

	Accuracy	Sensitivity	specificity	Precision	F1-score	MCC	AUC
PIDD	99.6%	99.62%	99.6%	99.25%	99.44%	99.14%	99.5%
Diabetes Dataset	99.8%	99.88%	99.02%	99.88%	99.88%	98.91%	99.8%

The prediction accuracy obtained from the novel machine learning model was compared against various classical machine learning models such as NB, LR, KNN, DT, RF, and XGB, as shown in Table 7. The concept of a novel machine learning model is work based on the majority voting classifier. The prediction performance obtained from the proposed novel machine learning model outperformed other classical machine learning models on both datasets, achieving an accuracy of 99.6% for PIDD and 99.8% for the Diabetes dataset. The novel machine learning has the potential to predict the diabetic and non-diabetic person. Therefore, it can be observed that the assumptions presented in this study support the notion that the proposed innovative machine learning model is capable of achieving greater accuracy compared to traditional machine learning models.

Table 7. Comparison of various machine learning models.

Model	Accuracy	
	PIDD	Diabetes dataset
Naïve Bayes	82.13	83.59
Logistic Regression	84.86	88.92
K- Nearest Neighbour	84.07	85.67
Decision Tree	80.12	82.79
Random Forest	89.73	93.54

XGBoost	89.07	91.22
Novel Machine Learning Model	99.6	99.8

Table 8 shows the internal accuracy of the novel machine learning model; each soft voting classifier model gives higher accuracy in comparison to the classical machine learning model, and all the output from each soft voting classifier passes to the hard voting classifier model which predicts the output based mode of all the prediction made by the output of the soft voting classifier. The achieved result of the proposed model is excellent in comparison to other classical machine learning models.

Table 8. Internal accuracy of novel machine learning model

Novel machine learning model	Accuracy	
	PIDD	Diabetes Dataset
Soft Voting classifier 1	96.59%	97.82%
Soft Voting classifier 2	97.07%	98.69%
Soft Voting classifier 3	96.98%	98.75%
Hard Voting classifier	99.6%	99.8%

The proposed novel machine learning model is contrasted with previously published models and approaches in Table. 9. It can be observed that the proposed novel machine learning model outperformed all of the other published methods and achieved an accuracy of 99.6% for PIDD and 99.8% for the Diabetes dataset.

Table 9. Results of the comparison table

Authors	Model/Approach	Dataset Used	Accuracy (%)	Miss Rate (%)
Aishwariya Dutta et al	Weighted Ensemble model	Bangladesh Demographic and Health Survey dataset	73.5%	26.5%
Yashi Srivastava et al	Azur AI Service	PIDD	77.8%	22.2%
Saloni Kumari et al	Ensemble soft voting classifier	PIDD	79.08%	20.92%
Umair Muneer Butt et al	Multilayer Perceptron	PIDD	86.08%	13.92%
Rassol Jader et al	Artificial Neural Network	Iraqi Kurdistan Region dataset	91.43%	8.57%
Rashi Rastogi et al	Logistic Regression	PIDD	82.48%	17.52%
Muhammet et al	Convolution Neural Network	PIDD	92.19%	7.81%
Pratya Nuankaew et al	Average Weighted Objective Distance based model	PIDD Diabetes dataset	93.22% 98.95%	6.78% 1.05%
Usama Ahmed et al	Fused Machine Learning	UCI repository diabetes dataset	94.87%	5.13%

	Technique			
Md. Mehed et al	Random Forest	PIDD	99.57%	0.43%
Proposed Method	Novel Machine Learning Model	PIDD Diabetes dataset	99.6% 99.8%	0.4% 0.2%

5. Conclusion and Future works

Diabetes is chronic disease with no permanent cure, only early diagnosis can controlled the spreading of disease. The scientific community has consistently emphasised the accuracy of illness prediction models in the context of diabetes. Various models have been put out for this purpose. Therefore the new technique and strategy are necessary to enhance the accuracy diabetes mellitus prediction. The principal objective of this research study is to create a highly accurate algorithm for the prognostication of diabetes in patients. Therefore, the author has proposed a novel machine learning model that uses cascading of soft voting and hard voting ensemble classifiers. In the soft voting classifier, the ensemble of NB, LR, DT, RF, SVM, KNN, and XGB. In the hard voting classifier, the ensemble of soft voting classifier mode predicts the output based on majority voting. The novel machine learning model has been applied to the PIDD and Diabetes dataset. The novel machine learning model has achieved 99.60% accuracy result on the PIDD and 99.80% correct result on the Diabetes dataset.

In future work, it is planned to increase the number of soft voting classifier models and classical ML classifier to enhance the prediction performance. Furthermore, future studies plan to apply different chronic disease datasets to novel machine learning models for early prediction of disease.

Decleration of Competing Interest

The authors do not have any conflict with other entitles or researches.

References

[1] A. K., Goyal, Y., Bhatt, D., Dev, K., Alsahli, M. A., Rahmani, A. H., & Almatroudi, A. Verma, "A Compendium of Perspectives on Diabetes: A Challenge for Sustainable Health in the Modern Era," *Diabetes, metabolic syndrome and obesity : targets and therapy*, vol. 14, pp. 2775-2787, June 2021.

[2] I. M., Bidikian, N. H., Hneiny, L., & Nasrallah, M. P. El-Kebbi, "Epidemiology of type 2 diabetes in the Middle East and North Africa: Challenges and call for action," *World journal of diabetes*, vol. 12, no. 9, pp. 1401-1425, September 2021.

[3] Karuranga S, Malanda B, et al Williams R, "Global and regional estimates and projections of diabetes-related health expenditure: Results from the International Diabetes Federation Diabetes Atlas," *Diabetes research and clinical practice* , vol. 162, Feburary 2020.

[4] G. Tripathi and R. Kumar, "Early Prediction of Diabetes Mellitus Using Machine Learning," in *8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, Noida, 2020, pp. 1009-1014.

[5] K. Qu, Y. Luo, D. Yin, Y. Ju, & H. Tang Q. Zou, "Predicting diabetes mellitus with machine learning techniques," *Front. Genet.*, vol. 9, no. 515, November 2018.

[6] H. K. Rana, M. S. Azam, M. S. Rana, M. R. Akhtar, M. R. Rahman, M. H. Rahman, M. A. Moni N. K. Podder, "A system biological approach to investigate the genetic profiling and comorbidities of type 2 diabetes," *Gene Reports*, vol. 21, December 2020.

[7] M.S Cheong, W.S Cheang Y. Tan, "Roles of Reactive Oxygen Species in Vascular Complications of Diabetes: Therapeutic Properties of Medicinal Plants and Food.," *Oxygen*, vol. 2, no. 3, pp. 246-268, July 2022.

[8] Rodger W, "nsulin-dependent (type I) diabetes mellitus," *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*, vol. 145, no. 10, pp. 1227-1237, November 1991.

[9] & D.J.P. Barker C.N. Hales, "Type 2 (non-insulin-dependent) diabetes mellitus: the thrifty phenotype hypothesis," *Diabetologia*, pp. 595-601, July 1992.

[10] A. H. Xiang, & K. A. Page T. A. Buchanan, "estational diabetes mellitus: risks and management during and after pregnancy," *Nature reviews. Endocrinology*, vol. 8, no. 11, pp. 639-649, July 2012.

[11] S.Mollick, & F. Yasmin M. M. Hassan, "An unsupervised cluster-based feature grouping model for early diabetes detection," *Healthcare Analytics*, vol. 2, September 2022.

[12] G. Issa, S. Aftab, M. F. Khan, R. Said, T. Ghazal, M. Ahmad, & M. Khan U. Ahmed, "Prediction of Diabetes Empowered With Fused Machine Learning," *IEEE Access*, vol. 10, pp. 8529-8538, January 2022.

[13] S. Chaising, & P. Temdee P. Nuankaew, "Average Weighted Objective Distance-Based Method for Type 2 Diabetes Prediction," *IEEE Access*, vol. 9, pp. 137015-137028, October 2021.

[14] & K. Sabanci M. F. Aslan, "A Novel Proposal for Deep Learning-Based Diabetes Prediction: Converting Clinical Data to Image Data," *Diagnostics* , vol. 13, no. 4, February 2023.

[15] & Mamta Bansal Rashi Rastogi, "Diabetes prediction model using data mining techniques," *Measurement: Sensors*, vol. 25, December 2022.

[16] S. Letchmunan, M. Ali, F. H. Hassan, A. Baqir, & H. H. R. Sherazi U. M. Butt, "Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications," *Journal of healthcare engineering*, September 2021.

[17] D. Kumar, M. Mittal S. Kumari, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," *International Journal of Cognitive Computing in Engineering,* vol. 2, pp. 40-46, January 2021.

[18] M. K. Hasan, M. Ahmad, M. A. Awal, M. A.Islam, M. Masud, & H. Meshref A. Dutta, "Early Prediction of Diabetes Using an Ensemble of Machine Learning Models," *International journal of environmental research and public health* , vol. 19, September 2022.

[19] & S. Aminifar R. Jader, "Fast and Accurate Artificial Neural Network Model for Diabetes Recobnition," *NeuroQuantology*, vol. 20, no. 10, pp. 2187-2196, August 2022.