Intelligent Medicine 4 (2024) 22–32

Contents lists available at ScienceDirect

Intelligent Medicine

journal homepage: www.elsevier.com/locate/imed

Research Article

Clinically adaptable machine learning model to identify early appreciable features of diabetes

Nurjahan Nipa[1], Mahmudul Hasan Riyad[2], Shahriare Satu[3], Walliullah[2], Koushik Chandra Howlader[4,5], Mohammad Ali Moni[6,*]

[1] Department of Information and Communication Technology, Bangabandhu Sheikh Mujibur Rahman Digital University, Bangladesh, Kaliakair, Gazipur, 1750 Bangladesh
[2] Department of Applied Mathematics, Noakhali Science and Technology University, Noakhali, 3814, Bangladesh
[3] Department of Management Information Systems, Noakhali Science and Technology University, Noakhali, 3814, Bangladesh
[4] Department of Computer Science and Telecommunication Engineering, Noakhali Science and Technology University, Noakhali, 3814, Bangladesh
[5] Department of Computer Science, North Dakota State University, Fargo, 58105, ND, United States
[6] School of Health and Rehabilitation Sciences, Faculty of Health and Behavioural Sciences, The University of Queensland, St Lucia, QLD 4072, Australia

ARTICLE INFO

Keywords:
Diabetes
Early features
Machine learning
Classification

ABSTRACT

Objective Diabetes mellitus is a serious disease where the body of affected patients are failed to produce enough insulin that causes an abnormality of blood sugar. This disease happens for a number of reasons including modern lifestyle, lethargic attitude, unhealthy food consumption, family history, age, overweight, etc. The aim of this study was to propose a machine learning based prediction model that detected diabetes at the beginning.

Methods In this work, we collected 520 patients records from the University of California, Irvine (UCI) machine learning repository of Sylhet Diabetes Hospital, Sylhet. Then, a similar questionnaire of that hospital was followed and assembled 558 patients records from all over Bangladesh through this questionnaire. However, we accumulated patient records of these two datasets. In the next step, these datasets were cleaned and applied thirty five state-of-arts classifiers such as logistic regression (LR), K nearest neighbors (KNN), support vector classifier (SVC), Nave Byes (NB), decision tree (DT), random forest (RF), stochastic gradient descent (SGD), Perceptron, AdaBoost, XGBoost, passive aggressive classifier (PAC), ridge classifier (RC), Nu-support vector classifier (Nu-SVC), linear support vector classifier (LSVC), calibrated classifier CV (CCCV), nearest centroid (NC), Gaussian process classifier (GPC), multinomial NB (MNB), complement NB, Bernoulli NB (BNB), categorical NB, Bagging, extra tree(ET), gradiant boosting classifier (GBC), Hist gradiant boosting classifier (HGBC), one vs rest classifier (OVsRC), multi-layer perceptron (MLP), label propagation (LP), label spreading (LS), stacking, ridge classifier CV (RCCV), logistic regression CV (LRCV), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and light gradient boosting machine (LGBM) to explore best stable predictive model. The performance of the classifiers has been measured using five metrics such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic. Finally, these outcomes were interpreted using Shapley additive explanations methods and identified relevant features for happening diabetes.

Results In this work, different classifiers were shown their performance where ET outperformed any other classifiers with 97.11% accuracy for the Sylhet Diabetes Hospital dataset (SDHD) and MLP shows the best accuracy (96.42%) for the collected dataset. Subsequently, HGBC and LGBM provide the highest 94.90% accuracy for the combined datasets individually.

Conclusion LGBM, stacking, HGBC, RF, ET, bagging, and GBC might represent more stable prediction results for each dataset.

1. Introduction

Diabetes mellitus is a chronic disease that happens for producing an insufficient amount of insulin by the pancreas or when the produced insulins are not properly utilized. The normal range of blood glucose is found 70–100 mg/dl for a healthy person. If the level goes above this range, it is known as diabetes [1]. According to International Diabetes Federation (IDF), approximately 463 million people are world-

* Corresponding author: Mohammad Ali Moni, School of Health and Rehabilitation Sciences, Faculty of Health and Behavioural Sciences, The University of Queensland, St Lucia, QLD 4072, Australia (Email: m.moni@uq.edu.au).

https://doi.org/10.1016/j.imed.2023.01.003
Received 23 March 2022; Received in revised form 12 November 2022; Accepted 8 January 2023

wide affected by this disease. It is an apprehension that this number will be raised up to 578 million by 2030 as well as 700 million by 2045 [2]. According to the study, 1 in 5 people aged above 65 years old is affected by this disease. The undiagnosed situation causes various complications like retinopathy, neuropathy, nephropathy, micro, and macro-angiopathies, etc. The patients of this disease are vulnerable to infect various diseases such as pneumonia, tuberculosis, lower limb amputation, and cardiovascular including kidney diseases [2]. In 2019, it caused the death of almost 4.2 million people worldwide. Besides, the rate of infection is constantly growing in low and middle-income countries where almost 79% of adults are carrying out of this disease [3]. Diabetes is divided into three categories namely Type 1, Type 2, and gestational diabetes (GDM). Type 1 occurs due to insufficient or no insulin produced and mostly develops at younger ages, though can develop at any age. Type 2 diabetes is frequently happened rather than other types of diabetes where the produced insulins are not properly utilized due to physical inactivity, sedentary behavior, unhealthy food consumption, etc. GDM occurs in the pregnant woman for high blood glucose, which increases complications for both mother and child. According to the study [3], 1 out of 6 children born alive with this disease. This type of diabetes is generally ended up after the pregnancy period, but later they have higher chances to be affected by type 2 diabetes.

Moreover, the cost due to diabetes health expenditure was 760 billion USD in 2019 worldwide. This amount will be increased to 825 billion USD by 2030 and 845 billion USD by 2045 [4]. In Bangladesh, the mean expenditure was $ 864.7 USD per person in 2017 [5]. In 2011, about 9.7% adults were being affected by this disease and it will be projected to be 13.7 million by 2045 [5]. The cost of diabetes creates a huge burden on natural expenditure in low and middle-income countries. However, oral glucose tolerance test (OGTT) and HbA1c methods are being used to detect this disease globally. But, these methods are costly, time incursive, as well as require the expert technician to perform this test [6]. However, these tests are not properly happened in rural areas. The delayed diagnosis and treatment increase the complexity of this disease to a great extent. Therefore, several factors such as age, glucose, body mass index (BMI), blood pressure, skin thickness, diabetes pedigree function, insulin, pregnancy, etc. are required to identify diabetes more efficiently. However, early diagnosis of diabetes minimizes the morbidity of the patients and helps to avoid any serious complications. Moreover, it is a challenging task because of the nonlinearity as well as the complexity of the data.

Data mining is required to explore various types of diabetes records to diagnose this disease more efficiently. These methods are not only decreasing fatality and complications but also saving time and effort for both the patients and health professionals. In this study, we investigated the potentiality of machine learning techniques to predict diabetes at an early stage. First, we collected patient instances of Sylhet Diabetes Hospital in Sylhet, Bangladesh from the UCI machine learning repository which is called SDHD. Therefore, we gathered diabetes patient records through a similar questionnaire by medical professionals (i.e. from SDHD). Then, we combined these two datasets and created a merged dataset (MDD). In the working steps, we cleaned these datasets and applied various classifiers such as logistic regression (LR), K nearest neighbors (KNN), support vector classifier (SVC), Nave Byes (NB), decision tree (DT), random forest (RF), stochastic gradient descent (SGD), Perceptron, AdaBoost, XGBoost, passive aggressive classifier (PAC), ridge classifier (RC), Nu-support vector classifier (Nu SVC), linear support vector classifier (LSVC), calibrated classifier CV (CCCV), nearest centroid (NC), Gaussian process classifier (GPC), multinomial NB (MNB), complement NB, Bernoulli NB (BNB), categorical NB, Bagging, extra tree(ET), gradiant boosting classifier (GBC), Hist gradiant boosting classifier (HGBC), one vs rest classifier (OVsRC), multi-layer perceptron (MLP), label propagation (LP), label spreading (LS), stacking, ridge classifier CV (RCCV), logistic regression CV (LRCV), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and light gradient boosting machine (LGBM) to these datasets respectively. Finally, we found the best stable predictive models for each dataset. Then, the significant features for each classifier have been interpreted using SHapley Additive exPlanations (SHAP) values.

Section 2 describes the datasets and methodologies used in this work, Section 3 contains the experimental results, and Section 4 includes related works of recent times and compares this work with state-of-the-art works and finally, Section 5 includes conclusion and future plans about this work.

2. Data and methods

The proposed methodology (Figure 1) for detecting diabetes is given in several sections as follows:

2.1. Dataset description

First, we collected an early-stage diabetes risk prediction dataset of Sylhet Diabetes Hospital (i.e., called SDHD) from the UCI machine learning repository [6]. It contains 520 records with 17 attributes which are briefly described in Table 1. This dataset contains information on newly affected patients who have signs and symptoms of diabetes. Among 520 instances, there are found 320 diabetes and 200 normal cases where the ratio of male and female is found as 63% : 37% respectively. The range of age is determined within 20 to 65 of the patients and all attributes are nominal except age. Then, we considered almost similar attributes of SDHD, and this questionnaire is reviewed and approved by the research cell, at Noakhali Science and Technology University. Then, we manually gathered 558 records (i.e., this dataset is titled prediagnosis diabetes (PDD)) that have 19 attributes where the age range of the patients is found within 10 to 90 years old. In this dataset, 191 cases are observed as diabetes and 367 cases are perceived as normal cases. A brief description of different attributes in these datasets is shown in Table 1. Afterward, we combined SDHD and PDD which is renamed MDD. To mix them, we considered similar 14 attributes between the two datasets (i.e., details in Table 1).

2.2. Data preprocessing

The raw instances are often contained noisy and missing values. Therefore, it is needed to preprocess primary data for generating good outcomes. In PDD, we imputed one missing value for age using the mean of age. Then, five missing values were imputed for smoking by the maximum occurrence of value ‘No’. Moreover, we converted weight and all categorical attributes into numeric respectively. Before going to the next step, we checked outliers of these datasets by utilizing inter-quartile range (IQR) method [7–9]. Further, we performed a correlation analysis and *t*-test on these datasets When two attributes are highly correlated, one of them is needed to omit to achieve better results.

2.3. Machine learning classifiers

We applied 35 classifiers namely, LR, KNN, SVC, NB, DT, RF, SGD, Perceptron, AdaBoost, XGBoost, PAC, RC, Nu-SVC, LSVC, CCCV, NC, GPC, MNB, Complement NB, BNB, Categorical NB, Bagging, ET, GBC, HGBC, OVsRC, MLP, LP, LS, Stacking, RCCV, LRCV, LDA, QDA, LGBM into these three datasets. Some of these good-performing ML models are described briefly as follows:

2.3.1. Extra tree (ET)

ET [10] is an ensemble method that consists of different decision trees like RF [11]. But, it differs from RF in two ways. ET minimizes biases and variances where bias is reduced by training whole data samples of each decision tree instead of bootstrapping samples, unlike RF. Besides, the reduction of variances is achieved by picking the cut points while splitting nodes is performed randomly. Random splitting reduces the execution time of the algorithm.

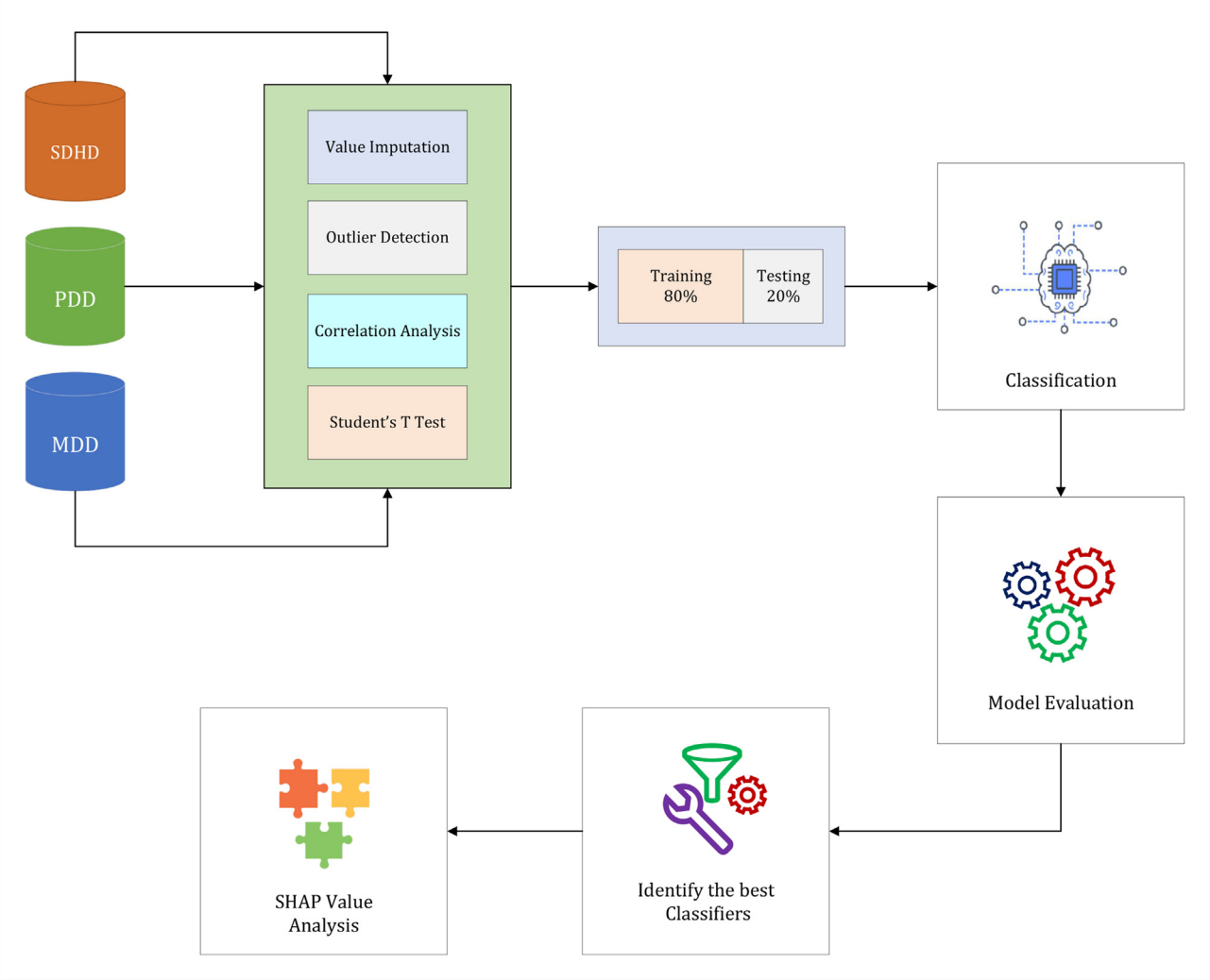

Figure 1. Workflow of the conceptual diagram. SDHD: Sylhet Diabetes Hospital dataset; PDD: prediagnosis diabetes dataset; MDD: combined SDHD and PDD.

Table 1 Different attributes of the datasets

Number	Attributes	Type	Value	SDHD	PDD	MDD
1	Age (years)	Numerical	(10–90)/(20–65)	✓	✓	✓
2	Gender	Nominal	Male or Female	✓	✓	✓
3	Polyuria	Nominal	Yes or No	✓	✓	✓
4	Polydipsia	Nominal	Yes or No	✓	✓	✓
5	Sudden weight loss	Nominal	Yes or No	✓	✓	✓
6	Weakness	Nominal	Yes or No	✓	✓	✓
7	Polyphagia	Nominal	Yes or No	✓	✓	✓
8	Genital thrush	Nominal	Yes or No	✓		
9	Blurred vision	Nominal	Yes or No	✓	✓	✓
10	Itching	Nominal	Yes or No	✓	✓	✓
11	Irritability	Nominal	Yes or No	✓	✓	✓
12	Delayed healing	Nominal	Yes or No	✓	✓	✓
13	Partial paresis	Nominal	Yes or No	✓	✓	✓
14	Muscle stiffness	Nominal	Yes or No	✓	✓	✓
15	Alopecia	Nominal	Yes or No	✓	✓	✓
16	Obesity	Nominal	Yes or No	✓		
17	Weight	Nominal	Yes or No		✓	
18	Smoking	Nominal	Yes or No		✓	
19	High blood pressure	Nominal	Yes or No		✓	
20	Parental diabetes	Nominal	Yes or No		✓	

SDHD: Sylhet Diabetes Hospital dataset; PDD: prediagnosis diabetes dataset; MDD: combined SDHD and PDD.

2.3.2. Bagging

Bagging called bootstrap aggregation is an ensemble method for minimizing bias and variance [12]. It constructs the multiple training sets by taking samples using the bootstrap method [13]. Then, different classification techniques are applied to these training subsets. Finally, the prediction results are obtained from each model where these outcomes are aggregated to generate the final output [14]. This technique is often very effective to provide higher performance than single classifiers.

2.3.3. Support vector machine (SVC)

SVC is one of the most popular and efficient supervised algorithms with excellent generalization capability used for both classification and regression problems [15–17]. It divides the data points by creating a hyperplane where some data points are closer to this line called support vectors. In linear SVC, it makes a differentiation between two classes in a n dimensional space with a maximum $n-1$ dimensional hyperplane. The line is chosen from several hyperplanes with maximum margin. Besides, the separation of data points is not easier. Some data points can fall under the 'grey' area. In such scenarios, SVC ignores the wrong position of data points depending on the user-specified parameter which balances classification error and margin maximization. It utilizes several kernel tricks such as linear, polynomial, sigmoid, and radial basis functions (RBF) which map samples from low to high dimensional space.

2.3.4. Multi-layer perceptron (MLP)

Perceptron is a simple classification model that is used for binary classification. In this method, the output is calculated by the weighted sum of input features and a bias term. A particular perception is activated depending on the value of the output result. A perceptron comprises one input and output layer. In contrast, MLP contains at least 3 layers including input, hidden, and output layers. This classifier is extensively used to perform various tasks such as predictive analysis [18-19], image recognition [20-21], speech recognition, machine translation, etc. It is a fully connected feed-forward neural network (FFNN) where input data transmits from the input, hidden to the output layer. For complex data analysis, The hidden and output layers are used in nonlinear transfer functions into MLP. However, it minimizes prediction error into an acceptable range using the backpropagation algorithm.

2.3.5. Extreme gradient boosting (XgBoost)

XgBoost [22], an updated variant of Gradient Boosting Machine (GBM) is an ensemble classifier that is extensively used in prediction, classification, as well as regression problems [23–26]. It integrates several weak learners to generate a strong learner in terms of scalability, execution speed, and performance. The subsequent weak learners reduce the residual error of previous learners by finding the second-order gradients.

2.3.6. Gradiant boosting classifier (GBC)

GBC [27] is a powerful ensemble method that combines weak learners to generate a strong learner for classification and predictive tasks [28]. It consists of three main parts: loss function, a number of weak learners, and an additive model. GBC improves its accuracy by reducing the losses of the previous base learners in each iteration.

2.3.7. Quadratic discriminant analysis (QDA)

QDA is the extension of LDA which separates the data points of every class by creating a hyperplane whereas QDA differentiates the data points of each point using a quadratic surface. When the data variance is relatively small, LDA delivers good results than QDA. While the data size is big and the variances become larger, QDA provides good results whilst LDA does not provide good outputs for a longer time. In two classifiers, the observations of these classes follow the Gaussian distribution and utilize Bayes theory for classification. Unlike LDA, the covariance of every class is not similar to QDA.

2.3.8. Light gradient boosting machine (LGBM)

LightGBM [29] is an ensemble method based on gradient boosting that is efficient for predictive tasks [26,30]. It is combined various decision trees and splitting is performed leafwise. To handle a large number of data samples and features, it uses the Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) technique.

2.3.9. Hist gradient boosting classifier (HGBC)

While gradient boosting is slower to train and inefficient for 10,000 data samples, HGBC is a robust technique to train a larger amount of samples.

2.3.10. Stacking

Stacking is an ensemble technique that predicts results by utilizing two layers of learning models. In the first layer, different base learners are trained by subsets of the primary dataset. Then, the outputs of these learners are used in the subset of 2nd layer learning model or meta learner. Besides, the dependent variable remains the same as a primary dataset, only the generated output from base models is given as input to the meta learner [31].

2.3.11. Decision tree (DT)

DT is a classification model that is easy to use and interpret, even for novice users. It analyzes records of different characteristics and divides input space hierarchically until it reaches a category. It has three types of nodes such as root, internal, and terminal/leaf node. The root node has zero or more outgoing edges whereas it has no incoming edges. Instead, The leaf nodes contain the incoming edges, but not any outgoing edges. The internal nodes have two or more outgoing edges but exactly contain one incoming edge. Both the root and internal nodes investigate instances based on attributes and splitting rules. This classifier analyzes unknown records by sorting them from root to leaf node.

2.3.12. K nearest neighbors (KNN)

KNN is a widely used classifier that gathers the instances of similar characteristics in their proximity. This algorithm is used to identify unknown records based on the class label of neighboring instances. It considers a number of neighboring instances by choosing the number of k and the unknown records are classified appropriately. In the small number of k, KNN is vulnerable to overfitting because of noisy training data. Instances are regarded as points in the n-dimensional space and affected in the labeling of k value. The distances of instances can be manipulated through various metrics. The actual position of records in n-dimensional space is not considered as the main issue rather than relative distances. In this method, the distance of similar instances is lower whereas the distance of different class instances is determined as higher.

2.3.13. Random forest (RF)

RF [32] is a popular ensemble learning method used for classification, regression, and other tasks. This technique constitutes a number of decision trees to solve a particular problem. In a general decision tree, RF randomly selects individual nodes by nth best splits and constructs trees from a different subset of a node. Then, a test sample is predicted by each tree and aggregated to predict it.

2.4. Model interpretation for feature importance

In this work, various machine learning classifiers were used to analyze diabetes data and determine more accurate results for identifying this disease. However, it is required to explore which properties/features are significant to derive these results. There are many techniques such as SHAP, Local Interpretable Model-agnostic Explanations (LIME), Kernel SHAP, DeepLIFT, etc. to interpret features of any machine learning model. In this work, we have used SHAP model to gain insights and

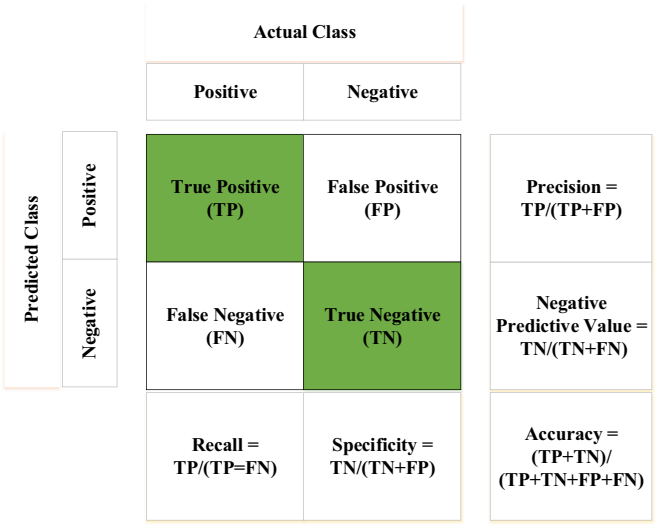

Figure 2. Confusion matrix.

knowledge about individual features of the model. This model was proposed by Lundberg and Lee [33], to interpret and rank different features according to their contribution to generating output. It uses local explanation methods [34] as well as game theory rules [35] to select features and make decision-making. The contribution of each feature a of a model is denoted by ϕ_a where the output is assigned by calculating their marginal contribution. Let M is a set of all input features where Shaply values are obtained through various axioms to allocate the contribution of each feature and predict output $f(M)$, by following Equation (1) where S represents the set of non-zero indexes in $y\prime$ as well as m represents the number of input features.

$$\phi_a = \sum_{S \subseteq M\{a\}} \frac{|S|!(m-|S|-1)!}{m!}[f(S \cup \{a\}) - f(S)] \tag{1}$$

A linear function l of a binary variable is modeled by an additive feature attribution method using the following Equation (2).

$$l(y\prime) = \phi_a + \sum_{a=1}^{P} \phi_a y\prime_a \tag{2}$$

In the aforementioned Equation (2), $\phi_a \epsilon \mathbb{R}$ and $y\prime \epsilon \{0,1\}^P$ is 1 if a feature is present, otherwise, it equals to 0.

2.5. Performance metric

The evaluation metrics such as accuracy, sensitivity, specificity, precision, and AUROC were used to determine the capability of the classifiers for detecting diabetes. This measurement is manipulated using a confusion matrix which is a matrix-like representation of the predicted class against the actual class. Therefore, some estimated values are provided as follows (Figure 2):

- True positive (TP): It estimates the positive instances of the predicted class where the actual class was also positive.
- True negative (TN): It estimates the negative instances of the predicted class where the actual class was also negative.
- False positive (FP): It estimates the positive instances of the predicted class where the actual class was negative.
- False negative (FN): It estimates the negative instances of the predicted class where the actual class was positive.

Then, different evaluation metrics are manipulated which are given as follows:

2.5.1. Accuracy

Accuracy is used to evaluate the performance of any classifier based on correctly predicted versus overall instances which are calculated using Equation (3).

$$Accuracy = \frac{(TP+TN)}{(TP+FP+TN+FN)} \tag{3}$$

When class is unbalanced, the highest accuracy is not enough to declare a classifier as the best model.

2.5.2. Precision

It calculates the ratio between true positive values and all positive predictions in Equution (4). The precision value decreases when the model makes more false positive assumptions.

$$Precision = \frac{TP}{TP+FP} \tag{4}$$

2.5.3. Recall

It computes the ratio between true positive values and all positive values of any predictive model in Equation (5).

$$Recall = \frac{TP}{TP+FN} \tag{5}$$

2.5.4. F1-score

F1-score which is defined in Equation (6) is a harmonic mean of precision and recall where the value of F1-score ranges from 0 to 1. The higher value of this metric is generated for low false negative and false positive values.

$$F-Measure = \frac{2 \times (Precision \times Recall)}{(Precision + Recall)} \tag{6}$$

2.5.5. Area under the receiver operating characteristics curve (AUROC)

AUROC is an evaluation metric that constructs a result by manipulating false positive and true positive rates respectively. This value is nearest to 1 which is considered by a good model.

3. Results

3.1. Model training and evaluation

We investigated SDHD, PDD, and MDD with 35 classifiers such as LR, KNN, SVC, NB, DT, RF, SGD, Perceptron, AdaBoost, XGBoost, PAC, RC, Nu-SVC, LSVC, CCCV, NC, GPC, MNB, CNB, BNB, CategoricalNB, Bagging, ET, GBC, HGBC, OneVSRest, MLP, LP, LS, Stacking, RCCV, LRCV, LDA, QDA, and LGBM using Scikit learn libraries in python. All the experiments were performed on Google Colaboratory. The performance of every classifier was evaluated with five evaluation metrics namely accuracy, precision, recall, F1-score, and AUROC. In this case, the ratio of training andtesting split is chosen as 80:20. The result of the experiments are shown in Table 2.

In SDHD, ET outperformed other classifiers where it gave the best accuracy (97.11%) and an F1-score (98.10%). It also gave good precision (98.52%), recall (97.10%), AUROC (96.67%). Other classifiers such as SC, RF, LGBM, GBC, Bagging, HGBC, LR, SGD, AdaBoost, and CCCV provided excellent performance in terms of all metrics. Similar aforesaid classifiers, XGBoost, DT, LP, LS, SGD, AdaBoost, and GPC also performed well, but these classifiers failed to identify some true diabetes instances as positive. In PDD, MLP performed best in terms of accuracy (96.42%) whereas the values of other metrics like precision, recall, F1-score, and AUROC were computed as 92.50%, 97.37%, 94.87%, and 92.18% respectively. Along with six classifiers, the rest of the classifiers yield an accuracy score of over 90%. In terms of accuracy, the subsequent classifiers produced better results like RF(96.40%), XGBoost (96.40%), SC (95.53%), LGBMC (95.53%), Bagging (94.64%), GBC (94.64%), HGBC (94.64%), LRCV(94.64%), LR (94.60%). Considering AUROC, ET and

Table 2 Experiments results of Sylhet Diabetes Hospital dataset and prediagnosis diabetes dataset (%)

	SDHD					PDD				
Classifiers	Accuracy	Precision	Recall	F1-score	AUROC	Accuracy	Precision	Recall	F1-score	AUROC
LR	92.30	94.02	94.02	94.02	92.75	94.60	87.10	93.10	90.00	92.68
KNN	80.70	100.00	70.15	82.40	87.66	93.70	86.60	89.60	88.10	90.33
SVC	64.40	64.40	**100**	78.30	50.00	92.80	83.80	89.60	86.60	90.33
NB	87.50	90.90	89.50	90.20	90.26	93.70	84.30	93.10	88.50	93.35
DT	91.30	96.70	89.50	93.02	94.71	91.90	83.30	86.20	84.70	88.39
RF	96.15	97.01	97.01	97.01	95.82	96.40	**93.10**	93.10	93.10	93.94
SGD	92.30	91.50	97.01	94.20	89.30	88.40	86.30	65.50	75.50	86.45
Perceptron	64.40	64.40	**100**	78.30	89.09	88.40	83.30	68.90	75.40	87.47
AdaBoost	90.40	92.50	92.50	92.50	92.49	92.80	83.90	89.60	86.60	91.50
XGBoost	91.30	96.70	89.50	93.02	95.82	96.40	**93.10**	93.10	93.10	93.35
PAC	80.70	96.22	73.91	83.60	91.85	73.21	78.57	28.94	42.30	92.18
RC	86.50	95.08	84.05	89.23	92.69	92.85	84.09	97.37	90.24	92.09
Nu-SVC	84.60	94.90	81.16	87.50	92.75	91.96	82.22	97.37	89.15	90.33
LSVC	83.60	83.30	94.20	88.40	90.79	79.46	62.29	**100**	76.76	67.65
CCCV	90.40	94.02	91.30	92.60	91.64	93.75	86.04	97.37	91.36	93.35
NC	58.65	77.08	53.62	63.24	61.88	91.96	82.22	97.36	89.15	90.33
GP	89.42	98.33	85.50	91.47	94.12	91.96	83.72	94.73	88.88	89.74
MNB	84.61	90.77	85.50	88.07	87.46	90.17	78.72	97.36	87.05	91.00
Complement NB	84.61	94.91	81.15	87.50	90.21	89.28	77.08	97.36	86.04	91.00
BNB	81.73	91.67	79.71	85.27	89.10	89.28	79.54	92.10	85.36	90.24
Categorical NB	81.73	91.67	79.71	85.27	90.21	91.96	83.72	94.73	88.88	92.68
Bagging	94.23	98.46	92.75	95.52	**96.93**	94.64	90.00	94.73	92.30	93.36
ET	**97.11**	98.52	97.10	**98.10**	96.67	93.75	89.74	92.10	90.90	**94.53**
GBC	95.20	97.05	95.65	96.35	95.82	94.64	90.00	94.73	92.30	93.35
HGBC	94.23	97.01	94.20	95.60	95.82	94.64	90.00	94.73	92.30	93.35
OVsRC	66.34	66.34	**100**	79.77	50.00	91.96	82.22	97.36	89.15	90.33
MLP	88.46	93.84	88.40	91.04	88.31	**96.42**	92.50	97.37	94.87	92.18
LP	91.34	98.38	88.40	93.13	93.60	91.96	83.72	94.73	88.88	89.74
LS	91.34	98.38	88.40	93.13	92.75	91.96	83.72	94.73	88.88	89.74
Stacking	96.15	97.10	97.10	97.10	95.82	95.53	90.24	97.37	93.67	90.09
RCCV	81.73	86.44	82.26	84.30	89.36	93.75	86.04	97.37	91.36	93.94
LRCV	88.46	86.76	95.16	90.76	92.75	94.64	88.09	97.37	92.50	**94.53**
LDA	82.70	87.93	82.25	85.00	92.69	92.85	84.09	97.36	90.24	92.09
QDA	87.50	84.50	96.77	90.22	93.33	90.17	78.72	97.37	**94.93**	93.35
LGBM	96.15	96.77	96.77	96.77	95.82	95.53	92.30	94.73	93.50	93.35

LRCV produced the highest results (94.53%). Meanwhile, QDA achieved the best F1-score (94.93%). Lastly, we investigated MDD dataset (see Table 3 where this dataset was almost balanced, HGBC, and LGBM showed the highest performance (94.90% accuracy, 95.87% precision, 93.00% recall, 94.41% F1-score, and 94.92% AUROC). LGBM provided almost similar results as HGBC with 94.90% accuracy, 95.87% precision, 93.00% recall, 94.41% F1-score, and 94.54% AUROC. In terms of AUROC, Stacking performed the best (AUC 95.47%). Thus, ET and GPC showed good performance than these classifiers. Some classifiers such as LSVC, SGD, NC, SVC, and Perceptron performed poorly for MDD datasets.

Besides, The outcomes of 35 models were compared in terms of accuracy, F1-score, and AUROC for three datasets that are depicted in Figures 3, 4, and 5 respectively. It is observed that LGBMC, SC, HGBC, RF, ET, Bagging, and GBC provided average and stable performance for all the datasets.

3.2. Feature importance using SHAP values

We adopted the SHAP value to interpret the outcomes of the best model in every dataset. The insights of feature contribution of the MLP output are depicted in Figure 6a where the X-axis denotes Shap values and the y-axis contains features. The color indicates lower and higher values for every observation of the feature. Purple indicates higher feature values whilst blue indicates lower feature values. The purple color on the left and right sides of the plot means negative and positive correlation with diabetes prediction respectively. After analyzing Figure 6a, it is said that age has a greater impact on MLP output, followed by delayed healing and polyphagia. The skewed SHAP values denote the most important features. However, the mean absolute SHAP values are shown in Figure 6b demonstrate the feature importance from MLP in descending order and we conclude that age, delayed healing, polyphagia, polyuria, irritability have a great impact on the results followed by parental diabetes mellitus, high blood pressure, muscle stiffness, smoking, and weakness. On the other hand, partial paresis, polydipsia, blurred vision, itching, sudden weight loss, weight, alopecia, and gender have less impact on the output.

Figures 7a and 7b show SHAP plots for ET in the case of SDHD. It is seen that polyuria is the highest influential feature on the output, followed by polydipsia, gender, itching, and sudden weight loss. The higher value of polyuria, polydipsia, sudden weight loss, partial paresis, irritability, visual blurring, and weakness lead to a higher risk of diabetes. In contrast, the lower value of gender, itching, delayed healing, alopecia, muscle stiffness, age, and obesity cause less risk of diabetes.

Figures 8a and 8b show SHAP plots of features for HGBC in the case of MDD. It is observed that age is the highest influence on the model output, followed by polyuria, delayed healing, and polyphagia. The higher value of age, polyuria, delayed healing, polyphagia, blurred vision, and sudden weight loss leads to the positive result of diabetes. In contrast, the lower value of itching, gender, alopecia, irritability, and weakness causes less risk of diabetes.

4. Discussion

Various machine learning models were proposed to explore and detect diabetes more accurately. From individual studies, many widely used methods like ET, MLP, SVM, NB, KNN, LR, LDA, QDA, GPC, RBF, GNB, and DT were employed to investigate this disease. However, it was noticed that ensemble learning-based methods such as bagging, boosting, decorate as well as stacking perform better than individual models in many works. Le et al. [7] proposed Grey Wolf Optimization (GWO) and an Adaptive Particle Swam Optimization (APSO) based MLP

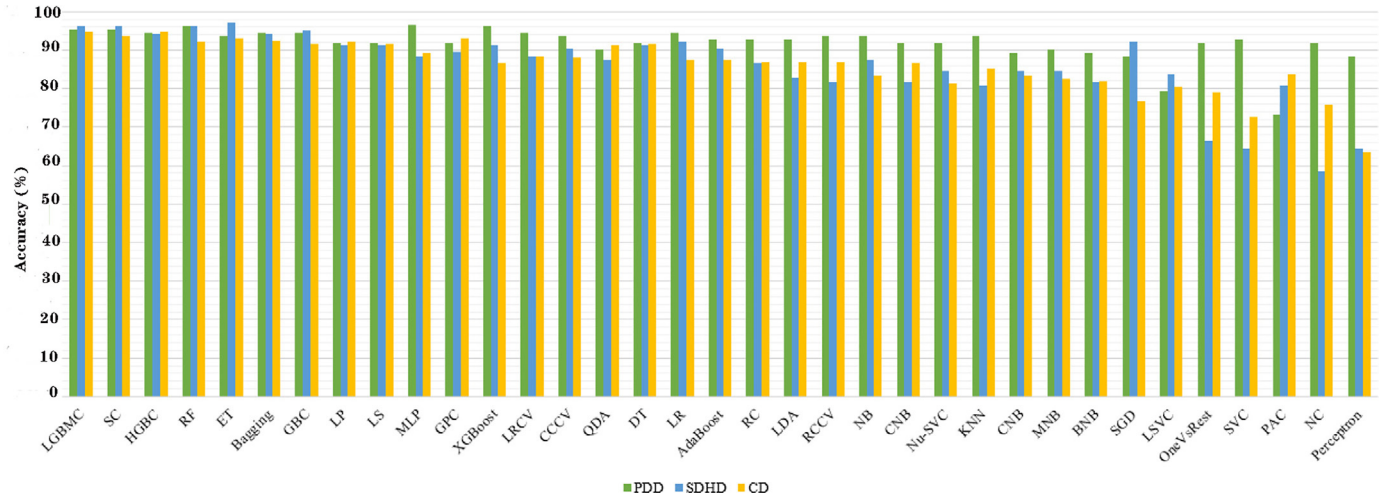

Figure 3. Comparison of the models in terms of accuracy.

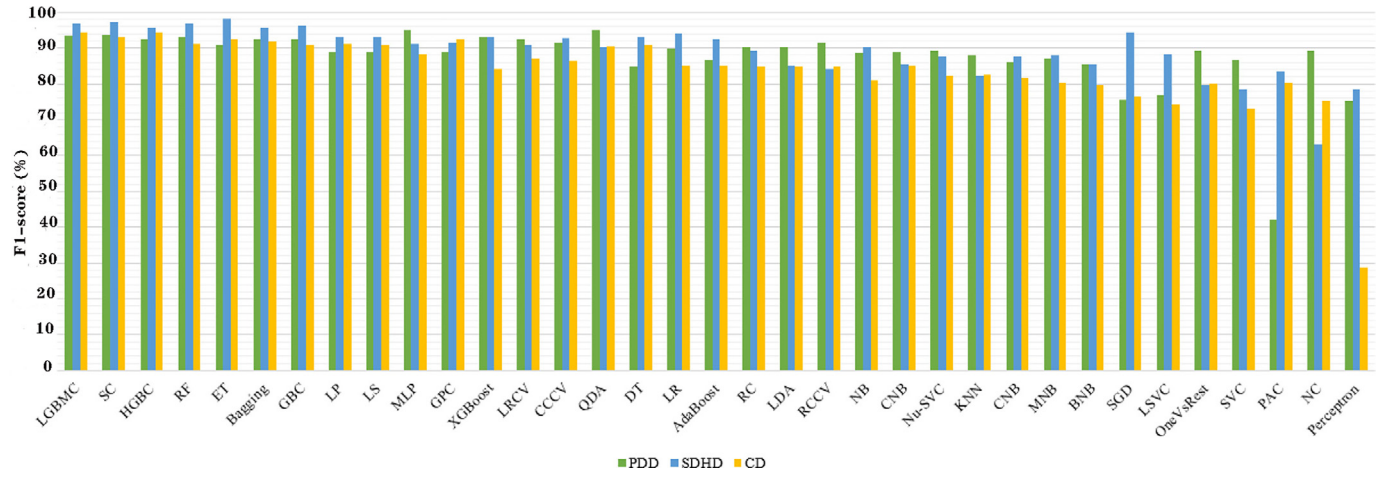

Figure 4. Comparison of the models in terms of F1-score.

method where they detected outliers using IQR and employed the proposed model along with SVM, DT, KNN, NB, RF, LR into SDHD diabetes dataset. In this case, APGWO-based MLP outperformed other classifiers with accuracy 97%, recall 97%, precision 99%, and F1-score 98%. Chaves and Marques [36] compared the performance of NB, NN, AdaBoost, KNN, RF, and SVM for SDHD where NN gave the highest accuracy (98.08%), F1-score (0.984), and AUC (0.983) with 10 fold cross-validation.

Yadav and Pal [12] implemented boosting and bagging techniques using Decision Table, OneR, and JRIP into Pima Indians Diabetes Database (PIDD). Before that, they imputed missing values, performed normalization, and employed the chi-square method to generate a feature subset. In that work, they found reported 98% accuracy, 98% precision, 98% recall, and 97% F1-score using the bagging approach. Islam et al. [14] gathered diabetes patient records from Khulna Diabetes Center, Khulna, Bangladesh, and employed two ensemble techniques such as Diverse Ensemble Creation by Oppositional Relabeling of Artificial Training Examples (DECORATE) [37-38] and bagging method on that dataset. However, DECORATE outperformed the bagging technique with the highest accuracy (98.53%). RF was found one of the efficient bagging algorithms for detecting this disease. Nurjahan et al. [39] employed DT, KNN, NB, SVM, LR, MLP, and XGB into various feature subsets of SDHD and PIDD where RF gave the best outcomes (i.e., 97.5% accuracy, 97.5% f-measure, AUC 99.80%) for GRAE feature subsets. However, LR provided 77.7% accuracy, 77% f-measure for IGAE, and AUC 83% for CSSSE and CAE in PIDD. Moreover, Islam et al. [6] investigated SDHD with NB, LR, and RF where RF shows the best accuracy (97.4%) and f-measure (0.974) for percentage split. Oladimeji et al. [40] preprocessed and balanced SDHD using Synthetic Minority Oversampling Technique (SMOTE). Afterward, Symmetrical Uncert Attribute Evaluator (SU), IGAE, GRAE, and CAE feature subsets were generated from that dataset, and RF, NB, J48, and KNN were applied to SDHD and its feature subsets. Thus, RF provided the best outcome with 98.31% Accuracy, 98.30% f-measure, and 99.90 AUROC respectively. Shahriare Satu et al. [9] implemented AdaBoost, NB, Bayes net (BN), MLP, LDA, QDA, KNN, sequential minimum optimization (SMO), simple logistic (SL), J48, and RF on PIDD where RF showed the best accuracy (99.067%), kappa statistics (98.09%), precision (99.10%), recall(99.10%), f-measure (99.10%), Matthews correlation coefficient (98.10%), AUROC (99.90%), area under the precision-recall curve (99.90%). Maniruzzaman et al. [41] applied NB, DT, AB, and RF into National Health and Nutrition Examination Survey(NHANES) dataset where RF gave the best outcomes including 94.25% accuracy, 96.88% f-measure and 95% AUROC. Different boosting algorithms such as AdaBoost, MultiBoost, real AdaBoost, Xgboost, GBM, LightGBM, and Catboost were also shown effective outcomes to detect diabetes. Kumar et al. [42] employed Catboost which achieved 100% accuracy to detect this disease. Taser [43] implemented tree-based classifiers namely C4.5, random tree, reduced error pruning tree (REPTree), decision stump, Hoeffding tree, NBTree, and some bagging and boosting approaches based on these classifiers on SDHD where AdaBoost and bagging with NBTree showed the best 98.65% accuracy.

Figure 5. Comparison of the models in terms of AUROC.

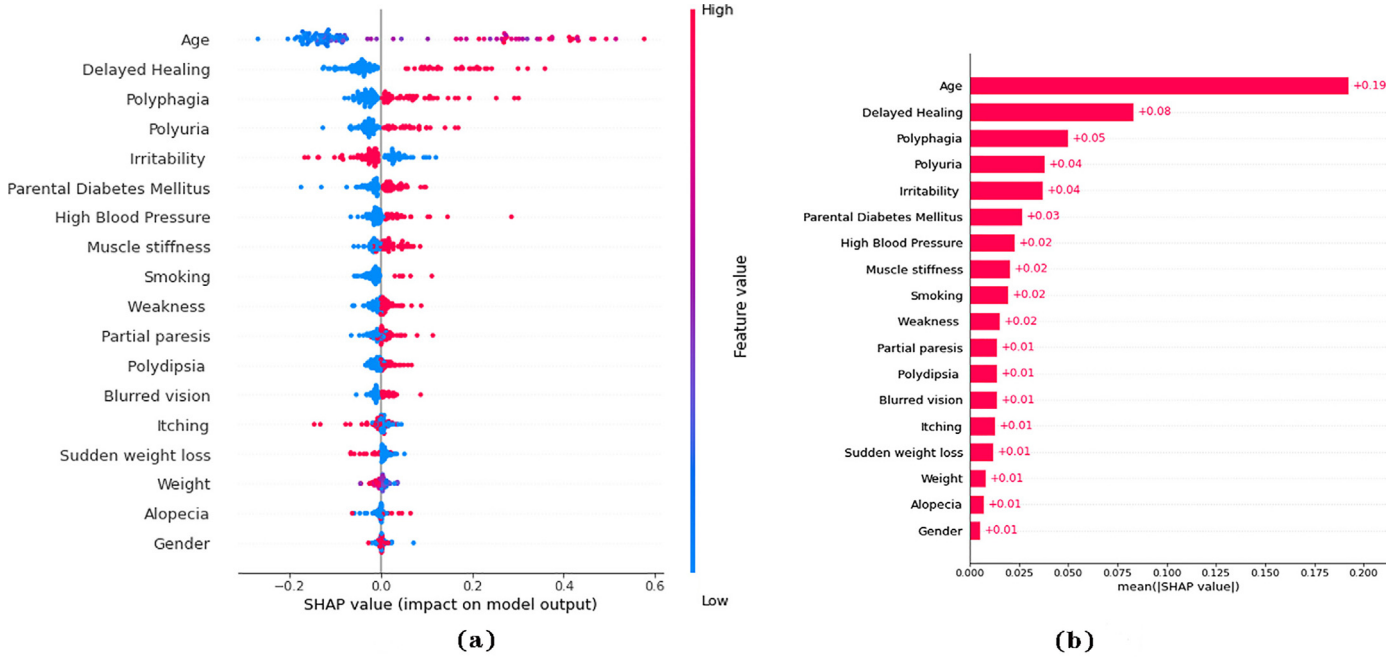

Figure 6. SHAP plots in case of PDD. (a) Feature importance impact using MLP. Each value is color coded, the blue color represents the lower value and the purple color represents the higher value of the attributes. (b) Feature importance plot for MLP. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Most of the works were happened based on PIDD which contains data from female patients whose age is above 21 years. Rahman et al. [44] proposed recurrent neural network-based convolutional long short-term memory (Conv-LSTM), convolutional neural network (CNN), traditional LSTM (T-LSTM), and CNN-LSTM on PIDD to detect diabetes. They used the Boruta algorithm for feature extraction and the Grid Search method to optimize parameters of individual classifiers and Conv-LSTM showed the best result with 97.26% accuracy. Naz and Ahuja [45] applied DT, ANN, NB, and deep learning (DL) after sampling PIDD for creating a balanced dataset and predicting diabetes. Also, Sahoo et al. [19] implemented seven classifiers such as KNN, LR, DT, RF, SVM, MLP, and CNN into PIDD and CNN gave the highest accuracy (93.2%). In PIDD, Zhu et al. [46] reduced dimensions using principal component analysis (PCA) and removed outliers using k-means, and finally applied LR (i.e. gave 97.40% accuracy) to detect diabetes. Apart from that, there are some studies where recent techniques have been employed to detect diabetes in the early days [47–50].

This study is explored various machine learning techniques to predict diabetes at the early stages. There were not happened sufficient works to investigate diabetes in developing countries like Bangladesh. Due to urbanization, the lifestyles of people are rapidly changing and most of them are not given priority to healthy habits. Therefore, the number of diabetes patients are increasing which accelerates the death rate and health expenditure of individuals. Due to a lack of datasets, more of the work did not happen. Hence, we focus on diabetes patients in such kinds of regions where we collected a variety of records from different locations in Bangladesh. Moreover, these instances were collected with very easy and confidential questions, therefore individuals found it better to respond accurately. In this work, we used numerous models to predict diabetes which did not occur in most of the works before.

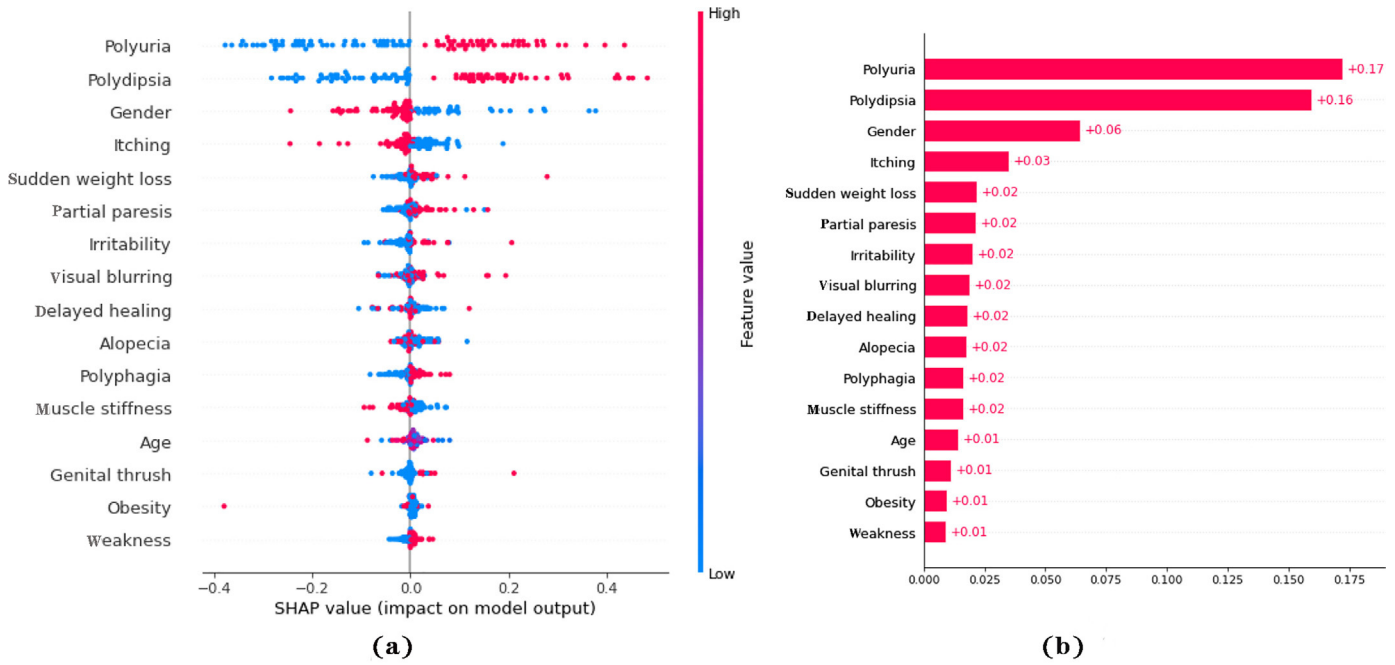

Figure 7. SHAP plots in case of SDHD. (a)Feature importance impact using ET. Each value is color-coded, the blue color represents the lower value and the purple color represents the higher value of the attributes. (b) Feature importance plot for ET. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

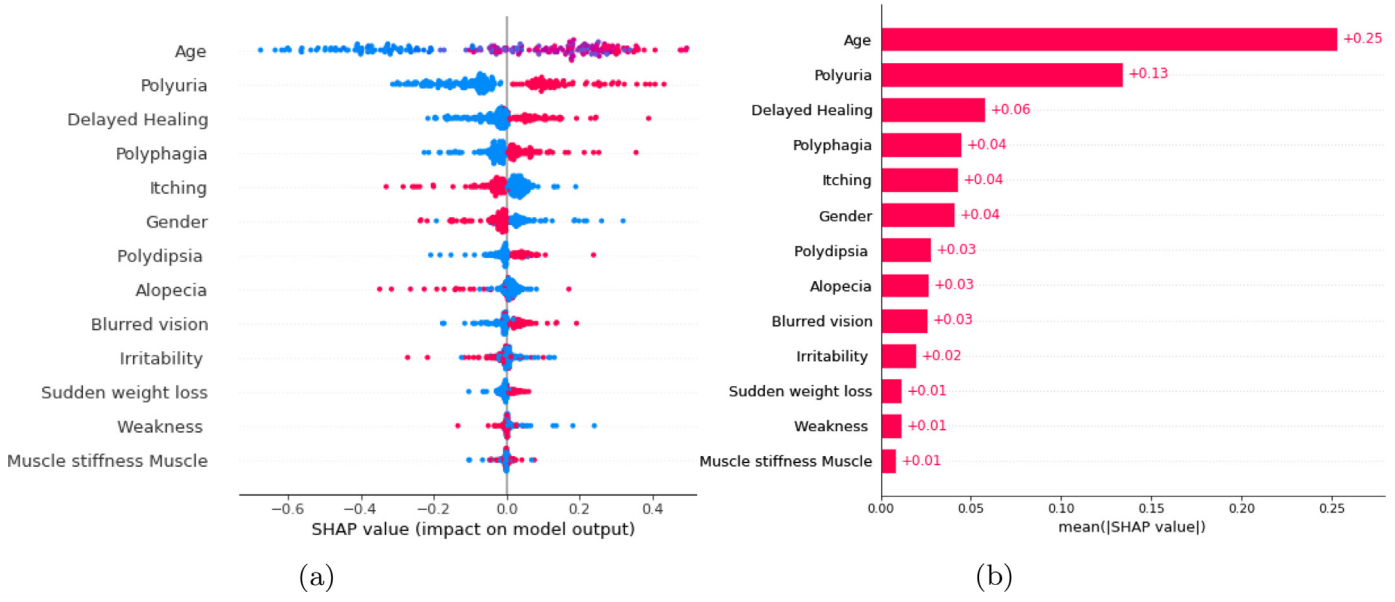

Figure 8. SHAP plots in case of MDD dataset. (a)Feature importance impact using HGBC. Each value is color-coded, the blue color represents the lower value and the purple color represents the higher value of the attributes. (b) Feature importance plot for HGBC. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Several works have happened on SDHD where a comparative analysis with some of them is depicted in Figure 9. It is observed that the proposed model provides better results than considering SDHD [7,51-53]. Finally, we interpreted the predicted results of these models by analyzing SHAP values and the influence of every feature have been identified more intuitively. In this work, we generated three distinct datasets where different classifiers provided some overfitted results for SDHD (See Table 2). Therefore, we collected more records using the almost similar questionnaire of SDHD where different classifiers are shown more stable results for PDD. Besides, individual classifiers are also given more stable results than SDHD. Therefore, we can reduce the overfitting issue of SDHD in this work. Again, we interpreted the best results for individual datasets using the SHAP method where different risk factors were determined which are highly responsible for happening diabetes. In this case, age, polyuria, polyphagia, delayed healing, and irritability are found as most dangerous factors for happening diabetes. On the other hand, gender, itching, alopecia, and weakness can be considered fewer risk factors of happening diabetes. Therefore, this analysis is helped physicians to detect diabetes more efficiently based on these factors. Due to these factors, individuals are being more careful about these factors and lead healthy lifestyles to protect against this disease. There are several points which are needed to consider such as the num-

Table 3 Experiments results of MDD (%)

Classifiers	Accuracy	Precision	Recall	F1-score	AUROC
LR	87.50	86.60	83.80	85.20	88.65
KNN	85.20	83.50	81.70	82.60	89.89
SVC	72.70	63.50	86.02	73.06	80.60
NB	83.30	79.40	82.80	81.05	84.07
DT	91.60	87.10	**94.60**	90.70	93.55
RF	92.10	88.00	**94.60**	91.20	94.11
SGD	76.80	67.70	88.20	76.60	85.09
Perceptron	63.40	88.90	17.20	28.80	69.07
AdaBoost	87.50	87.50	82.80	85.08	91.81
XGBoost	86.50	84.80	83.90	84.30	92.31
PAC	83.79	91.14	72.00	80.44	85.16
RC	87.03	91.86	79.00	84.94	87.34
Nu-SVC	81.48	73.80	93.00	82.30	83.52
LSVC	80.55	95.31	61.00	74.39	83.96
CCCV	87.96	90.21	83.00	86.46	90.07
NC	75.92	71.81	79.00	75.24	76.44
GP	93.05	92.07	93.00	92.53	89.71
MNB	82.41	82.98	78.00	80.41	82.02
Complement NB	83.33	83.33	80.00	81.63	81.03
BNB	81.94	82.79	77.00	79.79	81.34
Categorical NB	86.57	87.36	83.00	85.12	85.45
Bagging	92.59	94.68	89.00	91.75	92.80
ET	93.05	93.81	91.00	92.38	93.18
GBC	91.66	93.61	88.00	90.72	92.25
HGBC	**94.90**	**95.87**	93.00	**94.41**	94.92
OVsRC	79.16	72.00	90.00	79.99	80.60
MLP	89.35	90.52	86.00	88.20	90.88
LP	92.13	93.68	89.00	91.28	92.56
LS	91.66	92.70	89.00	90.81	92.07
Stacking	93.51	93.87	92.00	92.92	**95.47**
RCCV	87.04	91.86	79.00	84.94	87.75
LRCV	88.42	89.47	85.00	87.17	89.58
LDA	87.04	91.86	79.00	84.94	87.34
QDA	91.20	89.32	92.00	90.64	88.03
LGBM	**94.90**	**95.87**	93.00	**94.41**	94.54

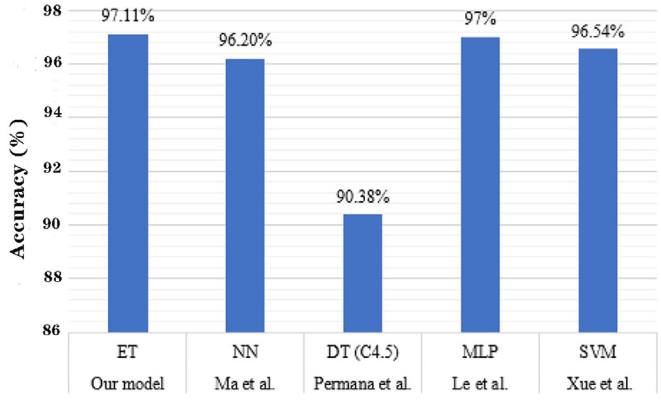

Figure 9. Comparison with existing works.

ber of data samples being low, no validation methods being used, and the performance of the predictive model is not verified using external data sources.

In conclusion, we investigated raw datasets of developing countries for detecting diabetes at an early stage. At first, we collected some data from UCI repository and generated a direct questionnaire using similar queries of SDHD and other related features. Then, this questionnaire was validated by experts and we collected instances from people throughout Bangladesh. Along with SDHD, we created PDD and MDD for further analysis. Then, we preprocessed these datasets and applied several classifiers which were employed in these datasets for predicting diabetes in patients. Then, the performance of every best classifier was interpreted by analyzing SHAP values. Therefore, we found several significant features which are extremely responsible for happening diabetes. We also compared this work with some other existing works where the proposed model showed better performance than others. But, we were not used more instances to investigate diabetes and required more clinical measurements to investigate diabetes. In the future, we will add more diverse instances as samples and investigate the early stage of diabetes in this region more accurately. On the other hand, there are some other important factors such as sleep deprivation and consumption of some prescribed drugs which need to consider for detection. Finally, we will design a web and mobile application to provide the advantage of the predictive machine learning model to a vast number of users at no cost.

Conflicts of interest statement

The authors declare that there are no conflicts of interest.

Funding

The data collection and data pre-processing of this research was supported by the University Grant Commission, Bangladesh under the research award (Award No: 37-01-0000-073-07-016-19/1759).

Author contributions

Nurjahan Nipa: Writing – original draft, Writing – review & editing. **Mahmudul Hasan Riyad:** Investigation. **Shahriare Satu:** Validation, Writing – review & editing. **Walliullah:** Investigation. **Koushik Chandra Howlader:** Data curation. **Mohammad Ali Moni:** Writing – review & editing.

References

[1] Gogebakan K, Sah M. A review of recent advances for preventing, diagnosis and treatment of diabetes mellitus using semantic web. In: Proceedings of 2021 3rd International congress on human-computer interaction, optimization and robotic applications (HORA). Ankara, Turkey: IEEE; 2021. doi:10.1109/HORA52670.2021.9461282.

[2] John JE, John NA. Imminent risk of COVID-19 in diabetes mellitus and undiagnosed diabetes mellitus patients. Pan Afr Med J 2020;36. doi:10.11604/pamj.2020.36.158.24011.

[3] Facts & figures. Available from https://idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html.

[4] Williams R, Karuranga S, Malanda B, et al. Global and regional estimates and projections of diabetes-related health expenditure: results from the international diabetes federation diabetes atlas, 9th edition. Diabetes Res Clin Pract 2020;162:108072. doi:10.1016/j.diabres.2020.108072.

[5] Afroz A, Alam K, Ali L, et al. Type 2 diabetes mellitus in Bangladesh: a prevalence based cost-of-illness study. BMC Health Serv Res 2019;19(1):601. doi:10.1186/s12913-019-4440-3.

[6] Islam MMF, Ferdousi R, Rahman S, et al. Likelihood prediction of diabetes at early stage using data mining techniques. Computer vision and machine intelligence in medical image analysis, 992. Singapore: Springer Singapore; 2020. p. 113–25. doi:10.1007/978-981-13-8798-2_12.

[7] Le TM, Vo TM, Pham TM, et al. A novel wrapper based feature selection for early diabetes prediction enhanced with a metaheuristic. IEEE Access 2021;9:7869–84. doi:10.1109/ACCESS.2020.3047942.

[8] Maniruzzaman M, Rahman MJ, Al-MehediHasan M, et al. Accurate diabetes risk stratification using machine learning: role of missing value and outliers. J Med Syst 2018;42(5):92. doi:10.1007/s10916-018-0940-7.

[9] Shahriare Satu M, Atik ST, Moni MA. A novel hybrid machine learning model to predict diabetes mellitus. Proceedings of international joint conference on computational intelligence. Singapore: Springer Singapore; 2020. p. 453–65. doi:10.1007/978-981-15-3607-6_36.

[10] Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn 2006;63(1):3–42. doi:10.1007/s10994-006-6226-1.

[11] Ishaq A, Sadiq S, Umer M, et al. Improving the prediction of heart failure patients survival using SMOTE and effective data mining techniques. IEEE Access 2021;9:39707–16. doi:10.1109/ACCESS.2021.3064084.

[12] Yadav DC, Pal S. An experimental study of diversity of diabetes disease features by bagging and boosting ensemble method with rule based machine learning classifier algorithms. SN Comput Sci 2021;2(1):50. doi:10.1007/s42979-020-00446-y.

[13] Kuo KM, Talley P, Kao Y, et al. A multi-class classification model for supporting the diagnosis of type II diabetes mellitus. PeerJ 2020;8:e9920. doi:10.7717/peerj.9920.

[14] Islam MT, Raihan M, Akash SRI, et al. Diabetes mellitus prediction using ensemble machine learning techniques. Advances in computational intelligence, security and internet of things. Vol, 1192. Singapore: Springer Singapore; 2020. p. 453–67. doi:10.1007/978-981-15-3666-3_37.

[15] Abbas HT, Alic L, Erraguntla M, et al. Predicting long-term type 2 diabetes with support vector machine using oral glucose tolerance test. PLoS ONE 2019;14(12):e0219636. doi:10.1371/journal.pone.0219636.
[16] Kaur H, Kumari V. Predictive modelling and analytics for diabetes using a machine learning approach. Appl Comput Inf 2020. doi:10.1016/j.aci.2018.12.004.
[17] Yu W, Liu T, Valdez R, et al. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC Med Inform Decis Mak 2010;10(1):16. doi:10.1186/1472-6947-10-16.
[18] Hasan MK, Alam MA, Das D, et al. Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access 2020;8:76516–31. doi:10.1109/ACCESS.2020.2989857.
[19] Sahoo AK, Pradhan C, Das H. Performance evaluation of different machine learning methods and deep-learning based convolutional neural network for health decision making. In: Nature inspired computing for data science, Vol. 871. Cham: Springer International Publishing; 2020. p. 201–12. doi:10.1007/978-3-030-33820-6_8.
[20] Hanbal IF, Ingosan JS, Oyam NAA, et al. Classifying wastes using random forests, gaussian nave bayes, support vector machine and multilayer perceptron. IOP Conf Ser Mater SciEng 2020;803:012017. doi:10.1088/1757-899X/803/1/012017.
[21] Cordeiro LS, Lima JS, Rocha Ribeiro AI, et al. Pill image classification using machine learning. 2019 8th Brazilian conference on intelligent systems (BRACIS). Salvador, Brazil: IEEE; 2019. p. 556–61. doi:10.1109/BRACIS.2019.00103.
[22] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. San Francisco California USA: ACM; 2016. p. 785–94. doi:10.1145/2939672.2939785.
[23] Athanasiou M, Sfrintzeri K, Zarkogianni K, et al. An explainable XGBoost based approach towards assessing the risk of cardiovascular disease in patients with Type 2 diabetes mellitus. Cincinnati, OH, USA: IEEE; 2020. p. 859–64. doi:10.1109/BIBE50027.2020.00146.
[24] Wang L, Wang X, Chen A, et al. Prediction of type 2 diabetes risk and its effect evaluation based on the XGBoost model. Healthcare 2020;8(3):247. doi:10.3390/healthcare8030247.
[25] Rashed-Al-Mahfuz M, Haque A, Azad A, et al. Clinically applicable machine learning approaches to identify attributes of chronic kidney disease (CKD) for use in low-cost diagnostic screening. IEEE J Transl Eng Health Med 2021;9:1–11. doi:10.1109/JTEHM.2021.3073629.
[26] Kopitar L, Kocbek P, Cilar L, et al. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. Sci Rep 2020;10(1):11981. doi:10.1038/s41598-020-68771-z.
[27] Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat 2001;29(5). doi:10.1214/aos/1013203451.
[28] Ghosh P, Azam S, Jonkman M, et al. Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques. IEEE Access 2021;9:19304–26. doi:10.1109/ACCESS.2021.3053759.
[29] Ke G, Meng Q, Finley T, et al. LightGBM: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst 2017;30:3146–54.
[30] Shobana G, Umamaheswari K. Prediction of liver disease using gradient boost machine learning techniques with feature scaling. Erode, India: IEEE; 2021. p. 1223–9. doi:10.1109/ICCMC51019.2021.9418333.
[31] Singh N, Singh P. A stacked generalization approach for diagnosis and prediction of type 2 diabetes mellitus. Advances in intelligent systems and computing. Singapore: Springer; 2020. p. 559–70. doi:10.1007/978-981-13-8676-3_47.
[32] Breiman L. Random forests. Mach Learn 2001;45(1):5–32. doi:10.1023/A:1010933404324.
[33] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in neural information processing systems, Vol 30. Curran Associates, Inc; 2017.
[34] Ribeiro MT, Singh S, Guestrin C. "Why should i trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. San Francisco California USA: ACM; 2016. p. 1135–44. doi:10.1145/2939672.2939778.
[35] Trumbelj E, Kononenko I. Explaining prediction models and individual predictions with feature contributions. Knowl Inf Syst 2014;41(3):647–65. doi:10.1007/s10115-013-0679-x.
[36] Chaves L, Marques G. Data mining techniques for early diagnosis of diabetes: a comparative study. Appl Sci 2021;11(5):2218. doi:10.3390/app11052218.
[37] Melville P, Mooney RJ. Constructing diverse classifier ensembles using artificial training examples. Eighteenth international joint conference on artificial intelligence; 2003. p. 505–10.
[38] Melville P, Mooney RJ. Creating diversity in ensembles using artificial data. Inf Fusion Special Issue on Diversity in Multiclassifier Syst 2004.
[39] Nurjahan, Rony MAT, Satu MS, et al. Mining significant features of diabetes through employing various classification methods. Dhaka, Bangladesh: IEEE; 2021. p. 240–4. doi:10.1109/ICICT4SD50815.2021.9397006.
[40] Oladimeji OO, Oladimeji A, Oladimeji O. Classification models for likelihood prediction of diabetes at early stage using feature selection. Appl Comput Inf 2021. doi:10.1108/ACI-01-2021-0022.
[41] Maniruzzaman M, Rahman MJ, Ahammed B, et al. Classification and prediction of diabetes disease using machine learning paradigm. Health Inf Sci Syst 2020;8(1):7. doi:10.1007/s13755-019-0095-z.
[42] Kumar PS, K AK, Mohapatra S, et al. CatBoost ensemble approach for diabetes risk prediction at early stages. Bhubaneswar, India: IEEE; 2021. p. 1–6. doi:10.1109/ODICON50556.2021.9428943.
[43] Taser PY. Application of bagging and boosting approaches using decision tree-based algorithms in diabetes risk prediction. Proceedings 2021;74(1):6. doi:10.3390/proceedings2021074006.
[44] Rahman M, Islam D, Mukti RJ, et al. A deep learning approach based on convolutional LSTM for detecting diabetes. Comput Biol Chem 2020;88:107329. doi:10.1016/j.compbiolchem.2020.107329.
[45] Naz H, Ahuja S. Deep learning approach for diabetes prediction using PIMA Indian dataset. J Diabetes Metab Disord 2020;19(1):391–403. doi:10.1007/s40200-020-00520-5.
[46] Zhu C, Idemudia CU, Feng W. Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. Inf Med Unlocked 2019;17:100179. doi:10.1016/j.imu.2019.100179.
[47] Hazarika BB, Gupta D. Random vector functional link with ϵ-insensitive Huber loss function for biomedical data classification. Comput Methods Programs Biomed 2022;215:106622. doi:10.1016/j.cmpb.2022.106622.
[48] Gupta D, Choudhury A, Gupta U, et al. Computational approach to clinical diagnosis of diabetes disease: a comparative study. Multimed Tools Appl 2021;80(20):30091–116. doi:10.1007/s11042-020-10242-8.
[49] Gupta D, Borah P, Sharma UM, et al. Data-driven mechanism based on fuzzy Lagrangian twin parametric-margin support vector machine for biomedical data analysis. Neural Comput Appl 2022;34(14):11335–45. doi:10.1007/s00521-021-05866-2.
[50] Kalita J, Balas VE, Borah S, et al. Recent developments in machine learning and data analytics: IC3 2018 Advances in intelligent systems and computing, 740. Singapore: Springer Singapore; 2019. doi:101007/978-981-13-1280-9.
[51] Ma J. Machine learning in predicting diabetes in the early stage. Proceedings of 2020 2nd International conference on machine learning, big data and business intelligence (MLBDBI). Taiyuan, China: IEEE; 2020. p. 167–72. doi:10.1109/MLBDBI51377.2020.00037.
[52] Permana BAC, Ahmad R, Bahtiar H, et al. Classification of diabetes disease using decision tree algorithm (C4.5). J Phys Conf Ser 2021;1869(1):012082. doi:10.1088/1742-6596/1869/1/012082.
[53] Xue J, Min F, Ma F. Research on diabetes prediction method based on machine learning. J Phys Conf Ser 2020;1684:012062. doi:10.1088/1742-6596/1684/1/012062.