Machine Learning Based Diabetes Detection Model for False Negative Reduction

Md Ashraf Uddin[1*], Md Manowarul Islam[2], Md. Alamin Talukder[2], Md. Al Amin Hossain[2], Arnisha Akhter[2], Sunil Aryal[1] and Maisha Muntaha[2]

[1]School of Information Technology, Deakin University, Geelong Waurn Ponds Campus, Australia.
[2]Department of Computer Science and Engineering, Jagannath University, Dhaka, Bangladesh.

*Corresponding author(s). E-mail(s): ashraf.uddin@deakin.edu.au;
Contributing authors: manowar@cse.jnu.ac.bd; alamintalukder.cse.jnu@gmail.com; hossainalamin980@gmail.com; arnisha@cse.jnu.ac.bd; sunil.aryal@deakin.edu.au; maishamuntahacsejnu@gmail.com;

Abstract

Diabetes is a chronic disease characterized by the inability of the pancreas to produce enough insulin or the body's inability to use insulin efficiently. This disease is becoming increasingly prevalent worldwide and can result in severe complications such as blindness, kidney failure, and stroke. Early detection of diabetes can potentially save millions of lives globally, making it a crucial focus of research. In this study, we propose a machine learning model to aid in predicting diabetes. The model comprises several machine learning methods: Linear Regression (LnR), Logistic Regression (LR), k-nearest neighbor (KNN), Naive Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), and Decision Tree (DT). Prior to feeding the pre-processed data into the machine learning model for evaluation, we conducted several pre-processing steps, such as removing null values, standardizing data using normalization, and labeling data using the label encoding process. Imbalanced datasets can adversely affect the accuracy of machine learning algorithms, and we address this problem by

balancing the datasets using the Synthetic Minority Oversampling Technique (SMOTE) method. We assessed the model's performance on two datasets and found that the random forest algorithm produced optimal results, with 97% accuracy on the diabetes dataset 2019 and 80% accuracy on the Pima Indian dataset. However, using a balanced dataset, we can significantly reduce the number of false-negative detections.

Keywords: Machine learning, pre-processed data, SMOTE, Random Forest, Balance data, Features selection

1 Introduction

Diabetes Mellitus, commonly known as diabetes, is a disease that can silently harm the body and cause various health complications. Among the reasons for diabetes are older age, poor diet, unhealthy lifestyle, and other factors. The normal range for a human's glucose level is typically between 70 and 99 milligrams per deciliter, and a person is considered diabetic if their fasting glucose level exceeds 126 mg/dL[1]. Currently, diabetes affects approximately 8.8% of the world's population, and this figure is expected to rise to 9.9% by the year 2045[2]. Common symptoms experienced by diabetic patients include weight loss, frequent weariness, irritation, dry mouth, burning, pain, numb feet, itching, reactive hypoglycemia, blurred vision, and erectile dysfunction, among others[3]. Although diabetes is not curable, early detection of the disease and adherence to a proper diet can help patients minimize the impact of this disease and prevent serious health problems in the future.

The advent of modern technology has facilitated the diagnosis and treatment of various illnesses. Machine learning and state-of-the-art technology have been employed to identify diseases such as diabetes, heart attacks, brain tumors, and cancer, among others [4–6]. Machine learning has enabled the development of models that can predict the likelihood of developing diabetes [7]. The use of machine learning models can aid in the early detection of diabetes.

Studies have been conducted to explore medical data and predict the likelihood of diabetes in the human body using machine learning and deep learning approaches. Ahmed et al. [7] evaluated several machine learning models to determine their accuracy in detecting diabetes in the Pima Indian diabetes dataset (PIDD) and diabetes dataset 2019. However, the effect of the imbalanced dataset was not considered in their study. Similarly, Islam et al. [8] used several machine learning models, including KNN, to diagnose diabetes in pregnant women. They extracted features using PCA (Principle Component Analysis) and a data mining approach. Maniruzzaman et al. [9] also investigated various machine learning models to detect diabetes by combining LR-based feature selection and RF-based classifier. Their works aimed to increase the accuracy of the diabetic detection model using different feature

selection methods. However, reducing the number of false negative cases is crucial. Therefore, our objective is to develop a machine-learning model that can balance the diabetes dataset to detect the disease at an early stage.

Several recent articles [10], [11], [12], [13] have explored various preprocessing and feature selection techniques to improve the accuracy of machine learning models in detecting diabetics. However, the available diabetic datasets typically lack an adequate number of features, which can result in the elimination of critical features during the selection process to boost the model's accuracy. Furthermore, a high accuracy rate does not necessarily indicate that most diabetics cases are being correctly classified, as diabetic datasets are often imbalanced, with a higher number of normal samples present. Miss-classifying a patient with diabetes as normal poses a greater danger than identifying a normal person as a diabetic patient. Thus, when training machine learning algorithms on imbalanced datasets with a larger number of normal samples, there is a risk of miss-classifying a diabetic patient. To mitigate this issue, we applied SMOTE data balancing techniques to reduce false negative cases. Our contributions are summarized below:

- To propose a new pipeline for diabetes prediction employing data balancing techniques using SMOTE and applying machine learning algorithms to predict diabetes.
- To implement the models and analyze the performance of machine learning algorithms in terms of different metrics, including accuracy, recall, precision, and confusion matrix. Finally, the best model is selected for diabetes prediction based on classifications' performance, reducing the risk of diabetes mortality.

The subsequent sections of this paper briefly present the study, which consists of a literature review in Section 2, a description of the proposed methodology and datasets in Section 3, an overview of the experiment setup and evaluation in Section 4, and lastly, conclusions and future work in Section 5.

2 Literature Review

This section provides a brief summary of works pertinent to diabetes prediction. The Pima Indians Diabetes Dataset has been used extensively in studies to predict diabetes in individuals at risk by employing a variety of machine learning approaches. We evaluate the approaches employed in these investigations and find the most accurate models. Given the global incidence of diabetes, there is an abundance of ongoing study on this topic.

Olisah et al.[10] proposed a framework comprising of a support vector machine, a decision tree, and their original twice-growth deep learning network. To select crucial features, they utilized Pearson correlation and polynomial regression techniques. The performance of the model was then evaluated using stratified k-fold cross-validation and grid-search hyperparameter tuning.

To prepare the data for a model, Hasan et al.[11] employed several preprocessing techniques, such as removing outliers, handling missing values, standardizing the data, and selecting crucial features. The model itself comprised of multiple machine learning models, including KNN, DT, RF, NB, AdaBoost, and XGBoost, and the results were generated through cross-validation. Similarly, After evaluating multiple machine learning models, including Artificial Neural Networks (ANNs) and NB, Naz et al. [12] recommended decision tree algorithms for predicting diabetic diseases. The experiment demonstrated that both DT and Deep Learning (DL) achieved an accuracy range of 90-98%. Notably, for the PIMA dataset, the Deep Learning model yielded the highest accuracy (98.07%) in predicting the onset of diabetes. A machine learning model was proposed by Krisha et al. [13] for identifying individuals with diabetes, which included a support vector machine and a decision tree. The model was trained on the PIDD dataset and achieved an 86% accuracy rate in detecting the disease.

Given the recent surge in the prevalence of diabetes, Decision Support Systems are aiding healthcare professionals in identifying this condition at an early stage. Yahyaoui et al. [14] evaluated the efficacy of two commonly used machine learning models, namely, support vector machine and random forest, and a deep learning model, CNN, in detecting diabetes using the Pima Indians Diabetes dataset. The results indicated an accuracy of 83.6% for the model. Interestingly, the study findings revealed that the machine learning algorithms outperformed the deep learning model in terms of detection accuracy.

Islam et al. [8] utilized various machine learning algorithms, including SVM, LR, NB, RF, and KNN, along with PCA and data mining techniques to detect diabetes in pregnant women. Their study reported a high accuracy rate of 87% using the SVM algorithm. Tapak et al. [15] used different algorithms, such as ANN, SVM, Fuzzy C-mean, and RF, and employed a 10-fold cross-validation method to determine the best accuracy rate. SVM showed the highest accuracy rate of 98.6%.

Swapna G. et al. [2] applied deep learning methods, such as Recurrent Neural Networks (RNN), Short-Term Long Memory (LSTM), Convolutional Neural Networks (CNN), Hybrid networks (CNN-LSTM), and SVM, to diagnose diabetes by analyzing ECG signal of heart rate. Their study reported that the CNN 5-LSTM with the SVM method showed the highest accuracy rate of 95.7%. Ahmed et al. [7] evaluated the performance of seven machine learning algorithms, namely, DT, NB, KNN, RF, GB, LR, and SVM, on four different datasets. The results showed varying levels of accuracy using different datasets.

Joshi et al. [16] compared SVM, Logistic Regression, and ANN to predict diabetes and proposed a technique that could detect the disease at an earlier stage. Meng et al. [17] compared LR, ANN, and DT methods for predicting the risk of diabetes and prediabetes based on 12 risk factors, including education level, work stress, BMI, age, sleep duration, gender, marital status, family history of diabetes, coffee consumption, preference for salty foods, physical activity, and fish consumption. DT was shown to be the most effective strategy of the three.

Nongyao et al. [18] examined four classification techniques: decision tree, artificial neural network, logistic regression, and naive Bayes. All were subjected to additional bagging, boosting, and random forest. All of the participants' greatest accuracy was between 84% and 86%. Choi et al. [19] used machine learning algorithms on patients with a history of non-diabetes and a high risk of cardiovascular disease. Korea University Guro Hospital has collected data in the form of an EMR over the past five years. Then, using 10-fold cross-validation, machine learning algorithms were used. The LR model has the highest accuracy. Tigga et al. [1] used six different machine-learning algorithms. They are Logistic regression, Naive Bayes, Support vector machine, Decision tree, KNN, and RF. And then, accuracy was compared. They found a greater accuracy of 94% using a random forest algorithm.

Maniruzzaman et al. [9] used LnR, LR, NB, RF model, KNN, DT, and AdaBoost and found 94.25% accuracy using the combination of LR-based feature selection and RF-based classifier. Zhang et al. [20] Early prediction of gestational diabetes mellitus in the Chinese population via advanced machine learning. They used different datasets and found 80% of the best accuracy using Deep Neural Network.

3 Methodology

In this section, we discuss the premise of our proposal and the machine learning models used to classify diabetes in the preprocessed dataset. Our proposed diabetic prediction model employs machine learning techniques to balance the dataset. We begin by labeling, scaling, and balancing the dataset to detect the potential presence of diabetes in the human body. Then, using the balanced dataset, we train several machine learning algorithms, including LR, LrR, SVM, NB, DT, RF, and k-kNN, and evaluate their accuracies. Next, we select the best model for prediction based on the model's performance. Finally, we apply the chosen model to other datasets to compare its performance with other models.

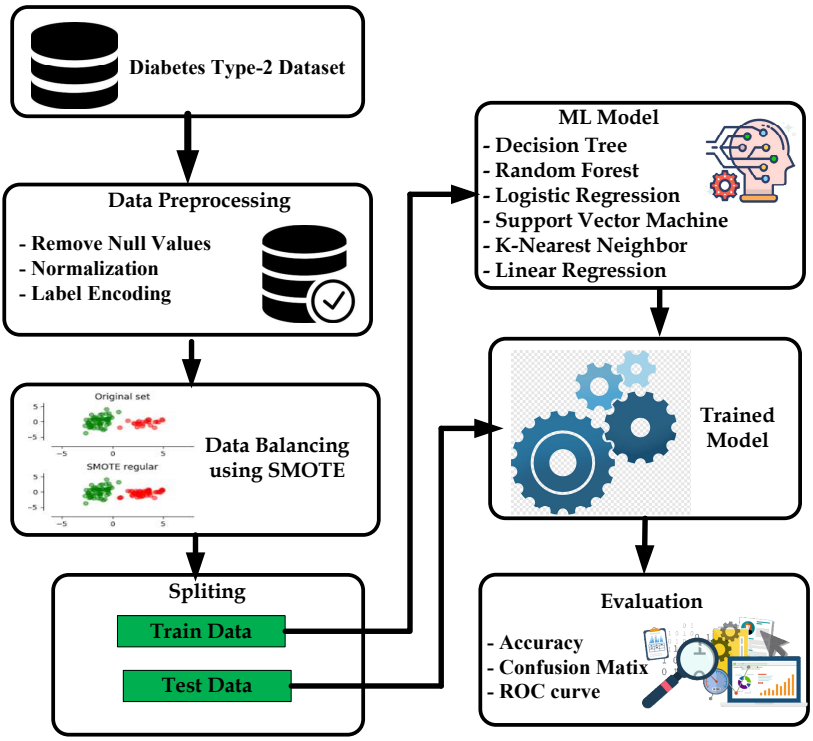


Fig. 1 The initial step in this study involved the acquisition of a type 2 diabetes dataset. Subsequently, the dataset underwent preprocessing to handle missing values and remove outliers. The preprocessed dataset was found to be imbalanced, which prompted the use of SMOTE to increase minority samples and decrease majority samples to obtain a balanced dataset. The balanced dataset was then split into training and testing sets. A range of machine learning algorithms, including decision tree, support vector machine, and random forest, were trained on the balanced datasets. Grid search techniques were employed to fine-tune the parameters of the machine learning algorithms.

Figure 1 depicts the proposal's workflow, in which we preprocessed the dataset by removing null values, data normalization using standardization, and data labeling using the label encoding process. Finally, the processed data is balanced using SMOTE and is split for feeding them into machine learning algorithms to analyze the performance.

- **Data collection :** We have collected two datasets for training the ML model. One is Pima Indian diabetes datasets, and another is diabetes datasets 2019. In Pima Indian datasets, there are eight features; in diabetes datasets 2019, there are seventeen features.
- **Data pre-processing :** After collecting the datasets, we preprocess data to train our models. First, we remove the null value from the datasets. Next, we visualize the characteristics of datasets. After that, data is encoded using

one hot encoding and label encoding. Label encoding entails changing each value in a column to a number.

- **Data scaling :** We scale the datasets using a min-max scaler to improve the model's performance.
- **Dataset balance:** The visualization of the Diabetes dataset 2019 shows that it is an imbalanced dataset. Therefore, to acquire an accurate outcome, we balance the dataset using SMOTE method.
  - **SMOTE :** SMOTE is a synthetic tabular data generation approach based on over-sampling. The main idea behind SMOTE is to create synthetic data between each minority class sample and its " k " closest neighbors.
- **Splitting Train and Test Data :** We split the dataset into train data and test data. We take a different amount of train and test data to fit into the model; each case gives an additional value.
- **Traing models:** We have used seven different models to train using our preprocessed, balanced, and imbalanced data and test the models with the test data.
- **Performance analysis:** After testing seven models using Test data, we found the accuracy of the models and showed the confusion matrix.

3.1 Machine learning Algorithms

In this module, we train several models using machine learning. The main functions of this module can be described as follows:

- Linear Regression: The linear regression algorithm illustrates the linear connection between one or more independent (X) variables and a dependent (Y) variable. Since linear regression reveals a linear correlation, it determines how the variation in the value of the dependent variable changes with the variation in the independent variable. Mathematically, this relationship depicted in Figure 2(a) can be expressed using the following equation:

$$Y = mX + b$$

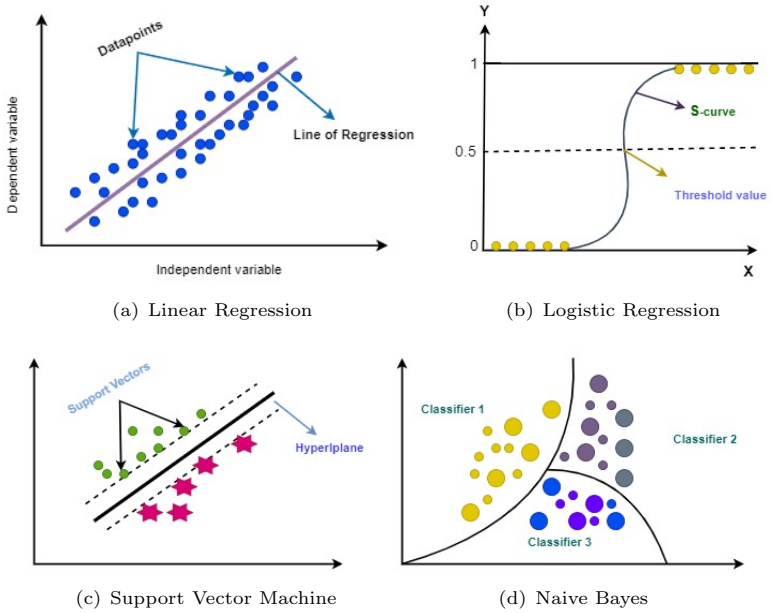
.

- Logistic regression: This algorithm is used for binary classification in ML. This finds the best-fit line between two types of data sets. For finding the best-fit line, there is a function called the sigmoid function is used. The best-fit line is like an S shape, and there is a middle point of the function the decision is made compared with this point. This performs better than linear regression. The model is depicted in Figure 2(b).

$$S_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

- Support Vector Machine: SVM is a well-known supervised learning technique that can be utilized for classification and regression tasks. This starts by finding a linear boundary between two classes of data and then calculates

the margin distance while drawing two extra hyperplanes. This helps in achieving a better decision-making outcome. The goal is to maximize the margin distance in this scenario. The support vectors are utilized to define the hyperplanes, and SVM leverages them to determine the hyperplanes. Figure 2(c) shows the diagram of support vector machine.



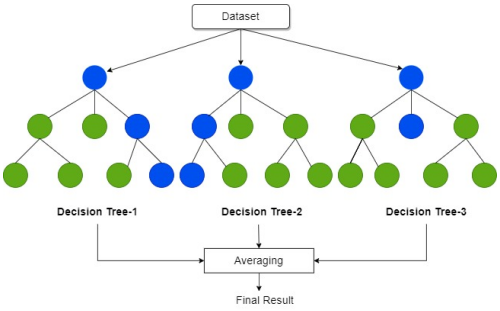
(a) Linear Regression (b) Logistic Regression

(c) Support Vector Machine (d) Naive Bayes

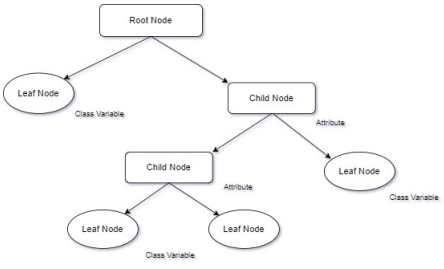
Fig. 2 Machine Learning model

- Naive Bayes (NB): NB algorithm is a classification method that employs Bayes' theorem and assumes that predictors are independent. In simpler terms, this assumes that the presence of one feature in a class is unrelated to the presence of any other feature. The Naive Bayes model is easy to construct and is highly advantageous for large datasets. Despite its simplicity, Naive Bayes has been known to outperform even the most advanced classification methods. There are three types of Naive Bayes algorithms, which are listed below. Figure 2 (d) show the model of Niave Bayes.
  - **Gaussian**: The assumption of the Gaussian model is that the features adhere to a normal distribution.
  - **Multinomial:**The Multinomial Naive Bayes classifier is employed when the data conforms to a multinomial distribution.
  - **Bernoulli:** The Bernoulli classifier operates in a similar fashion to the Multinomial classifier, but it assumes that the predictor variables are independent Boolean variables. However, for our study, we have employed the Multinomial Naive Bayes classifier.

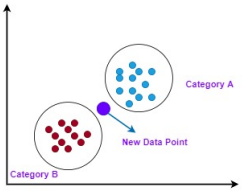
- Random forest: RF is a technique based on decision trees that are used in modeling predictions and behavior analysis. It contains many decision trees, each of which represents a unique instance of the classification of data input into the random forest depicted in figure 3(a). The random forest technique considers each instance individually, selecting the one with the most votes as the selected prediction.
- Decision tree: The DT is the most prominent classification and prediction tool. A Decision tree is a structure similar to a chart, in which each internal node represents a test on an attribute, each branch represents a test outcome, and each leaf node (terminal node) holds a class value. A Decision Tree is a supervised learning method that can be used to solve both classification and regression problems, while it is most typically utilized to address classification issues. Nodes in the network represent dataset variables, branching represents decision rules, and each leaf node provides the conclusion in this tree-structured classification. The Decision Node and the Child Nodes are the two types of nodes of a Decision tree. Child nodes are the result of those judgments and do not hold any additional branches, whereas Judgment nodes are used to make a decision and have multiple branches. The figure 3 (b) shows the picture of decision tree.



(a) Random Forest



(b) Decision Tree



(c) K Nearest Neighbour

Fig. 3 Machine Learning model

- K-Nearest Neighbour: KNN, short for K-Nearest Neighbors, is a popular Supervised Learning algorithm that predicts the output of new data points based on a labeled set of known inputs. This simple yet effective machine learning approach can be applied to diverse problems and relies on feature similarity. KNN is sometimes referred to as a lazy learner algorithm since it does not immediately learn from the training set. Instead, it stores the dataset and applies a computation during classification. Figure 3(c) shows the total picture of knn algorithm.

In the upcoming section, we will elaborate on the architectures of different deep learning models that we have utilized for training and learning the system.

4 Experimental setup and Evaluations

This section presents the evaluation of our proposed approach utilizing a range of machine learning algorithms, including NB, DT, RF, SVM, LR, LnR, and KNN. Initially, we provide a description of the dataset employed in the experiments. Subsequently, we conduct various experiments to assess our approach and present the results in detail. We employ accuracy bar diagrams, confusion matrices, and classification reports to compare the outcomes and arrive at an informed decision.

4.1 Dataset Description

In our study, we utilized two datasets to assess our model's performance: the Pima Indian Diabetes dataset and the Diabetes Datasets 2019. We obtained both datasets from Kaggle.com[21], an online platform for data science tutorials and competitions. One dataset was balanced, while the other was imbalanced. The Pima Indian Diabetes dataset was balanced, whereas the Diabetes Datasets 2019 was imbalanced. In this section, we provide a detailed description of both datasets, including their features and characteristics.

4.1.1 Pima Indian dataset

The Pima Indian dataset comprises eight features and a single class label indicating whether the individual has diabetes or not. With 768 rows, each feature has distinct characteristics, which we outline below.

- **Pregnancies:**Number of times being pregnant.
- **Glucose:** Plasma glucose concentration of 2 hours in an oral glucose tolerance test.
- **Blood Pressure:**Diastolic blood pressure (mm Hg).
- **Skin Thickness:** Triceps skin fold thickness (mm).
- **Insulin:** 2-Hour serum insulin (mu U/ml).
- **BMI:** Body mass index.
- **Diabetes Pedigree Function:** The Diabetic Pedigree Function (DPF) is a mathematical value that shows the probability that a patient would develop

diabetes based on their family history of the condition. It is determined by analysing the patient's family tree and taking into account the number and degree of diabetic relatives. The DPF is often utilised as a predictor in diabetes risk assessment models. A larger DPF number suggests an increased risk for getting diabetes.

- **Age :** Age in years
- **Outcome :** Class variable (0 or 1)

4.1.2 Anylysis of PIMA Indian datasets:

We now move on to exploratory data analysis of the variables of this dataset that are involved in the study. Table 1 describes the first 10 records of datasets.

Table 1 PIMA Indian dataset

Pregnancies	Glucose	Blood pressure	Skink thickness	Insulin	BMI	DPF	Age	Target
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	23.5	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.2288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31.0	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0.0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0

The dataset comprises 768 records and a total of 9 characteristics, each of which can be of integer or float data type. Some features, including blood pressure, skin thickness, insulin, and BMI, have missing values indicated by zero. The outcome column denotes 1 for positive diabetes detection and 0 for negative. Out of 768 data points, 500 are labeled as 0, and 268 are labeled as 1, as depicted in Figure 4. The development and assessment of classification models has been done using the Pima Indian Diabetes dataset in a number of studies and research papers. This dataset has been subjected to the application of a number of well-known classification techniques, including decision trees, random forests, logistic regression, and support vector machines. The models' reported accuracy vary, with some obtaining rates as high as 80–85%.

Figure 4 displays the correlation among the variables in the dataset. The heatmap indicates that none of the parameters are actually dependent on one another.

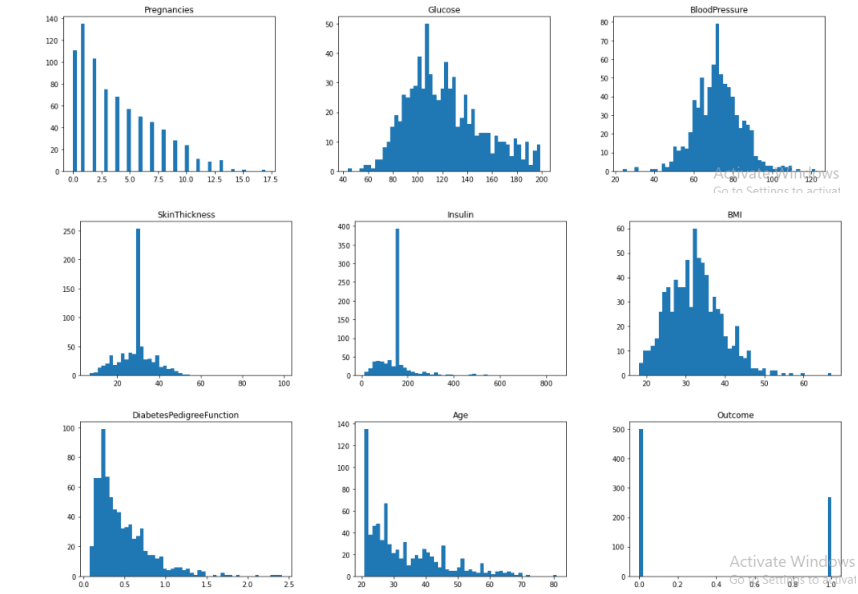
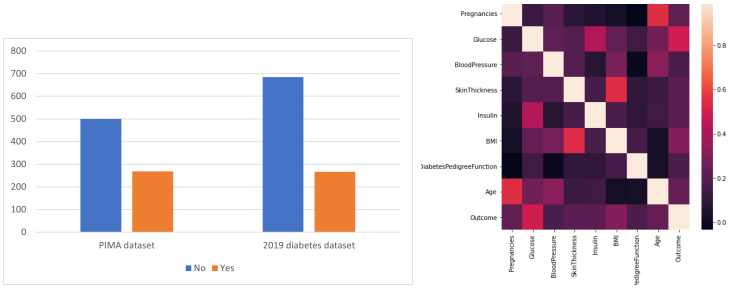


Fig. 4 PIMA Indian dataset histogram

The histogram of the dataset is presented in Figure 4. Based on the histogram, This can be inferred that the data does not contain a significant number of outliers.



(a) PIMA and 2019 Diabetes dataset (b) Heatmap for PIMA

Fig. 5 Data distribution

Figure 5(b) displays the correlation among the variables in the dataset. Based on the heat map, this can be concluded that none of the parameters are dependent on each other.

Table 2 2019 Diabetes dataset

Features	row-1	row-2	row-3
Age	50-59	50-59	40-49
Gender	Male	Male	Male
Diabetes History	no	no	no
High BP	yes	yes	no
Physically Active	1 hr or more	less than 1.5hr	1 hr or more
BMI	39	28	23
Smoking	no	no	no
Alcohol	no	no	no
Sleep	8	8	8
Sound Sleep	6	6	6
Regular Medication	no	yes	no
Junk Food Consumption	occasionally	very often	occasionally
Stress	sometimes	sometimes	sometimes
BP level	high	normal	normal

4.1.3 Analysis of Diabetes Dataset 2019

The diabetes datasets 2019 comprises of seventeen features and one class label, with a total of 952 rows. However, the dataset is imbalanced, with a significant number of nondiabetic data compared to diabetic data. Table 2 displays a segment of the dataset, which consists of both numeric and string values.

The diabetes dataset of 2019 has a total of seventeen features and one class label, with 952 rows in total. The Diabetes 2019 dataset is a collection of medical records of patients who were tested for diabetes. The dataset contains 17 features including age, sex, BMI, blood pressure, serum measurements, and others, and a binary outcome indicating whether or not the patient tested positive for diabetes. The dataset consists of 952 records and includes both numeric and string values. The class label is binary and contains values "yes" or "no," indicating the presence or absence of diabetes in the body. The dataset is imbalanced, with a significantly higher number of nondiabetic cases compared to diabetic ones. Table 2 displays a subset of the dataset containing both numeric and string values. In particular, the "diabetic" column has two values, "yes" and "no," indicating the presence or absence of diabetes, respectively. Counting the values of diabetic and nondiabetic instances, there are 684 persons who are diabetic and 266 who are not, indicating that the dataset is imbalanced. The diabetic and nondiabetic value count is displayed in Figure 5.

4.2 Experimental Setup

We performed data preprocessing to improve the quality of our datasets before training our model. The details of this preprocessing are described in the following section.

4.2.1 Data Preprocessing

Data preprocessing is a crucial step to ensure effective training of our models. In this section, we eliminate null values from the datasets by either using

the mean or median value of the columns. This results in well-defined and compatible data for training our models. In the Pima Indian datasets, several columns contain zero values which are treated as null and are removed. For the diabetes dataset 2019, we replace null values with mean and median values. Additionally, we label the data using label encoding and one hot encoding to improve the model's performance. We also scale the dataset using min-max scaler and standard scaler.

4.2.2 Data Balancing

Imbalanced datasets refer to datasets that have a lower ratio between the class-level data. The Diabetes dataset 2019 is an example of an imbalanced dataset. To balance the dataset, there are several techniques available. Two common approaches are undersampling and oversampling. In undersampling, the dataset is balanced by deleting some records from the majority class to match the minority class. On the other hand, oversampling increases the minority class by replicating the data or creating synthetic data. In this study, oversampling was performed using SMOTE (Synthetic Minority Oversampling Technique). This method was proposed in 2002 and involves generating synthetic examples in the minority class by interpolating between existing examples. After balancing the dataset, the count for both class labels became the same.

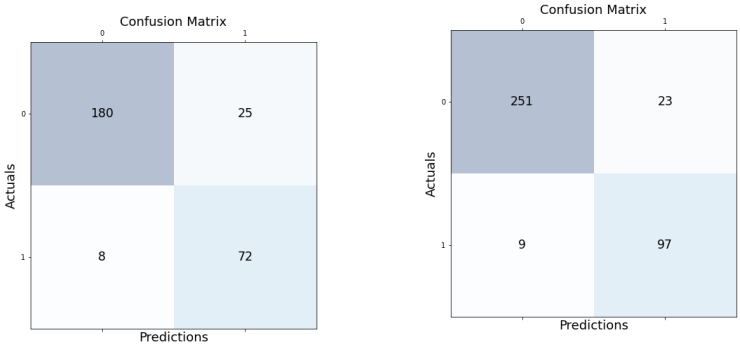
4.3 Experiments

We have split the data into different training and testing sets. Train the model with both balanced and imbalanced data and observe the result. We split the data into four different ratios and apply the same model and measure the performance in each case. In the next part, we will the observation in different cases.

4.3.1 Result of Experiment 1

In the initial case, we evaluate the performance of the model using balanced datasets. We train our model using the diabetes dataset 2019 extracted from the PIMA Indian dataset. After balancing the dataset, we fit it into our model and measure the accuracy of each classification algorithm. The confusion matrix is also presented to evaluate the performance.

Next, we examine the confusion matrix for our models. The confusion matrix for the balanced dataset using the Logistic Regression algorithm is presented in Figure 6 (a), while the one for the SVM algorithm is shown in Figure 6(b).



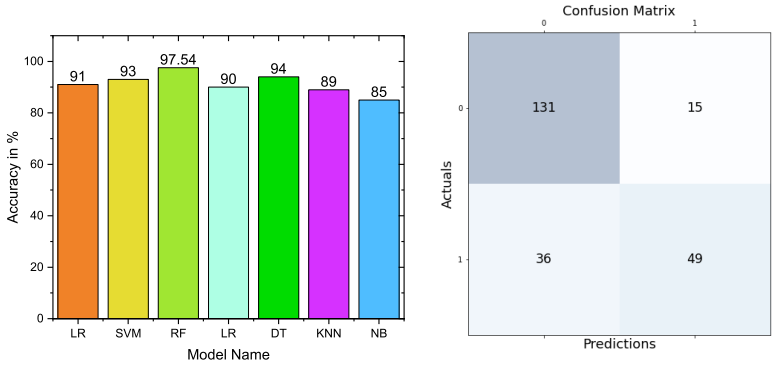
(a) Logistic Regression model (b) SVM model

Fig. 6 Confusion matrix for balanced 2019 dataset

In the Logistic regression confusion matrix number of false negatives is only eight as we balanced the dataset. In SVM, there are only nine false negatives.

Now we observe the best accuracy of the Pima Indian dataset. We found the best accuracy when we used 25 % test data and 75 % train data. The accuracy comparison is shown in Figure 7(a).

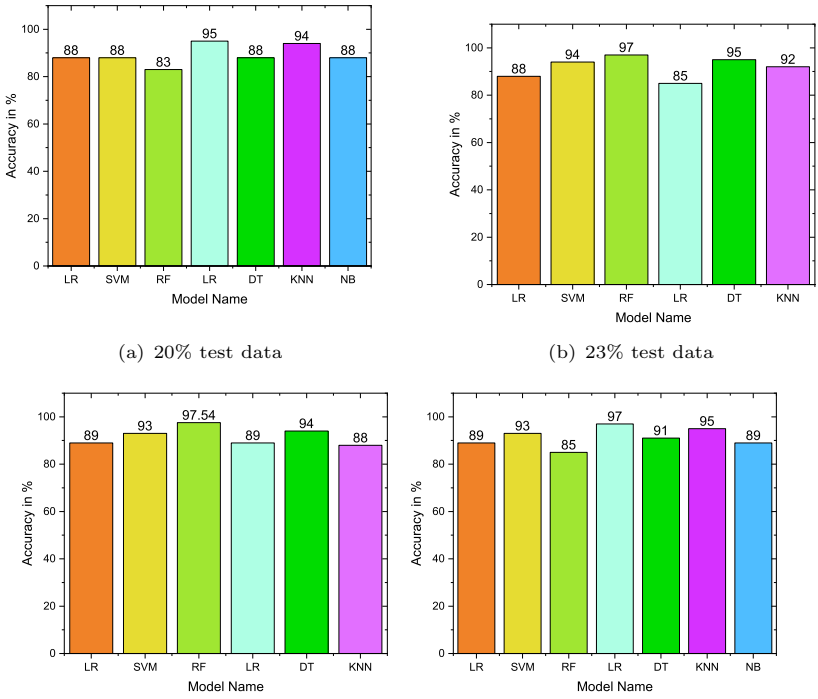
We can find the best accuracy of 97.54 % using the Random Forest algorithm for the Pima Indian datasets. The second best classification algorithm for this dataset is Decision Tree. Figure 7 (b) shows the pima indian datasets as one of the better-performed confusion matrix. In this case, the number of false negative cases is 39.



(a) Accuracy (b) Confusion matrix

Fig. 7 Performance of Pima Indian dataset

In this section, we compare the accuracy of the balanced 2019 diabetics datasets in different train-test split scenarios (testing data portion: 20%, 25%, 30%, and 40%). Figure 8 illustrates the comparison of accuracy among algorithms with 30% test data in the balanced dataset of 2019.



(a) 20% test data

(b) 23% test data

(c) 30% test data

(d) 40% test data

Fig. 8 The accuracy of balanced 2019 Diabetics dataset

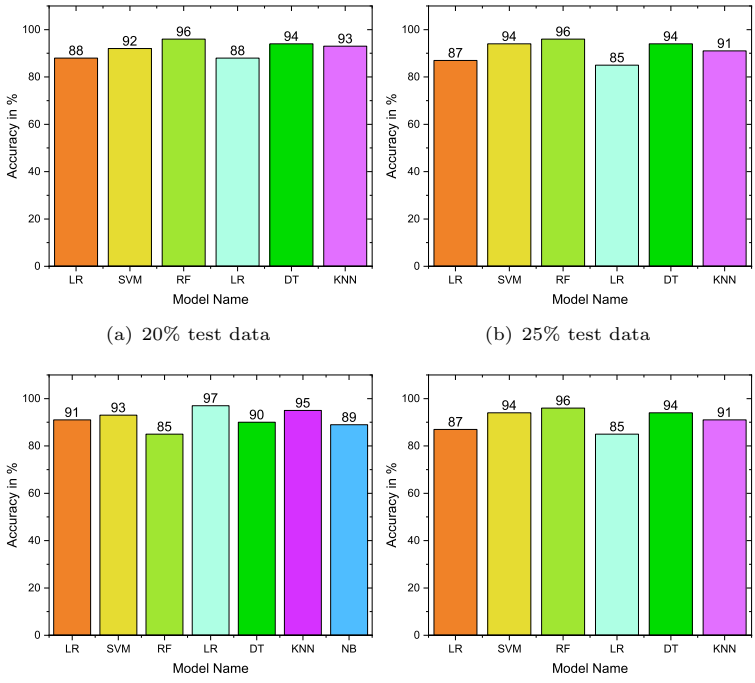
4.3.2 Result of Experiment 2

In the second case, we exclusively use the imbalanced diabetes dataset of 2019. We split the dataset into different train-test cases and analyze the model's characteristics and their performance. Figure 9 below presents the accuracy comparison of each model in different split cases.

Figure 8 and 9 depict the accuracy of various classifiers when applied to diverse datasets with varying levels of data balance and test data size. In the analysis, LR, SVM, RF, LnR, DT, KNN, and NB are used as classifiers. Random forest obtains the highest accuracy score of 97.54% when the data is balanced with 30% test data, followed by SVM and decision tree with 93% and 94% accuracy, respectively. Nave Bayes performs poorly with balanced data.

In contrast, when the data is imbalanced with 30% test data, random forest achieves the highest accuracy with 97.54 percent, followed by SVM and decision tree with 93 percent accuracy. With an accuracy rating of 85%, Naive Bayes performs inadequately.

When the test data size is reduced to 25%, the accuracy scores of all classifiers decrease marginally. The random forest obtains an accuracy score of 97% with balanced data, which is still its best performance. For unbalanced data,



(a) 20% test data (b) 25% test data

(c) 30% test data (d) 40% test data

Fig. 9 The accuracy of imbalanced 2019 Diabetics dataset

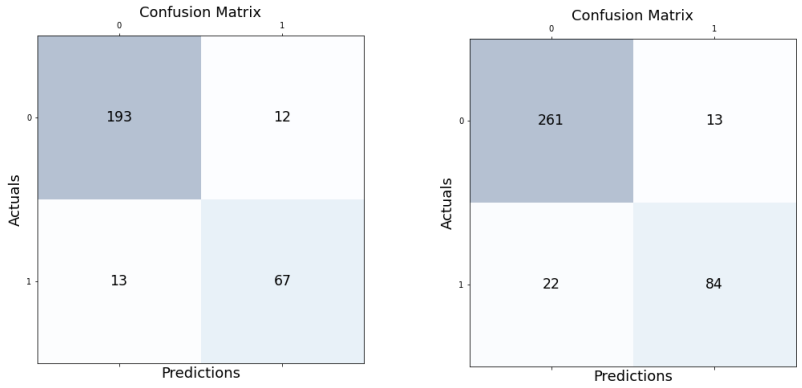
decision tree and random forest obtain 95% accuracy with the highest accuracy score. When the size of the test data is increased to 40%, the accuracy scores of all classifiers diminish in comparison to the 30% test data scenario. With balanced data, random forest continues to perform best, obtaining an accuracy score of 96%. Again, decision tree and random forest have the maximum accuracy for unbalanced data at 94%.

Finally, when the test data size is reduced to 20%, the accuracy scores for all classifiers, with the exception of linear regression, remain relatively stable. With balanced data, random forest achieves an accuracy score of 96%, while random forest and linear regression achieve an accuracy score of 97% for imbalanced data.

Overall, the results indicate that random forest performs the most consistently across various levels of data balance and test data size, followed by decision tree and SVM. The Nave Bayes algorithm consistently yields low accuracy scores, particularly with balanced data. The results also indicate that accuracy scores decrease as the size of the test data increases, which is to be anticipated given that the model has less training data.

The confusion matrix of logistic regression and svm model for imbalanced data set is shown in Figure 10. For imbalanced 2019 diabetics dataset, the

model showed higher number of false negative cases than that of balanced 2019 diabetics dataset.



(a) Logistic Regression model (b) SVM mode

Fig. 10 Confusion matrix for imbalanced 2019 dataset.

Figure 11 illustrates the error rate of six machine learning models for the 2019 diabetics dataset at various test data proportions. Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Linear Regression (LinR), Decision Tree (DT), and K-Nearest Neighbours (KNN) are among the models. The error rate is expressed as a percentage for each model and test data proportion. The Random Forest model has the lowest error rate for all proportions of test data compared to the other models, as shown in Figure 11. Additionally, the Decision Tree model has a relatively low error rate. SVM also has a low error rate, but its performance is marginally inferior to that of the Random Forest and Decision Tree models. The error rates of the Logistic Regression, Linear Regression, and KNN models are greater than those of the other models. In addition, we can observe that increasing the proportion of test data reduces the error rate of all models.

Table 3 Accuracy Results for diabetes dataset 2019

Model Name	Training Data and Testing data	Accuracy using balanced dataset	Accuracy using imbalanced dataset
Linear regression	70%:30%	89.47%	90.88%
Logistic regression	70%:30%	89%	91%
SVM	70%:30%	93%	93%
Random forest	70%:30%	97.54%	97.54%
Naive Bayes	70%:30%		85%
Decision tree	70%:30%	95.43%	95.08%
KNN	80%:20%	92%	91%

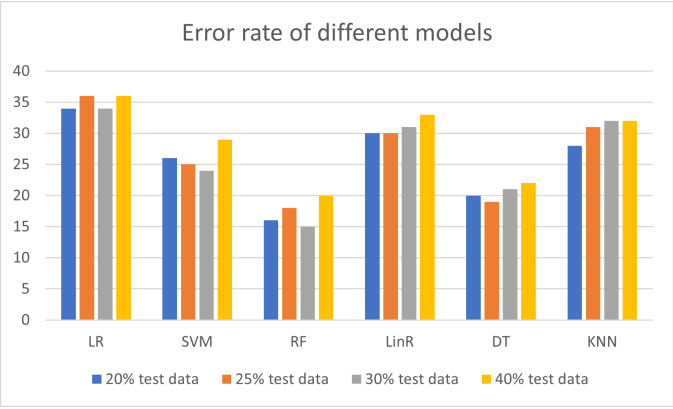


Fig. 11 Error rate of different machine learning models

4.3.3 Cross Validation Result

We show the score comparison of pima indian dataset and diabetes dataset 2019 for 10-fold cross validaiton.

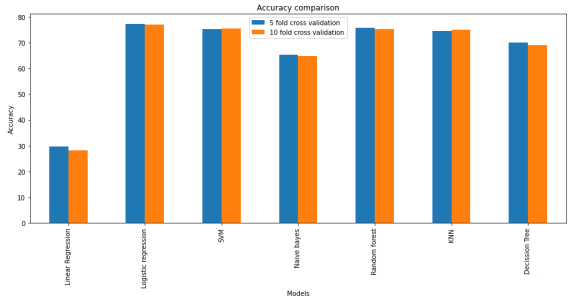
We have utilized k-fold cross-validation technique to evaluate the accuracy of our models. For the Pima Indian dataset, we conducted 5-fold and 10-fold cross-validation, as depicted in Figure 12 (a). This figure shows the comparison of accuracy obtained through cross-validation.

We conducted cross-validation on the diabetes dataset 2019 using 4-fold, 5-fold, and 10-fold methods to validate the accuracy of the models. Figure 12 illustrates the accuracy comparison we obtained from the cross-validation process.

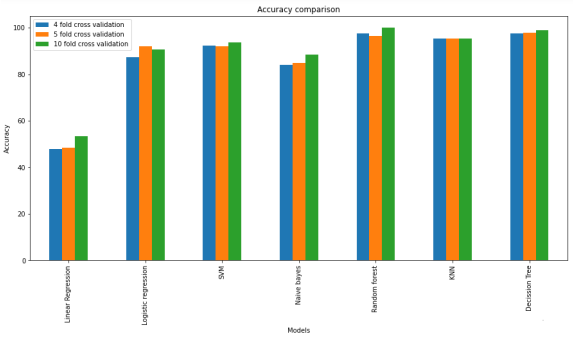
4.3.4 Result comparison of experiment 1 and Experiment 2

We compare the accuracy in the Table 3 of the diabetes dataset 2019 in different cases and highlight the best accuracy.

The accuracy results for the 2019 diabetes dataset show that linear regression, logistic regression, and SVM all perform reasonably well, with SVM having the highest accuracy of 93%. However, random forest and KNN perform much better, with an accuracy of 97.54%. This is true for both balanced and



(a) Pima Indian dataset



(b) 2019 Diabetics dataset

Fig. 12 Accuracy comparison of Cross validation.

imbalanced datasets. Naive Bayes and linear regression have relatively lower accuracy than other models on the balanced dataset. This is likely because these models are not as good at handling imbalanced datasets. Overall, random forest is the best model for predicting diabetes, followed by SVM.

In the table 4, we provided accuracy comparison of the Pima Indian dataset. In this study, six machine learning models were trained on the Indian Pima dataset using a 75%:25% training-testing data split ratio. The models were then evaluated on the testing data using the accuracy metric. Table 4 results showed that the Random Forest model achieved the highest accuracy of 79%, followed by Logistic Regression (78%), SVM (78%), KNN (77%), NB (69%), and DT (67%).

4.3.5 Balanced dataset effects on result

We have evaluated the performance of both balanced and imbalanced datasets using various performance metrics. The accuracy of both datasets is similar, but the confusion matrix shows a higher number of false negative reports for the imbalanced dataset, which significantly decreases in the balanced dataset. The random forest algorithm achieved 97% accuracy in the diabetes dataset 2019 and 80% in the Pima Indian dataset for both balanced and imbalanced

Table 4 Accuracy Results for Pima Indian dataset

Model Name	Training Data and Testing data	accuracy
Linear regression	75%:25%	77%
Logistic regresion	75%:25%	78%
SVM	75%:25%	78%
Random forest	75%:25%	79%
Naive bayes	75%:25%	69%
Decision tree	75%:25%	67%
KNN	75%:25%	77%

datasets. The comparison of confusion matrix for four classifiers such as Decision Tree, Random Forest, Logistic Regression and KNN between our model and an existing work is shown in Figure 13.For instance,

Comparison of False Negative rare for DT						
Number of test data: 76				**Number of test data: 298**		False Negative rate in %
Actual 0	40	9	Actual 0	263	11	**Nazin et al, 11.84**
Actual 1	9	18	Actual 1	9	15	
	predicted 0	predicted 1		predicted 0	predicted 1	**Proposal 3.02**
Nazin et al,			Proposal			

(a) Decision Tree

Comparison of False Negative rare for RF						
Number of test data: 76				**Number of test data: 256**		False Negative rate in %
Actual 0	44	5	Actual 0	202	3	**Nazin et al, 13.16**
Actual 1	10	17	Actual 1	4	47	
	predicted 0	predicted 1		predicted 0	predicted 1	**Proposal 1.56**
Nazin et al,			Proposal			

(b) Random Forest

Comparison of False Negative rare for LR						
Number of test data: 76				**Number of test data: 285**		False Negative rate in %
Actual 0	44	12	Actual 0	183	22	**Nazin et al, 6.58**
Actual 1	5	15	Actual 1	9	71	
	predicted 0	predicted 1		predicted 0	predicted 1	**Proposal 3.16**
Nazin et al,			Proposal			

(c) Logistic Regression

Comparison of False Negative rare for KNN						
Number of test data: 76				**Number of test data: 238**		False Negative rate in %
Actual 0	41	5	Actual 0	151	20	**Nazin et al, 15.07**
Actual 1	11	16	Actual 1	3	64	
	predicted 0	predicted 1		predicted 0	predicted 1	**Proposal 1.26**
Nazin et al,			Proposal			

(d) K-Nearest Neighbour

Fig. 13 Comparison of False Negative rate of the proposal for DT, RF, LR and KNN classifications model.

Figure 13(a)-(d) depicts a comparison of the confusion matrices for four machine learning algorithms (Decision Tree, Random Forest, Logistic Regression, and K-Nearest Neighbors) between our proposed model and an existing

Table 5 Comparison between the existing models and the proposed model

Reference	Method	Accuracy
Rahman et al. [8]	SVM, Logistic Regression, Naive Bayes, Random Forest model, KNN.	SVM showed higher accuracy: 87%.
Kavakiotis et al. [22]	LR, LDA, NB, SVM, ANN, KNN, RF.	SVM showed 98.6% using 10-fold cross validation.
Swapna et al. [2]	Recurrent neural network (RNN), Long short-term memory (LSTM), Convolutional neural network (CNN), Hybrid networks (CNN-LSTM), Support vector machine (SVM).	CNN 5-LSTM with SVM displayed the best result: 95.7%
Maniruzzaman et al. [9]	LR, Logistic Regression, Naive Bayes, Random Forest Model, KNN, Decision tree, Adaboost.	LR-based feature selection and RF-based classifier showed 94.25%
Ahmed et al. [7]	Decision tree (DT), Naive Bayes (NB), k-nearest neighbor (KNN), Random Forest (RF), Gradient Boosting (GB), Logistic Regression (LR) and SVM.	DT and RF both provided the highest accuracy respectively 96.81% and 96.81%
Tigga et al. [1]	Logistic Regression, SVM, KNN, Naïve bayes, Decision tree, Random Forest.	Accuracy of 94.1% and after 10-fold cross validation, the accuracy was 96.9%
Proposed model	Linear Regression, Logistic Regression, SVM, Random Forest, Naïve Bayes, Decision Tree, KNN.	The proposed model demonstrated accuracy at 97.54% using Random Forest algorithm over balanced data.

work. In comparison to Nazin et al.'s model [7] for the same dataset, our model demonstrated lower rates of false negatives for most classifiers. Specifically, for Decision Tree, our model produced only 3.02% false negatives, whereas Nazin et al.'s model yielded 11.84% as shown in Figure 13(a). Similarly, for Random Forest and Logistic Regression, our model also exhibited a lower rate of false negatives compared to the existing model. For K-Nearest Neighbors, our model achieved an impressively low rate of 1.26% false negatives, whereas Nazin et al.'s model had a much higher percentage of false negative cases (15.07%) as illustrated in Figure 13 (d). Additionally, we compare our work with other existing studies in Table 5, which indicates that our work achieves a higher level of accuracy than the majority of the studies included.

The advantages of the proposed machine learning-based diabetes detection model are: 1) SMOTE technique: The use of the SMOTE technique in the proposed approach helps to reduce the number of false negative cases and improve the model's performance on the imbalanced 2019 dataset. 2) Multiple algorithms: The proposed approach trained and tested several machine learning algorithms, including DT, RF, LR, and SVM, providing a comprehensive comparison of their performance on the given datasets. 3) Two datasets: The proposed approach used two datasets, PIMA Indian dataset, and 2019 diabetics dataset, to test the model's performance, which helps to evaluate the

model's generalizability. However, the proposed approach may also have some disadvantages, such as: 1) Small datasets: Both datasets used in the proposed approach are comparatively smaller, which may limit the model's generalizability to larger datasets. 2). Limited features: The proposed approach did not consider all possible features that may affect diabetes prediction. Therefore, there may be other factors that the model has not taken into account, limiting the model's accuracy.

5 Conclusion

Diabetes is a significant global medical issue that causes numerous fatalities every year. Early detection of diabetes intensity and risk factors can greatly reduce its impact. In this study, we developed a machine-learning technique for diabetes prognosis that predicts whether a patient has diabetes or not, utilizing data balancing and multiple machine-learning algorithms. We compared the performance of various classifiers using Pima Indian and diabetes 2019 datasets. Our findings show that the random forest classifier outperforms other classifiers in predicting diabetes for both balanced and unbalanced datasets. A balanced dataset helps to minimize false negative cases, making it a superior prediction model for gestational diabetes mellitus in healthcare. By using SMOTE technique, the number of minority samples in the dataset is increased which helps in avoiding bias towards a particular class in machine learning algorithms. This further leads to a reduction in the false negative rate. Based on the result analysis, it can be observed that logistic regression has a higher percentage of false negative cases (3.16%). However, it is still lower than the existing models. Therefore, the Random Forest algorithm is considered the best classification tool for this research, as it provides greater accuracy than other classification techniques. In the future, we aim to collect data from individuals and healthcare systems to develop a smart application that can predict diabetes with greater precision and efficiency.

Declarations

Conflict of interest

The authors have no conflicts of interest to declare that they are relevant to the content of this article.

Acknowledgement

This research was supported by Jagannath University Research Grant, Dhaka, Bangladesh (JnU/Research/Gapro/2022-2023/Science/18)

Ethics approval

Not applicable

Consent to participate

Not applicable

Consent to Publish

Not applicable

Availability of data and materials

The selected datasets are sourced from free and open-access sources such as Pima Indian Dataset: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database and Diabetes 2019 Dataset: https://www.kaggle.com/datasets/tigganeha4/diabetes-dataset-2019

Author Contribution

Md Ashraf Uddin, Md Manowarul Islam, Md. Al Amin Hossain: Conceptualization, Methodology, Software, Resources, Writing - Original Draft, Visualization; Md. Alamin Talukder: Methodology, Investigation, Formal analysis, Visualization, Writing- Reviewing and Editing; Arnisha Akher, Maisha Muntaha: Visualization; Sunil Aryal: Validation, Writing - Review & Editing.

References

[1] Tigga, N.P., Garg, S.: Prediction of type 2 diabetes using machine learning classification methods. Procedia Computer Science **167**, 706–716 (2020)

[2] Swapna, G., Vinayakumar, R., Soman, K.: Diabetes detection using deep learning algorithms. ICT express **4**(4), 243–246 (2018)

[3] Ramachandran, A.: Know the signs and symptoms of diabetes. The Indian journal of medical research **140**(5), 579 (2014)

[4] Talukder, M.A., Islam, M.M., Uddin, M.A., Akhter, A., Pramanik, M.A.J., Aryal, S., Almoyad, M.A.A., Hasan, K.F., Moni, M.A.: An efficient deep learning model to categorize brain tumor using reconstruction and fine-tuning. Expert Systems with Applications, 120534 (2023)

[5] Talukder, M.A., Islam, M.M., Uddin, M.A., Akhter, A., Hasan, K.F., Moni, M.A.: Machine learning-based lung and colon cancer detection using deep feature extraction and ensemble learning. Expert Systems with Applications **205**, 117695 (2022)

[6] Talukder, M.A., Hasan, K.F., Islam, M.M., Uddin, M.A., Akhter, A., Yousuf, M.A., Alharbi, F., Moni, M.A.: A dependable hybrid machine learning model for network intrusion detection. Journal of Information Security and Applications **72**, 103405 (2023)

[7] Ahmed, N., Ahammed, R., Islam, M.M., Uddin, M.A., Akhter, A., Talukder, M.A., Paul, B.K.: Machine learning based diabetes prediction and development of smart web application. International Journal of Cognitive Computing in Engineering **2**, 229–241 (2021)

[8] Rahman, M., Islam, L.: Diabetes recognition in pregnant women by extracting features using pca and data mining algorithms. In: 2019 IEEE Pune Section International Conference (PuneCon), pp. 1–6 (2019). IEEE

[9] Maniruzzaman, M., Rahman, M., Ahammed, B., Abedin, M., *et al.*: Classification and prediction of diabetes disease using machine learning paradigm. Health information science and systems **8**(1), 1–14 (2020)

[10] Olisah, C.C., Smith, L., Smith, M.: Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. Computer Methods and Programs in Biomedicine **220**, 106773 (2022)

[11] Hasan, M.K., Alam, M.A., Das, D., Hossain, E., Hasan, M.: Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access **8**, 76516–76531 (2020)

[12] Naz, H., Ahuja, S.: Deep learning approach for diabetes prediction using pima indian dataset. Journal of Diabetes & Metabolic Disorders **19**, 391–403 (2020)

[13] Krishnamoorthi, R., Joshi, S., Almarzouki, H.Z., Shukla, P.K., Rizwan, A., Kalpana, C., Tiwari, B., et al.: A novel diabetes healthcare disease prediction framework using machine learning techniques. Journal of Healthcare Engineering **2022** (2022)

[14] Yahyaoui, A., Jamil, A., Rasheed, J., Yesiltepe, M.: A decision support system for diabetes prediction using machine learning and deep learning techniques. In: 2019 1st International Informatics and Software Engineering Conference (UBMYK), pp. 1–4 (2019). IEEE

[15] Tapak, L., Mahjub, H., Hamidi, O., Poorolajal, J.: Real-data comparison of data mining methods in prediction of diabetes in iran. Healthcare informatics research **19**(3), 177–185 (2013)

[16] Joshi, T.N., Chawan, P.: Diabetes prediction using machine learning techniques. Ijera **8**(1), 9–13 (2018)

[17] Meng, X.-H., Huang, Y.-X., Rao, D.-P., Zhang, Q., Liu, Q.: Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. The Kaohsiung journal of medical sciences **29**(2), 93–99 (2013)

[18] Nai-arun, N., Moungmai, R.: Comparison of classifiers for the risk of

diabetes prediction. Procedia Computer Science **69**, 132–142 (2015)

[19] Choi, B.G., Rha, S.-W., Kim, S.W., Kang, J.H., Park, J.Y., Noh, Y.-K.: Machine learning for the prediction of new-onset diabetes mellitus during 5-year follow-up in non-diabetic patients with cardiovascular risks. Yonsei medical journal **60**(2), 191–199 (2019)

[20] Wu, Y.-T., Zhang, C.-J., Mol, B.W., Kawai, A., Li, C., Chen, L., Wang, Y., Sheng, J.-Z., Fan, J.-X., Shi, Y., *et al.*: Early prediction of gestational diabetes mellitus in the chinese population via advanced machine learning. The Journal of Clinical Endocrinology & Metabolism **106**(3), 1191–1205 (2021)

[21] Kaggle: Kaggle: Your Machine Learning and Data Science Community. https://www.kaggle.com. Accessed on March 22, 2023 (2000-2023)

[22] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., Chouvarda, I.: Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal **15**, 104–116 (2017)