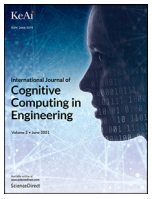
International Journal of Cognitive Computing in Engineering 2 (2021) 229–241

Contents lists available at ScienceDirect

International Journal of Cognitive Computing in Engineering

journal homepage: https://www.keaipublishing.com/en/journals/international-journal-of-cognitive-computing-in-engineering/



Machine learning based diabetes prediction and development of smart web application

Nazin Ahmed[a], Rayhan Ahammed[a], Md. Manowarul Islam[a,*], Md. Ashraf Uddin[a], Arnisha Akhter[a], Md. Al-Amin Talukder[a], Bikash Kumar Paul[b]

[a] Department of Computer Science and Engineering, Jagannath University, Dhaka, Bangladesh
[b] Department of Information and Communication Technology, Mawlana Bhashani Science and Technology University, Bangladesh

ARTICLE INFO

Keywords:
Diabetes prediction
Machine learning
Flask
Accuracy
Random Forest (RF)
Support Vector Machines (SVM)
Logistic regression (LR)
Gradient boosting (GB)
k-nearest neighbor (k-NN)

ABSTRACT

Diabetes is a very common disease affecting individuals worldwide. Diabetes increases the risk of long-term complications including heart disease, and kidney failure among others. People might live longer and lead healthier lives if this disease is detected early. Different supervised machine learning models trained with appropriate datasets can aid in diagnosing the diabetes at the primary stage. The goal of this work is to find effective machine-learning-based classifier models for detecting diabetes in individuals utilizing clinical data. The machine learning algorithms to be trained with several datasets in this article include Decision tree (DT), Naive Bayes (NB), k-nearest neighbor (KNN), Random Forest (RF), Gradient Boosting (GB), Logistic Regression (LR) and Support Vector Machine (SVM). We have applied efficient pre-processing techniques including label-encoding and normalization that improve the accuracy of the models. Further, using various feature selection approaches, we have identified and prioritized a number of risk factors. Extensive experiments have been conducted to analyze the performance of the model using two different datasets. Our model is compared with some recent study and the results show that the proposed model can provide better accuracy of 2.71% to 13.13% depending on the dataset and the adopted ML algorithm. Finally, a machine learning algorithm showing the highest accuracy is selected for further development. We integrate this model in a web application using python flask web development framework. The results of this study suggest that an appropriate preprocessing pipeline on clinical data and applying ML-based classification may predict diabetes accurately and efficiently.

1. Introduction

The disease “Diabetes Mellitus” is one of the most common critical diseases in the world. According to the World Health Organization (WHO), diabetes affects 8.5% of adults over the age of 18 and is responsible for 1.6 million deaths worldwide (World Health Organization, 2021). Although the rate of diabetes-related premature death in many developing countries fell from 2000 to 2010, the statistics again increased between 2010 and 2016. The four primary diseases, namely cardiovascular diseases, cancer, chronic respiratory diseases, and diabetes, kill over 18% of people worldwide and have become a serious public health concern. For example, in 2000, deaths from diabetes climbed by 70%, and in 2020, mortality among males are expected to grow to 80%. Diabetes mellitus can result from obesity, age, lack of exercise, lifestyle, hereditary diabetes, high blood pressure, poor diet, etc. Over time, people with diabetes have a high risk of diseases such as heart disease, stroke, kidney failure, nerve damage, eye issues, etc.

The present clinical practice consists in collecting the data necessary to detect diabetes through a number of tests and then providing an appropriate diagnostic drug (Gadekallu et al., 2020). In the healthcare sector (Haq et al., 2020; Yu et al., 2010), supervised and non-supervised machine learning (ML) approaches are utilized to diagnose various kinds of diseases. ML methods enable researchers to investigate hidden patterns of medical datasets for predicting expected outcomes and reducing costs of identifying complex diseases (Ahamed et al., 2021). ML algorithms are trained with real medical datasets having different features and external variables.

If diabetes can be discovered at the primary stage, harmful effects can be avoided with adequate medical care (Dinh et al., 2019). ML approaches can aid in early detection of this disease. Machine learning algorithms are adopted to create a prediction model since ML methods allow computers to learn and gain intelligence from previous experience or a pre-defined dataset (Gulshan et al., 2016; Vinayakumar et al., 2019). The predictive model can identify and understand the incoming data, allowing it to make more precise decisions.

* Corresponding author.
E-mail address: manowar@cse.jnu.ac.bd (Md.M. Islam).

https://doi.org/10.1016/j.ijcce.2021.12.001
Received 28 March 2021; Received in revised form 4 December 2021; Accepted 6 December 2021
Available online 7 December 2021

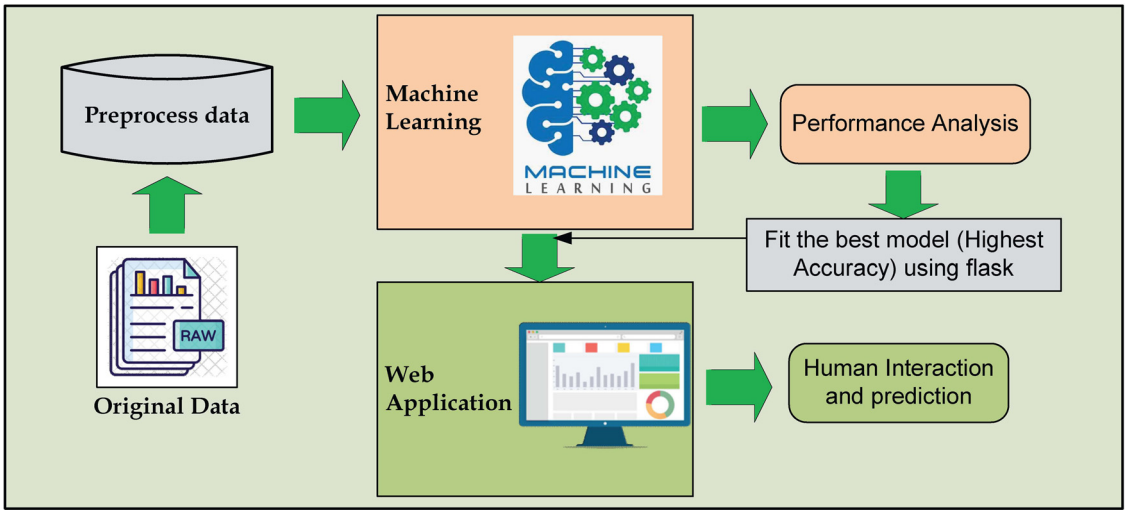


Fig. 1. Overview of the proposal.

Researchers such as Albahli (2020), Dinh et al. (2019), Kaur & Kumari, 2020, Kopitar et al. (2020), Maniruzzaman et al. (2020) and Vinayakumar et al. (2019) designed and proposed different kinds of machine learning based models to diagnose diabetes disease. However, only a few studies have concentrated on integrating the trained model into a user's app and designing a user interface so that consumers may monitor their health status on their smartphones. Furthermore, those models were trained using only one or two datasets, which does not guarantee that the model would perform as expected in real-world scenarios. We have focused on addressing these gaps in our proposed machine learning model.

This work aims to apply machine learning algorithms to present the analytical results regarding physical components and circumstances that contribute to the development of diabetes in human body. The pre-processing on the datasets provides better accuracy of the model than the existing research works. The correlation based feature selection method discovers necessary attributes in the dataset. The classifier showing better performance for all the datasets is integrated in a web application. Considering the current research gap, we set the following research questions. *How does the accuracy of machine learning algorithms vary in predicting diabetes for different datasets including PIMA Indian dataset? What factors influence the most in detecting diabetics? How can an effective machine learning algorithm be discovered to integrate that in a web application?*

To predict diabetes of individuals, we have adopted several machine learning algorithms and evaluated their performances. We have examined the performance of seven different models, namely Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), Support Vector Machines (SVM), Logistic Regression (LR), Gradient Boosting (GB) and k-nearest neighbor (k-NN) algorithm to develop a predictive model. Two datasets with various attributes including glucose level, insulin level, blood pressure, BMI and age are used for training the ML algorithms. Results of the algorithms are evaluated based on several performance metrics. Finally, based on the accuracy level, a web application is developed to predict the diabetes of any individual. Any users can get the prediction of diabetes by using this application on their smartphones or computers.

Fig. 1 shows each phase of the proposed ML based diabetes prediction model. In the first phase, every dataset is pre-processed. In the second stage, the pre-processed datasets are feed into the different machine learning algorithms. In the third phase, the output of the models is then analyzed using various metrics. In the later phase, the model that provides the highest accuracy is adopted to detect diabetes of any individual and integrated with a web-based application. This web-based application is developed using Flask of python programming language.

In a nutshell, the contributions of this research are listed as follows:

- Our first contribution is to train several machine learning algorithms using four different clinical datasets for detecting diabetes. All the datasets are pre-processed by applying different pre-processing techniques.
- Secondly, the performances of each ML algorithms with four datasets are analyzed with respect to several parameters like precision, Recall, f1-score, ROC curve and accuracy. Further, we have identified several important features or attributes using different feature selection methods such as correlation, chi-square, etc. The feature selection methods find out the mostly correlated features to diabetes disease. The performances of the ML algorithms were also analyzed on the reduced set of attributes.
- Thirdly, based on the performance results, web-based application is developed to predict the diabetes of individuals.

The rest of the paper is organized as follows. In Section 2, we review the related literature for diabetes detection. In Section 3 and 4, we describe our proposal along with various machine learning algorithms adopted in the predictive learning. Experimental results are presented in Section 5. Then in Section 6, we present the implementation details of the web-based application and Finally in Section 7, we conclude the paper with the future enhancement of this study.

2. Literature survey

In this section, some closely related works are discussed briefly. In most of the reseach works, Pima Indians Diabetes Dataset (PIDD) have been used by many researchers for diabetes prediction. Various supervised machine learning algorithms were used to predict diabetes (Kaur & Kumari, 2020). Radial basis function (RBF) kernel SVM, artificial neural network (ANN), multifactor dimensionality reduction (MDR), linear SVM and k-NN are some of them to mention. Based on p value and odds ratio (OR), Logistic Regression (LR) has been used to recognize the risk factors for diabetes (Maniruzzaman et al., 2020). Four classifiers have been adopted to predict diabetic patients, such as NB, DT, Adaboost, and RF. Partition protocols like- K2, K5, and K10 were also adopted, repeating these protocols into 20 trails. For the performance measurement of the classifiers, accuracy (ACC) and area under the curve (AUC) were analyzed.

Kopitar et al. (2020) showed a comparison of widely utilized regression models such as Glmnet, RF, XGBoost, LightGBM for predicting type 2 diabetes mellitus. The goal of this work was to examine if innovative machine learning methodologies gave any advantages in early predic-

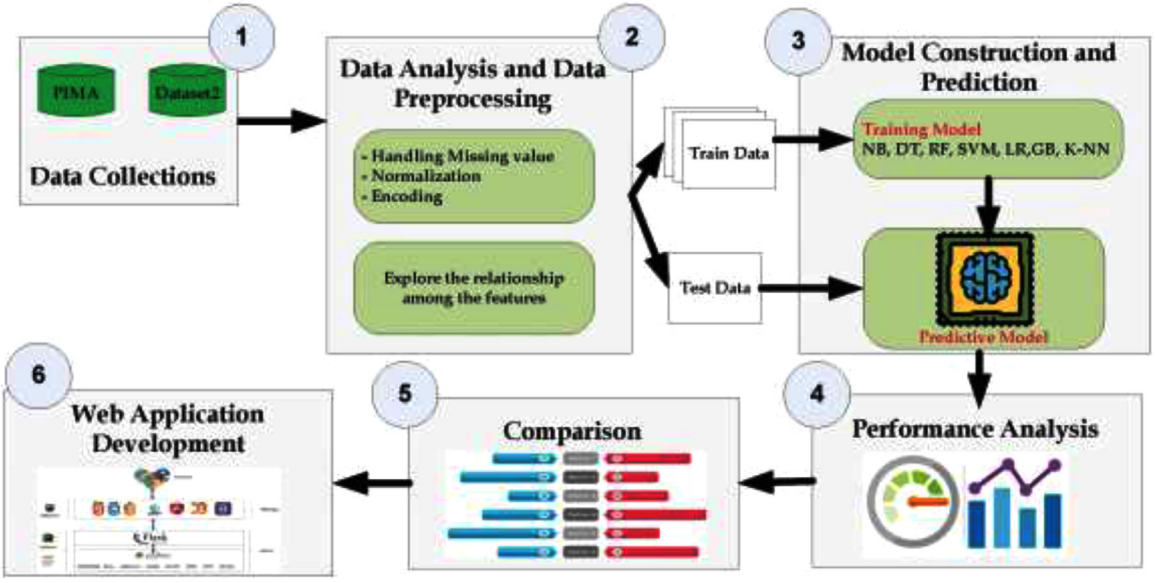


Fig. 2. Work-flow diagram of the proposal.

tion of impaired fast glucose and fasting plasma glucose (FPGL) levels compared to classic regression techniques.

For the prediction of diabetic patients, Maniruzzaman et al. (2020) have chosen four classifications such as naive bays (NB), decision tree (DT), adaboost and random forest. These methods were also implemented by three types of partition protocols (K2, K5, and K10). These classifers' performances are measured with precision (ACC) and curve surface (AUC).

A hybrid model to detect type 2 diabetes was suggested by Albahli (2020). In order to extract unknown, hidden property from the dataset and to obtain more exact results, we use K-mean clustering, which is followed by the execution of a Random Forest and XGBoost classifier.

Yahyaoui et al. (2019) suggested a Machine Learning Techniques (ML) DSS for anticipating diabetes. They compared traditional machine learning with approaches to the deep learning. The authors applied the classifiers most typically used for a standard machine learning method: SVM and the Random Forest (RF). In contrast, they used a full-scale neural network (CNN) for Deep Learning (DL) to forecast and identify patients who suffer from diabetes.

Zou et al. (2018) predicted diabetes using the decision tree, random forests, and neural network. The dataset is collected from the Luzhou physical exams in China. The PCA was applied to reduce the dimension of the dataset. They selected several ML approaches to execute independent test to verify the universal applicability of method.

Supervised machine learning models which explore data-driven approaches were used to identify patients with diabetes diseases (Dinh et al., 2019). A complete research was conducted based on the National Health and Nutrition Examination Survey (NHANES) dataset. To develop models for cardiovascular, prediabetes, and diabetes detection, they have used all available feature variables within the data. Using various time frames and set of features within the data, different machine learning models, namely Support Vector Machines, logistic regression, gradient boosting and random forest were evaluated for the classification.

In Choubey et al. (2017) the authors used NBs for the classification on all the attributes. Afterwards GA was used as an attribute selector and NBs used the selected attributes for classification. The experimental results show the performance of this work on PIDD and provide better classification for diagnosis. Three specific supervised machine learning methods are used by Joshi and Chawan (2018), namely SVM, Logistic regression and ANN. His goal for research was to predict diabetes patients and he has also proposed an effective model for the prior detection of diabetes disease. Rajeswari and Prabhu (2019) focused on machine learning classification algorithms for predicting diabetes disease with more accuracy. Their study in SVM classification algorithm achieved highest accuracy. Various measures have been used to calculate the performance of classification algorithms.

An intelligent model using machine learning practices is developed (Nilashi et al., 2017) to identify diabetes disease. This model is constructed using approaches like clustering, removal of noise and classification, each of which made use of SOM, PCA and NN, respectively. The adaboost and bagging ensemble techniques are used to detect diabetes (Perveen et al., 2016). Along with standalone data mining technique, a base learner is used to identify patients with diabetes mellitus, namely J48 (c4.5) decision tree that makes use of multiple diabetes risk factors. In the Canadian Primary Care Sentinel Surveillance Network, three different ordinal adult groups are selected for classification. Experimental result shows that, the adaboost ensemble method shows better performance than both bagging and standalone J48 decision tree. For diagnosing T2DM, Kazerouni et al. (2020) has taken in consideration four different classification models, namely SVM, K-NN, ANN and LR. A comparison is done among these algorithms to measure the diagnostic power of this algorithms. The algorithms are performed on six LncRNA variables and demographic data.

3. Methodology

Fig. 2 depicts the proposed framework for diabetes prediction. Firstly, we pre-process two separate datasets. In the pre-processing stage, correlation between attributes of the datasets is analyzed for finding useful features in detecting diabetes. After that, the data is divided into two sets: training and testing. The training set is utilized to develop predictive ML models using a variety of machine learning algorithms. Next, we assess the proposal's performance with respect to different metrics. Finally, the best ML model is deployed in a web application using flask. Following this, we describe the workflow of each part briefly:

1 *Data Collection:* We collected two alternative datasets, each with a different number of factors or features, to ensure the model's robustness. The datasets were compiled from a wide variety of sources, including diabetes statistics and health characteristics obtained from people around the world and from various health institutes.

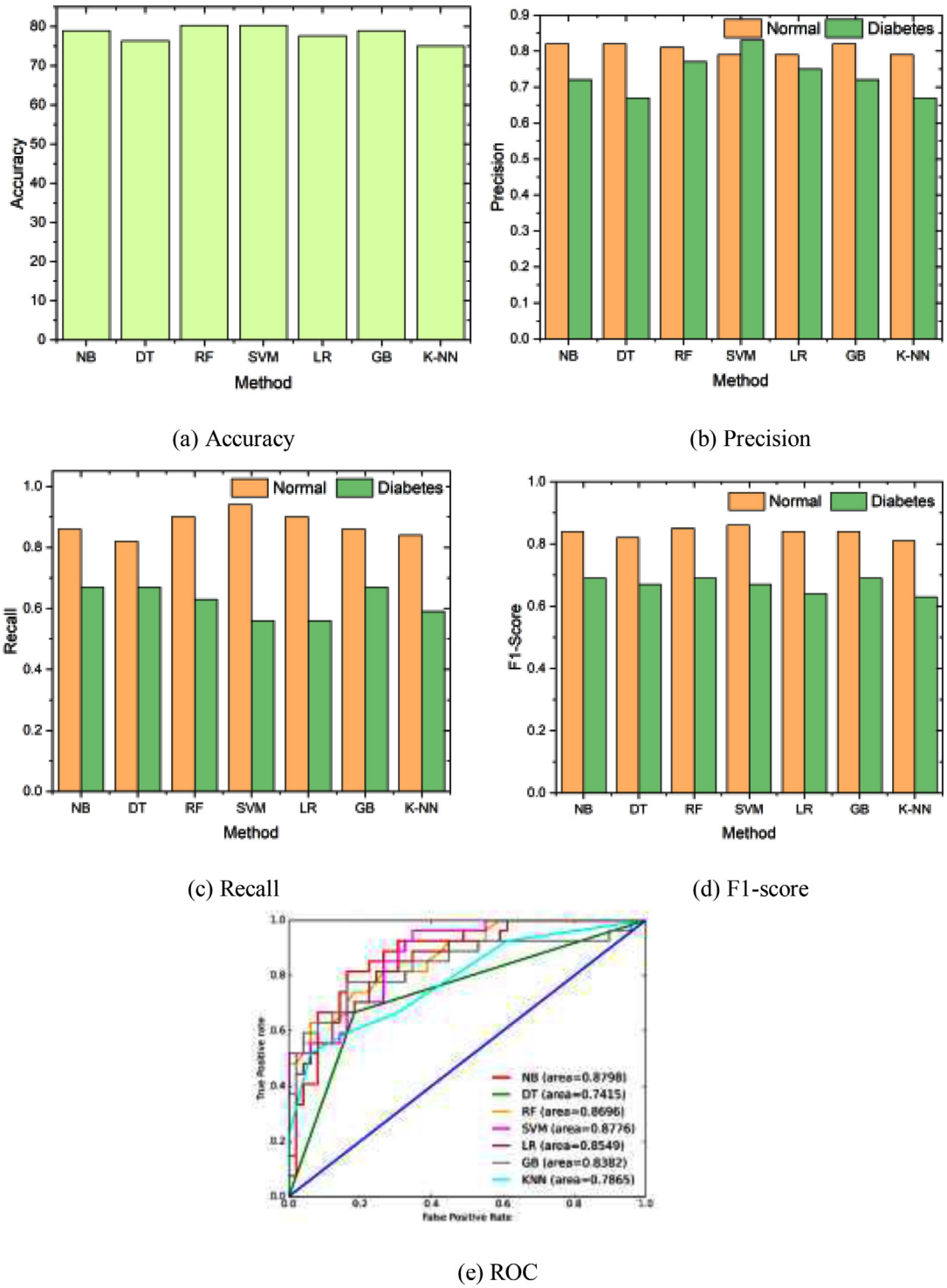


Fig. 3. The performance results of dataset- 1.

2 *Data Analysis and Data Preprocessing:* Several pre-processing techniques are applied on the datasets before feeding these datasets into the machine learning model so that the performance of the model is improved. The pre-processing tasks include removing outliers and dealing with missing values, data standardization, encoding, and so on.

- *Outliers Removal* - Attributes' values that are beyond acceptable boundaries and have high variation from the rest of the respective attribute's value might be present in the dataset. Such attributes' value might degrade the machine learning algorithm's performance. To eliminate such outliers, we applied the IQR (Inter-quartile Range) approach.
- *Missing value Handling* - To improve model performance, the mean value of each attribute was employed for handling the missing values.
- *Label Encoding* - Label encoding is the process of converting the labels of text/categorical values into a numerical format that ML algorithms can interpret. For example, the categorical values of

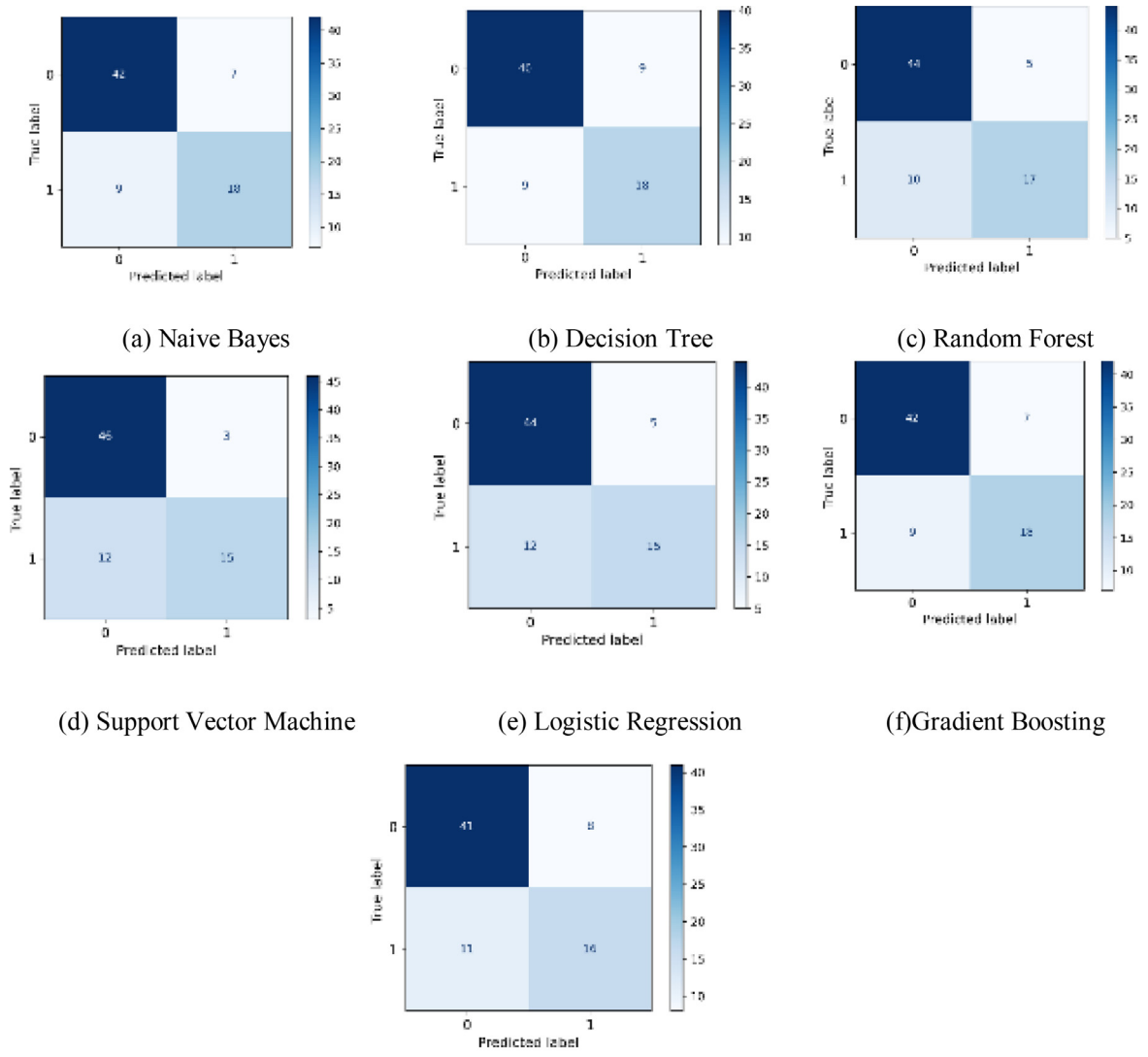


Fig. 4. Confusion matrix of dataset-1.

Junkfood consumption status yes to '1' and No to '0' have been converted.

3 *Model Construction and Prediction:* To construct the predictive model, 80% of the pre-processed data has been used for training while the remaining 20% data is used for the testing purpose.

4 *Performance Analysis:* We have analyzed the results of the proposed model in terms of several performance metrics. The algorithm that provides highest prediction accuracy is selected as the best algorithm for the web application development.

5 *Performance Comparison:* In this step, the accuracy of the proposal has been compared with some recent works related to diabetes prediction. The performance results indicate that the proposal can improve the performance compared to the recent related research.

6 *Web Application development:* To develop a smart web application, we have used the Flask micro-framework and integrated the best model. To predict diabetes, a user is required to submit a form with necessary numbers of diabetes related parameters. The application uploaded in a server predicts the results using the adopted machine learning model. We describe the adopted machine learning algorithms in the following sections.

4. Adopted machine learning algorithms

In this section, we will describe various machine learning algorithms that are used in the predictive model.

- *Naive Bayes* - The Naive Bayes classification model is designed based on Bayes' Theorem (Gandhi, 2018). The assumption of independence among predictors is considered in this model. Naive Bayes can classify a given problem instance using a conditional probability model. The probability of an entity having $x=(x_1,x_2,x_3,\ldots x_n)$, n number of features (independent variables) is calculated as $P(C_k|x_1, x_2, \ldots x_n)$ for each of K possible outcomes or classes C_k. The conditional probability can be represented as follows:

$$p(C_k|x) = \frac{p(C_k) \times p(x|C_k)}{p(k)} \quad (1)$$

Here, $p(C_k)$represents the prior probability of class C, $p(k)$ is the prior probability of predictor and $p(x|Ck)$ is the likelihood which is the probability of predictor given class.

Assuming each feature x_i as conditionally independent of all other features x_j and considering any category C_k, $j \neq i$, this model can be

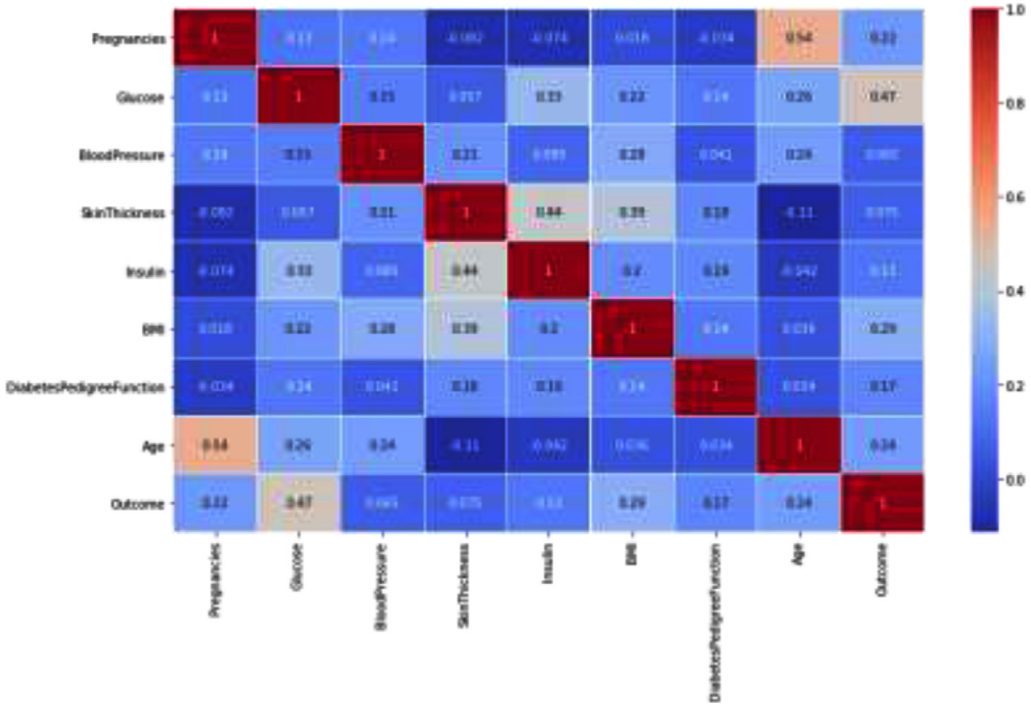


Fig. 5. Correlation matrix for correlation analysis.

represented as follows:

$$p(C_k|x_1, x_2, x_3, ...x_n) = p(C_k)\prod_{i=1}^{n} p(x_i|C_k) \tag{2}$$

- *Decision Tree* - For classification and prediction (Gupta, 2019), Decision Tree is one of the most robust and prevalent algorithms. Internal nodes in a decision tree architecture represent tests on an attribute. The outcome of the respective tests is represented by the leaf nodes (terminal node) bearing respective class labels.
- *Random Forest Classifier* - Another classification model used in modeling forecasts and analyzing behavioral characteristics is random forest (Yiu, 2019). The random forest algorithm is composed mostly of many decision trees, each of which represents a distinct instance. The instances assist in the classification of the data input into the random forest. The random forest technique evaluates each instance independently and returns the most voted prediction.
- *Support Vector Machine* - Support Vector Machines are used to handle classification and regression problems (SVM) (Ray, 2017). The decision boundary that the SVM returns is defined by the following equation:

$$f(X) = w^b + T \tag{3}$$

Here, w is the weight vector, X is the data dataset to be classified, and b is the linear coefficient.

- *Logistic Regression* - In logistic regression, the probability determines whether a given data entry falls into the category denoted by the number "1"(Brownlee, 2016c). The sigmoid function in logistic regression is used to model the data as follows:

$$P(X) = \frac{1}{1 + e^{-y}} \tag{4}$$

Here, e is the base of the natural logarithms, y is the actual numerical value and $P(X)$ is the probability that can be any value between 0 and 1.

- *Gradient Boosting* - Gradient boosting is another machine learning strategy for handling regression and classification problems. This prediction model is created by combining several weak prediction models, which are generally decision trees. The model is created in a stage-by-stage way, similar to how other boosting approaches operate. Next, the model is simplified using an arbitrary differentiable loss function optimization approach (Brownlee, 2016a).
- *K-Nearest Neighbor*- K-NN is one of the mostly used basic Machine Learning algorithms based on the Supervised Learning methodology. K-NN is used for both Regression and mostly classification (Brownlee, 2016b). The K-NN algorithm considers the similarity between the new case/data and the existing case/data. The new case is then assigned to the category that is the most similar among the available options.

5. Experimental results analysis

Our proposed model is tested and evaluated in this section using a variety of machine learning algorithms, including NB, DT, RF, SVM, LR, GB, and K-NN. To find the effectiveness we have used 4 different datasets and each of them contains different types and number of attributes.

5.1. Experimental setup

The proposed model is built in Python and executes on a computer having an Intel Core i7 processor with a 4 GB graphics card, 16GB RAM and a 64-bit Windows operating system running at 1.80 GHz. To test the efficiency of our model, we have used a 10-fold cross validation process. The dataset is shuffled and divided into 10 segments at random, with one segment serving as the test set and the others serving as the training set in turn. The average of the results from multiple experiments is considered as the final output of the experiment.

5.2. Performance metrics

The performance of the proposed approach has measured using confusion matrix shown in Table 1. The confusion matrix has four different outcomes: true positive(TP), true negative(TN), false positive(FP), and false negative(FN), as follows:

Next, we consider the following metrics to analysis the suggested model.

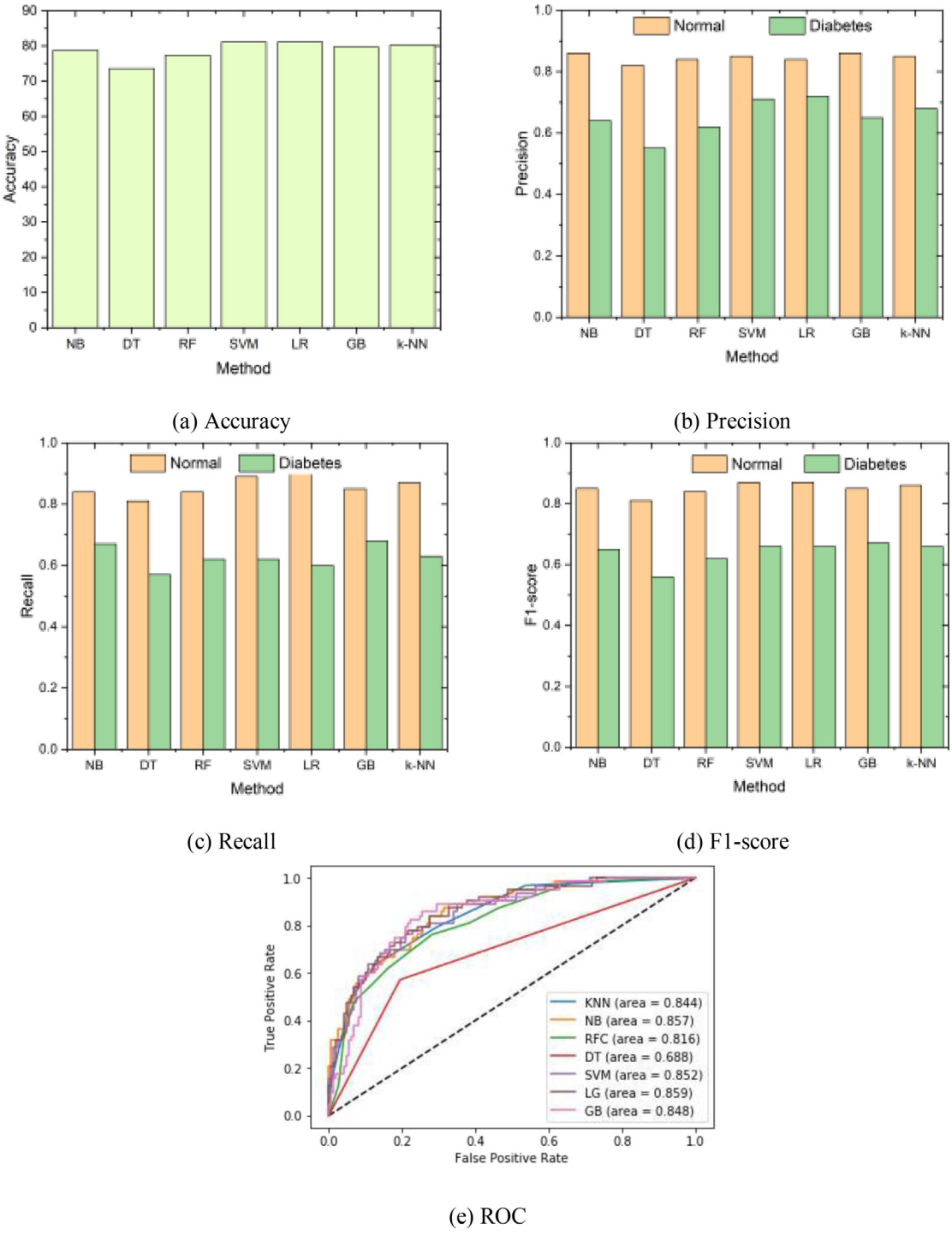


Fig. 6. The performance results with 5 variables.

Table 1
Confusion matrix.

Predicted Results	Actual Positive	Actual Negative
Yes	TP	FP
No	FN	TN

- *Accuracy:* It measures the model's total number of accurate predictions and can be measured as a ratio between the number of correct prediction and total number of test cases of any model as follows:

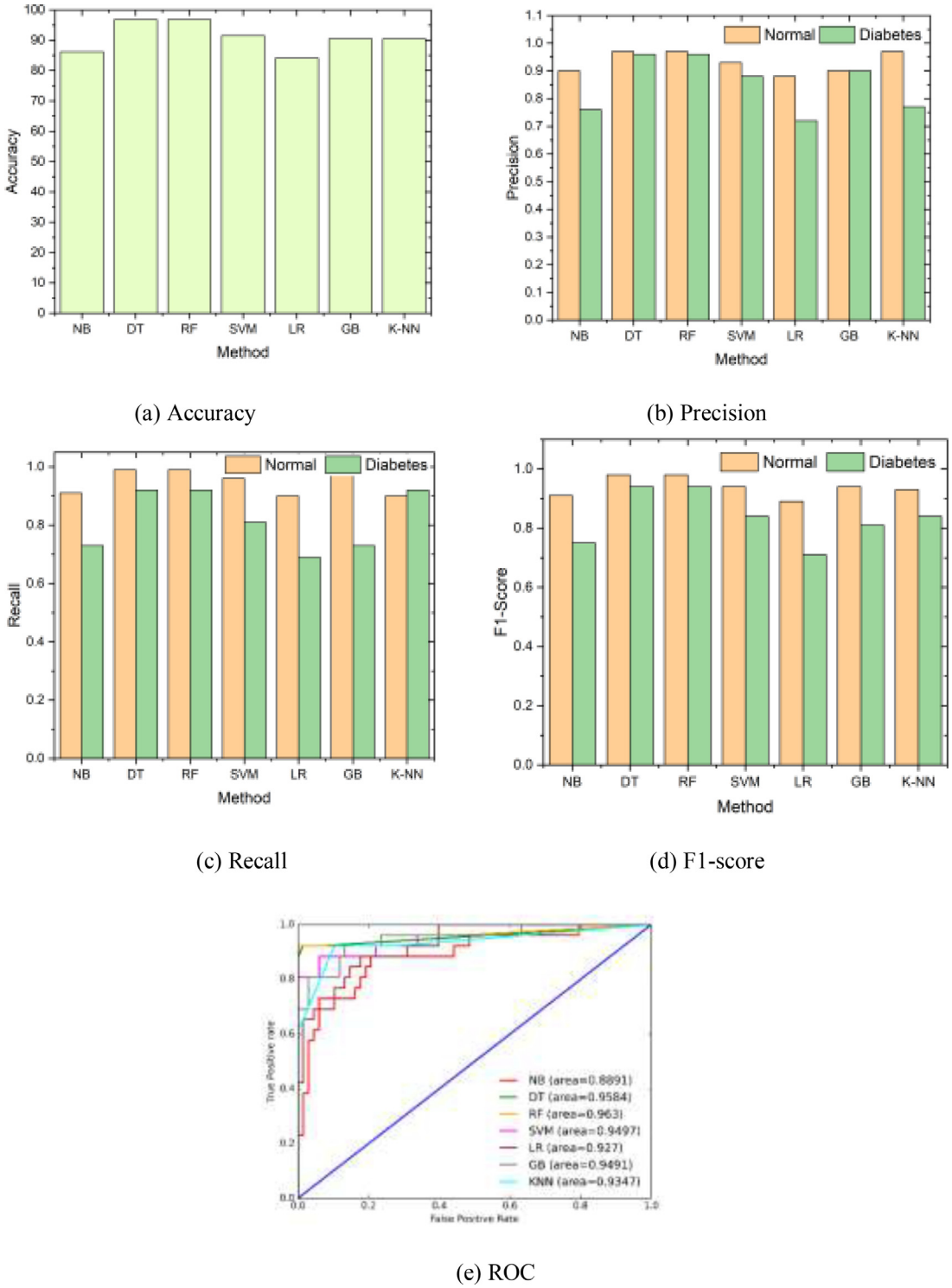
$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{5}$$

- *Precision:* The proportion of correct positive predictions to total positive predictions is known as precision:

$$\text{Pr}\,ecison = \frac{TP}{TP + TN} \tag{6}$$

- *Recall:* Total positive predictions vs. actual positive values is known as recall:

$$\text{Re}call = \frac{TP}{TP + FN} \tag{7}$$



(a) Accuracy

(b) Precision

(c) Recall

(d) F1-score

(e) ROC

Fig. 7. The performance results of dataset-2.

- *F1-score:* F1-score takes precision and recall into account and can be described as follows:

$$F1\_score = 2 * \left(\frac{precion * recall}{precion + recall} \right) \tag{8}$$

5.3. Results of dataset −1

Pima Indian Diabetes Dataset is one of the ideal datasets for evaluating machine learning algorithms for predicting diabetes (UCI Machine Learning Repository, 1998). The National Institute of Diabetes and Digestive and Kidney Diseases provided the Pima Indian dataset to determine if a patient has diabetes based on diagnostic measures like Pregnancies, Glucose level, Blood Pressure, Skin Thickness, Diabetes Pedigree Function, Insulin, BMI and Age (Table 2).

Before training the proposed model, missing values stated in Table 3 were filled using the mean statistical method. Also, outliers are important to be removed if distance-based algorithm is used like logistic re-

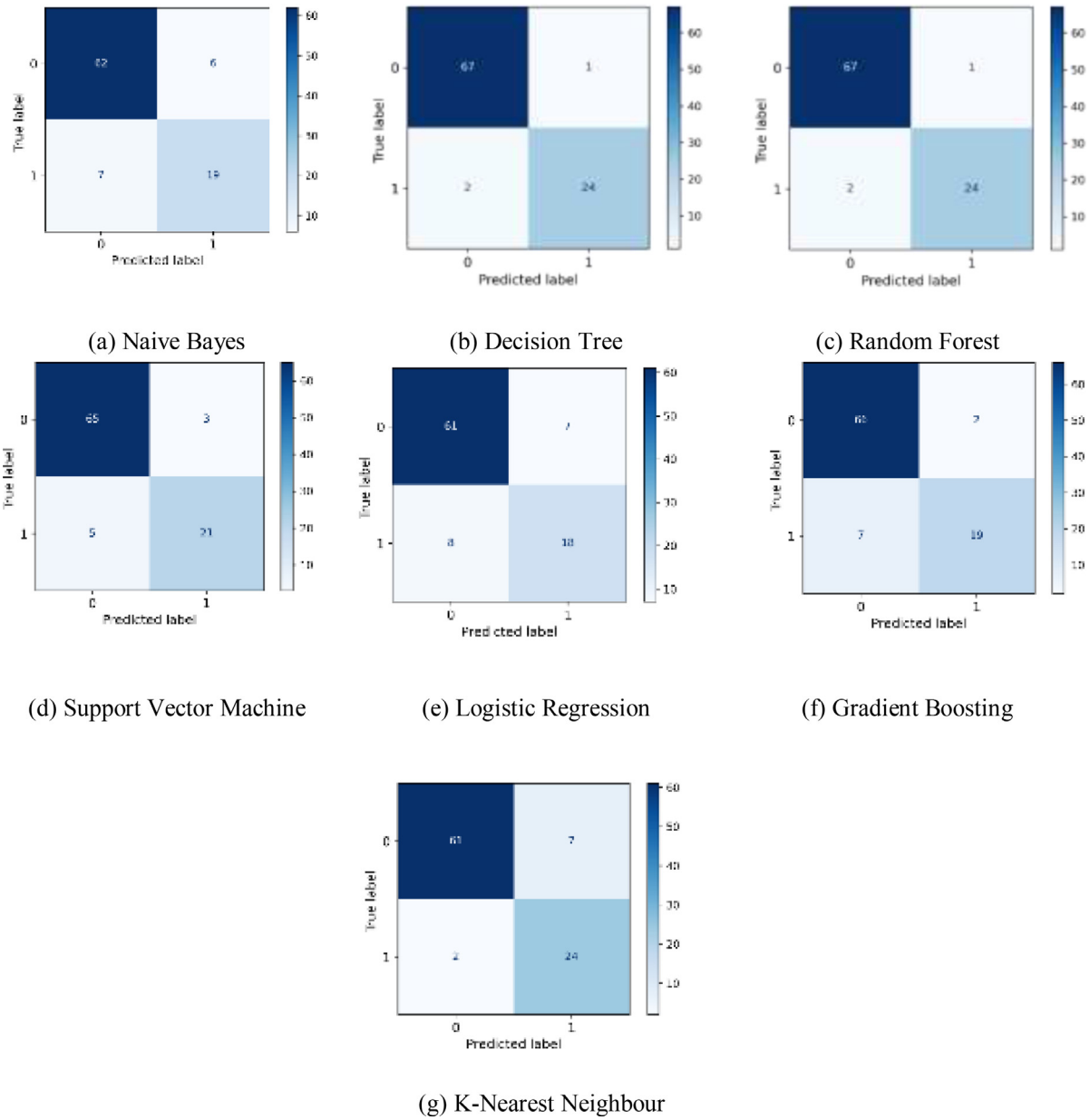


Fig. 8. Confusion matrix of dataset-2.

Table 2
Features of dataset-1.

Pima Indian Dataset		
Number of Records		768
Number of Attributes		9
01	Pregnancies:	Number of occurrences of pregnancy
02	Glucose:	In a glucose tolerance measure, the plasma glucose concentration after 2 h
03	Blood Pressure	The number of times the heart beats per minute is called diastolic blood pressure (mm Hg)
04	Skin Thickness	The thickness of the skin folds on the triceps (mm)
05	Insulin:	serum insulin (mu U/ml) after 2 h
06	BMI	Body mass index
07	Diabetes Pedigree Function	Diabetes Pedigree Function
08	Age	Age of the person in years
09	Outcome	Class variable as a result (0 or 1)

Table 3
Features of dataset-1.

Feature	Number of Missing value
Glucose	5
Blood Pressure	35
Skin Thickness	227
Insulin	374
BMI	11

gression, SVM, etc. For better accuracy, we have removed the outliers using IQR (interquartile range) method.

Fig. 3 represents the overall results of the experiment in terms of accuracy, precision, recall and f1-score. The accuracy of the various models is 78.95%, 76.32%, 80.25%, 80.26%, 77.63%, 78.95% and 75% NB, DT, RF, SVM, LR, GB, KNN, respectively, and SVM outperforms the other methods. Graph 3(e) shows the ROC curve (receiver operating characteristic curve) for various model and we found SVM provide better performance. The efficient preprocessing pipeline- outliers removal, missing value handling, and label encoding result in unexpected model accuracy gains, ensuring the effectiveness of the proposal. As a result, any machine learning algorithms applied in this research work outperformed the existing state-of-the-artwork discussed in Section 5.5.

Fig. 4 depicts the confusion matrix for all the adopted machine learning algorithms.

In next experiment, we find most influential attributes using correlation matrix that states how the features are related to each other on the target variable. Fig. 5 shows the correlation matrix between each of the attributes to the class variable. The relationship between the parameters is depicted in the correlation plot.

- The most associated parameters with the Outcome are glucose, age, BMI, and pregnancies.
- Insulin and Diabetes Pedigree Function have no bearing on the final result.
- There is only a slight connection between blood pressure and skin thickness and the outcome.
- Age and Pregnancy, Insulin and Skin Thickness, BMI and Skin Thickness, Insulin and Glucose all have a small association.

Then 5 features namely Pregnancies, Glucose, BMI, Diabetes Pedigree Function and Age have been selected as most influential parameters. Fig. 6 shows the overall results for the selected 5 variables. It is clear that SVM provides better accuracy of 81.13% than other algorithms.

5.4. Results of dataset –2

Tigga and Garg (2020) conducted a survey and collected a dataset contains 950 records and 19 attributes that have measurable influence on diabetes such as Family Diabetes history, Blood Pressure, Exercise, BMI, Smoking level, Alcohol consumption, Sleeping hours, Food habits, Pregnancy, Urination frequency, Stress level and so on (Table 4)..

The entire outcomes of the experiment in terms of accuracy, precision, recall, and f1-score are presented in Fig. 7. For NB, DT, RF, SVM, LR, GB, and KNN, the accuracy of these models is 86.17%, 96.81%, 96.81%, 91.49%, 84.04%, 90.43%, and 90.43%, respectively. This table illustrates that DT and RF both provide the highest level of accuracy and exceed the other approaches. The ROC curve (receiver operating characteristic curve) for various models is shown in graph 7(e), and this graph demonstrates that RF offers higher performance. The confusion matrix for all of the adopted machine learning techniques is shown in Fig. 8. In any ML algorithms, the data preparation step improves the data quality and facilitates the retrieval of relevant insights from the dataset. Thus, all the classifier algorithms can detect the diabetes patients more accurately with the normalized and preprocessed dataset.

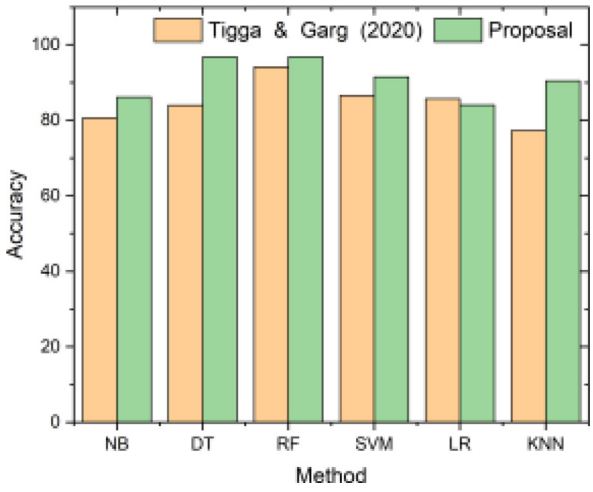


Fig. 9. Performance comparison with Tigga & Garg, 2020 for dataset-2.

Table 4
Features of dataset-2.

Diabetes dataset 2		
Number of Records		950
Number of attributes		19
01	Age	Age in Year
02	Gender	Gender of the participant
03	Family Diabetes	Family history with diabetes
04	highBP	Diagnosed with high blood pressure
05	PhysicallyActive	Walk/run/physically active
06	BMI	Body Mass Index
07	Smoking	Smoking
08	Alcohol	Alcohol consumption
09	Sleep	Hours of sleep
10	SoundSleep	Hours of sound sleep
11	RegularMedicine	Regular intake of medicine?
12	JunkFood	Junk food consumption
13	Stress	Not at all, Sometimes, Often, Always
14	BPLevel	Blood pressure level
15	Pregancies	Number of pregnancies
16	Pdiabetes	Gestation diabetes
17	UriationFreq	Frequency of urination
18	Diabetic	Yes or No

5.5. Comparison with other researches

The authors (Tigga & Garg, 2020) have surveyed to collect the dataset-3 and conducted their research for the detection of Diabetes disease. They have used several attributes and six machine learning algorithms for the detection process. In our paper, we have used their dataset and done preprocessing like- level encoding and normalization to improve the quality of the data. After that, machine learning models are implemented and we have found better accuracy than them. Fig. 9 shows the performance comparison on accuracy of dataset-2.. Except for LR, all other methods show better accuracy; 5.57%, 12.81%, 2.71%, 4.99%, 13.13% more for NB, DT, RF, SVM and KNN, respectively (Table 5).

Pima Indian Diabetes Dataset has several features and contains a huge number of records to be used for diabetes detection. Missing values are filled using Mean values of the respective feature and outliers are removed using IQR method from the dataset. ML models are applied on the dataset after preprocessing the dataset and that yields better accuracy than previous works. Fig. 10 shows the performance comparison on accuracy of the dataset-1 for Pranto et al. (2020) and our proposed method. It shows 6.75%, 3.22%, 2.36% better accuracy in NB, DT, RF and KNN, respectively, for our proposed model than Pranto et al. (2020).

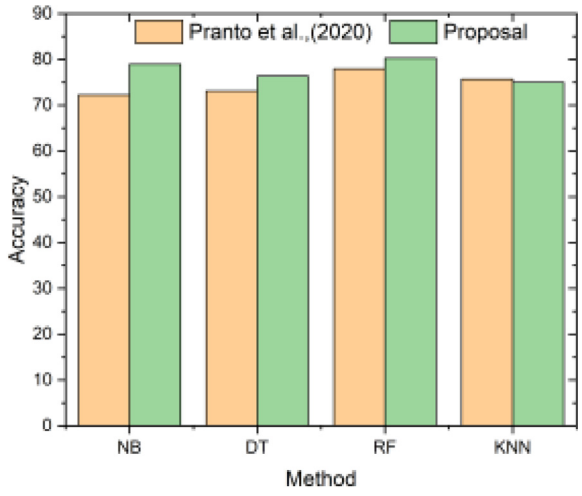


Fig. 10. Performance comparison with (Pranto et al., 2020) for dataset-1.

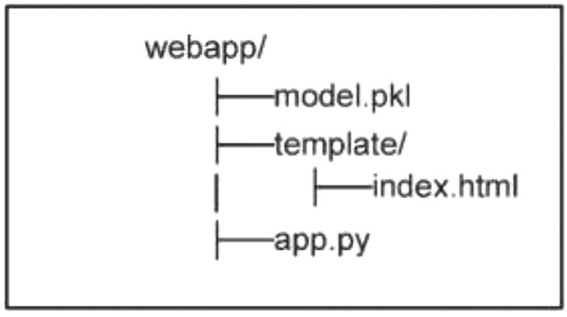


Fig. 11. File structure of the web application.

Table 5
Accuracy performance comparison with other works.

Method Name	Accuracy in% for Dataset-2		Accuracy in% for Dataset-1	
	Tigga and Garg (2020)	Proposal	Pranto et al. (2020)	Proposal
NB	80.60	86.17	72.20	78.95
DT	84.00	96.81	73.10	76.32
RF	94.10	96.81	77.90	80.26
SVM	86.50	91.49	–	80.26
LR	85.70	84.04	–	77.63
GB	–	91.00	–	78.95
KNN	77.30	90.43	75.70	75.00

Table 5 shows the accuracy comparison of different ML methods compared with existing works for two datasets. It is clear that the proposal outperforms the other existing recent works.

Many previous studies trained their machine learning models using specific datasets and minimal pre-processing. As a result, existing approaches performed better on specific datasets but did not demonstrate promising accuracy on other datasets. To address these issues, we are motivated to investigate numerous machine learning models, multiple datasets, and various pre-processing strategies such as standardization, null value removal, and outlier removal. Preprocessing and training the model on several datasets improves our model's accuracy.

6. Implementation of web application

The construction of a web application for diabetes prediction is described in this section. First, we provide a brief description of development of the web application. Then, we provide the uses of the application by simple experiment.

6.1. Web application development using flask

Flask is a Python-based microweb platform that allows users to add application functionality as if they were built into the framework itself. Fig. 11 shows the basic file structures of the developed application and this development process comprises of four different program modules as follows:

- *model.pkl*- This contains the machine learning model to predict diabetes. As SVM provided the highest accuracy of 78.125% with all the features, we will integrate this as predictive model in the model.pkl file.
- *app.py*- This package includes Flask APIs that receive Diabetes information through GUI or API calls, compute the predicted value using our model, and return it.
- *Template*- The HTML form (index.html) in this folder allows the user to enter diabetes information and shows the expected outcome.
- *Static*- This folder contains the css file which has the styling required for our HTML form.

The application workflow of the proposal is described in Fig. 12 has the following steps:

- The user sends the necessary information required by the application in a Webpage (Step-1).
- The information is sent to the back-end (Step-2:).
- The flask server adopted with the machine learning algorithm predict the results (Step-3 and Step-4).
- Finally, the predicted result is shown in the webpage (Step-5).

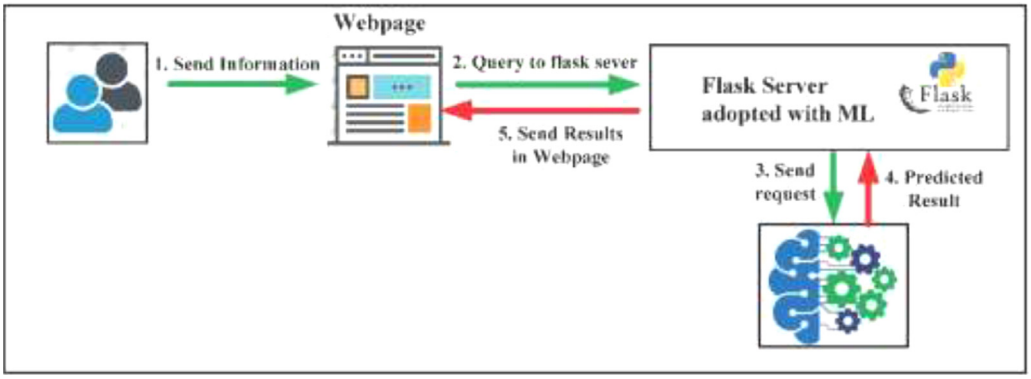


Fig. 12. Working flow of the web application.

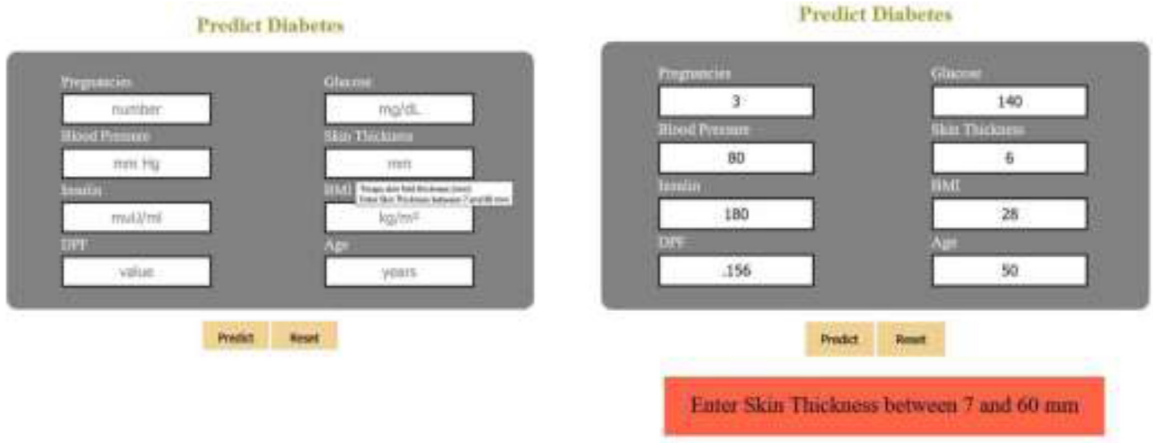


Fig. 13. User input validation in web application.

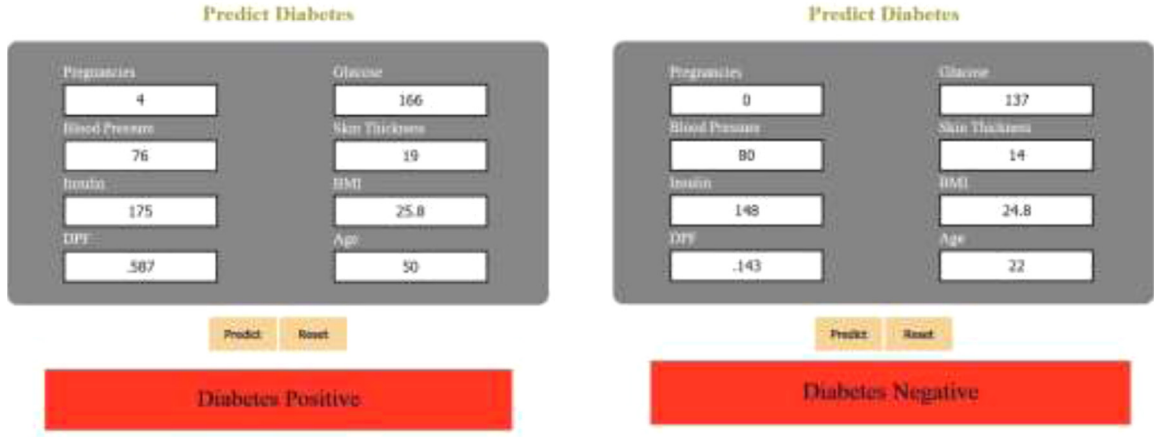


Fig. 14. Prediction results in web application.

6.2. *Prediction results of web application*

When a user runs the application, a page will appear as shown in Fig. 13. The application can check for the valid input for every fields. If the user enters an invalid value for any of the parameters, a warning message is displayed. If the user provides valid information, the application will predict whether the user has diabetes or not, as illustrated in Fig. 14.

7. Conclusion

In our research, firstly, we have adopted several machine learning algorithms and evaluated their performances to predict the diabetes of individuals. Secondly, we have conducted several experiments and evaluated the performances of the proposal. We found that SVM outperforms the other algorithms. Finally, based on our observed results, a smart web application is developed for predicting the diabetes accordingly. Any individual can submit clinical data to this web application, which can then forecast the existence or absence of diabetes. Individuals who are unsure or simply want a routine checkup may consider this application. Our model is compared with two recent studies, and the findings reveal that, depending on the dataset and the ML method used, the suggested model can offer greater accuracy ranging from 2.71% to 13.13%. While we have conducted several experiments utilizing two distinct datasets, this study still has room for additional research and development using a variety of deep learning methods. Additionally, in the future, we will examine a larger and deeper dataset for Bangladeshi patients with additional attributes for improved accuracy, and we will publish the web application on a cloud platform such as Heroku or AWS that is freely available and can be evaluated by real users.

References

Ahamed, K. U., Islam, M., Uddin, A., Akhter, A., Paul, B. K., Yousuf, M. A., . . . Moni, M. A., et al. (2021). A deep learning approach using effective preprocessing techniques to detect covid-19 from chest CT-scan and X-ray images. *Computers in Biology and Medicine, 139*, Article 105014. 10.1016/j.compbiomed.2021.105014.

Albahli, S. (2020). Type 2 machine learning: An effective hybrid prediction model for early type 2 diabetes detection. *Journal of Medical Imaging and Health Informatics, 10*, 1069–1075.

Brownlee, J. (2016a). A gentle introduction to the gradient boosting algorithm for machine learning. https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/ Accessed: 2021-03-20.

Brownlee, J. (2016b). K-nearest neighbors for machine learning. https://machinelearningmastery.com/k-nearest-neighbors-for-machine-learning/ Accessed: 2021-03-20.

Brownlee, J. (2016c). Logistic regression for machine learning. https://www.geeksforgeeks.org/understanding-logistic-regression/ Accessed: 2021-03-20.

Choubey, D. K., Paul, S., Kumar, S., & Kumar, S. (2017). Classification of Pima Indian diabetes dataset using naive Bayes with genetic algorithm as an attribute selection. In *Proceedings of the international conference on communication and computing system (ICCCS 2016)* (pp. 451–455).

Dinh, A., Miertschin, S., Young, A., & Mohanty, S. D. (2019). A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Medical Informatics and Decision Making, 19*, 211.

Gadekallu, T. R., Khare, N., Bhattacharya, S., Singh, S., Reddy Maddikunta, P. K., Ra, I. H., et al. (2020). Early detection of diabetic retinopathy using pca-firefly based deep learning model. *Electronics, 9*, 274.

Gandhi, R. (2018). Naive bayes classifier. https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c Accessed: 2021-03-20.

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Journal of the American Medical Association, 316*, 2402–2410.

Gupta, S. (2019). Decision tree. https://www.geeksforgeeks.org/decision-tree/ Accessed: 2021-03-20.

Haq, A. U., Li, J. P., Khan, J., Memon, M. H., Nazir, S., Ahmad, S., et al. (2020). Intelligent machine learning approach for effective recognition of diabetes in e-healthcare using clinical data. *Sensors, 20*, 2649.

Joshi, T. N., & Chawan, P. (2018). Diabetes prediction using machine learning techniques. *International Journal of Engineering Research and Applications, 8*, 9–13.

Kaur, H., & Kumari, V. (2020). Predictive modelling and analytics for diabetes using a machine learning approach. *Applied Computing and Informatics*. 10.1016/j.aci.2018.12.004.

Kazerouni, F., Bayani, A., Asadi, F., Saeidi, L., Parvizi, N., & Mansoori, Z. (2020). Type2 diabetes mellitus prediction using data mining algorithms based on the long-noncoding rnas expression: A comparison of four data mining approaches. *BMC Bioinformatics, 21*, 1–13.

Kopitar, L., Kocbek, P., Cilar, L., Sheikh, A., & Stiglic, G. (2020). Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Scientific Reports, 10*, 1–12.

Maniruzzaman, M., Rahman, M. J., Ahammed, B., & Abedin, M. M. (2020). Classification and prediction of diabetes disease using machine learning paradigm. *Health Information Science and Systems, 8*, 1–14.

Nilashi, M., Ibrahim, O., Dalvi, M., Ahmadi, H., & Shahmoradi, L. (2017). Accuracy improvement for diabetes disease classification: A case on a public medical dataset. *Fuzzy Information and Engineering, 9*, 345–357.

Perveen, S., Shahbaz, M., Guergachi, A., & Keshavjee, K. (2016). Performance analysis of data mining classification techniques to predict diabetes. *Procedia Computer Science, 82*, 115–121.

Pranto, B., Mehnaz, S., Mahid, E. B., Sadman, I. M., Rahman, A., Momen, S., et al. (2020). Evaluating machine learning methods for predicting diabetes among female patients in Bangladesh. *Information, 11*, 374.

Rajeswari, M., & Prabhu, P. (2019). A review of diabetic prediction using machine learning techniques. *International Journal of Engineering and Techniques, 5*, 1–7.

Ray, S. (2017). Understanding support vector machine(svm). https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/ Accessed: 2021-03-20.

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

UCI Machine Learning Repository. (1998). Diabetes data set. https://archive.ics.uci.edu/ml/datasets/diabetes Accessed: 2021-03-20.

Vinayakumar, R., Alazab, M., Soman, K., Poornachandran, P., Al-Nemrat, A., & Venkatraman, S. (2019). Deep learning approach for intelligent intrusion detection system. *IEEE Access : Practical Innovations, Open Solutions, 7*, 41525–41550.

World Health Organization. (2021). *Diabetes*. World Health Organization https://www.who.int/news-room/fact-sheets/detail/diabetes Accessed: 2021-04-20.

Yahyaoui, A., Jamil, A., Rasheed, J., & Yesiltepe, M. (2019). A decision support system for diabetes prediction using machine learning and deep learning techniques. In *Proceedings of the 1st international informatics and software engineering conference (UBMYK)* (pp. 1–4).

Yiu, T. (2019). Understanding random forest. https://towardsdatascience.com/understanding-random-forest-58381e0602d2 Accessed: 2021-03-20.

Yu, W., Liu, T., Valdez, R., Gwinn, M., & Khoury, M. J. (2010). Application of support vector machine modeling for prediction of common diseases: The case of diabetes and pre-diabetes. *BMC Medical Informatics and Decision Making, 10*, 16.

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. *Frontiers in Genetics, 9*, 515.