STM JOURNALS

Research & Reviews: A Journal of
Bioinformatics

ISSN: 2393-8722
Volume 12, Issue 2, 2025
May—August
DOI (Journal): 10.37591/RRJoBI

https://journals.stmjournals.com/rrjobi

Review RRJoBI

Enhanced Diabetes Prediction: A Comparative Study of Machine Learning Models

Nuzmul Hossain Nahid[1], Abdul Based[2,*]

Abstract

Excessively high blood glucose levels lead to diabetes, a condition that can be better managed with early detection, resulting in a longer life and improved health. Machine learning models are essential tools in diagnosing diabetes, especially when trained on appropriate and relevant datasets. In this study, a combination of ensemble methods and nine distinct machine learning algorithms were utilized to develop a predictive model for diabetes diagnosis based on a publicly accessible dataset. Among the models tested, the Random Forest algorithm demonstrated superior performance, achieving the highest prediction accuracy of 99.75%. This highlights the effectiveness of ensemble-based approaches in enhancing diagnostic precision and underscores the potential of machine learning in supporting clinical decision-making for diabetes detection. The study emphasizes the value of data-driven techniques in improving the early identification and management of diabetes. A comparison with existing studies highlights the strength and superiority of our approach. Additionally, a user-friendly web application has been developed using the best-performing model, providing users with diabetes predictions and relevant educational videos.

Keywords: Diabetes, Machine Learning, Random Forest, accuracy, web application

INTRODUCTION

Major contributing factors to diabetes include obesity, an unhealthy lifestyle, and high blood pressure. It affects millions of people worldwide and can impact nearly every part of the body, leading to complications in vital organs, such as the kidneys, heart, and eyes. However, early detection can significantly reduce these risks. Machine Learning (ML) tools offer a promising solution for identifying diabetes in its early stages, helping individuals take preventive measures before severe complications arise. Unfortunately, a significant number of people worldwide suffer from diabetes without being diagnosed in time, increasing their risk of severe health issues, such as blindness and kidney failure.

Diabetes is diagnosed when blood sugar is at or above average. People with diabetes are prone to heart, eye, and even kidney disease. Diabetes is mainly of two types of Type 1 occur in people under 30, who can be treated with medication and therapy. Type 2 diabetes commonly affects older adults, though it can also develop in individuals during middle age. If identified early, the condition can be managed effectively, or further treatment can be pursued. Adopting a healthy diet and maintaining a balanced lifestyle play a key role in preventing the onset of this disease.

***Author for Correspondence**
Abdul Based
E-mail: based@diu.ac

[1]Researcher, Department of CSE, Dhaka International University, Dhaka, Bangladesh
[2]Professor & Chairman, Department of CSE, Dhaka International University, Dhaka, Bangladesh

Received Date: May 27, 2025
Accepted Date: June 17, 2025
Published Date: July 04, 2025

Citation: Nuzmul Hossain Nahid, Abdul Based. Enhanced Diabetes Prediction: A Comparative Study of Machine Learning Models. Research & Reviews: A Journal of Bioinformatics. 2025; 12(2): 1–10p.

Numerous nations are expected to have significant social and economic consequences from the rising prevalence of diabetes. We must maintain healthy food chart if we want to keep our physical and mental stability while managing diabetes. Diabetes causes high blood and urine glucose levels, which over time can cause serious additional health

issues like blindness and renal failure. The management of diabetes depends on the early and correct identification of the condition. However, if diabetes could be predicted early, so many lives may have been saved.

As the number of individuals affected by diabetes continues to rise and the demand for testing increases, the medical field increasingly relies on Artificial Intelligence (AI) for support. AI technology enables individuals to assess their risk of developing diabetes without incurring financial costs.

This research explores various ML classification methods to develop a predictive model using clinical dataset. The data set was collected first, and the obtained dataset was pre-processed. The pre-processed dataset was then supplied to nine ML models. Each model's output is assessed using various evaluation metrics. Ultimately, the model with the highest accuracy in detecting diabetes is selected for the development of a web-based application.

The key contributions of this study include:

- 99.75% accuracy.
- High performance.
- A user-interactive web application.

Outline of the Paper: Section 2 briefly discusses the relevant existing works. Section 3 presents the methodology of this research work. Section 4 provides the results followed by analysis and comparison with others. Section 5 concludes the paper with a discussion of the findings agent.

RELATED WORKS

A model that depends on various Machine Learning (ML) Algorithms is published in Bothra R. (2021) [1]. The models used in this work achieved 90% accuracy. A framework-based ML algorithm is proposed in Sahoo et al. (2020) [2]. The highest accuracy of this work is 78% using Logistic Regression (LR). Jitranjan et al. [3] proposed a ML model utilizing classification algorithms with 79.17% accuracy using LR. Mitsubishi Soni [4] developed a model using six classifiers and achieved the best accuracy at 77% with Random Forest (RF).

Md. Faisal Faruque et al. [5] applied five ML algorithms in their work and got 74% accuracy as the highest with C4.5. In another study, Jingyu Xue et al. [6] introduced a model that combined SVM, NB, and LightGBM (LGBM), with the Naïve Bayes classifier outperforming the others by achieving 96.54% accuracy as the highest.

N. Sneha et al. [7] achieved 98.20% accuracy as the highest by applying the Decision Tree (DT). Ritik [8] presented a model using two classifiers: RF and XGBoost. The dataset contained 768 samples with nine attributes. The XGBoost algorithm outperformed RF, achieving the highest accuracy of 85.24%. Sallah Shafi et al. [9] proposed a model incorporating NB, SVM, and DT. Among these, NB attained the highest accuracy of 74.28%.

KM Jyoti et al. [10] proposed a machine learning framework that employed five different algorithms. Among them, the Decision Tree (DT) classifier delivered the highest performance, achieving 98% accuracy during training and 99% during testing. Avantika Nahar et al. [11] attained an accuracy of 98% using the K-Nearest Neighbor (KNN) algorithm. Olta Llaha et al. [12] achieved 79% accuracy as the highest using the DT.

Compared to the existing works, we implement nine ML algorithms in this work and get 99.75% accuracy using the Random Forest (RF).

METHODOLOGY

To predict diabetes, nine different models were developed and used to create a predictive system for diabetes diagnosis. The effectiveness of the algorithms was assessed using confusion matrix-based metrics to evaluate their performance. Based on accuracy levels, the best-performing model was selected and integrated into a user-friendly web application, enabling users to check their diabetes risk on smartphones or computers.

As illustrated in Figure 1, the dataset was collected from an online source and underwent a pre-processing phase. During this stage, essential features were identified using feature selection techniques. Then the data was divided into training and testing sets. Finally, a web application is developed.

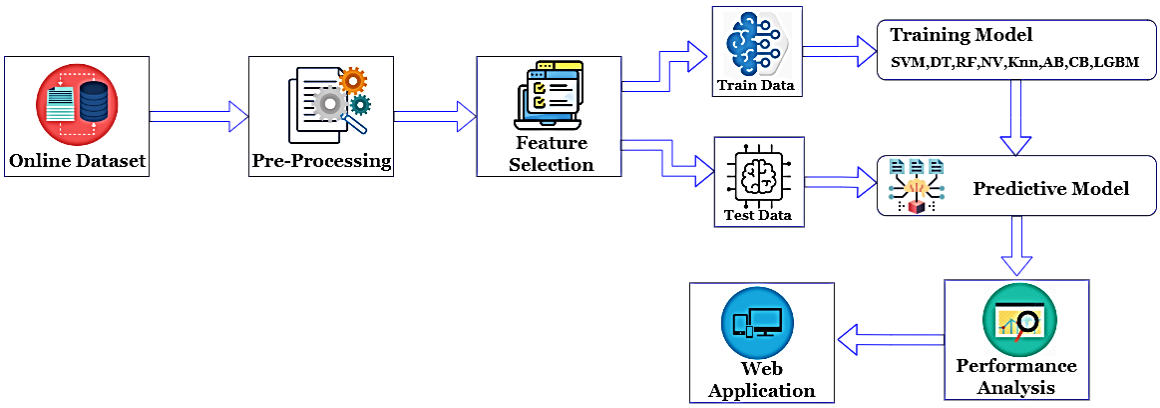

Figure 1. The methodology.

Data Collection

The dataset consists of 2,000 records with a total of ten columns. The data was sourced from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) [13]. Specific selection criteria were applied to extract relevant instances from a larger database. Among the 2,000 patient records, 1,316 cases have a positive diabetes diagnosis, while 684 cases are negative. The dataset contains nine predictive features.

Data Pre-Processing

This technique is applied to clean, format, and optimize the dataset to ensure accurate predictions and reliable performance. In many cases, patient records contain incomplete information, which can affect model performance. To identify missing values within a dataset, the isna().sum() command is used.

If the correlation is 0, the correlation is neural, and if the correlation is –1, then the correlation is negative or not strong. If the correlation is 1, it means the correlation is strong. Figure 2 shows the correlation matrix. Here, Pregnancies has the highest correlation (1) and Age has a strong correlation (0.54). Thickness has a less strong correlation (–0.11).

To enhance model performance while reducing overfitting and computational overhead, feature selection is employed. This approach focuses on identifying the most relevant attributes, thereby, improving the efficiency and accuracy of machine learning models.

Filter-based feature selection methods work independently of the learning algorithm, choosing features that exhibit a strong correlation with the target variable. To assess the relevance of selected features, these methods use standard evaluation criteria separate from the learning process. In this study, a commonly used information-based criterion was applied for feature selection. Mutual information, which measures the dependency between random variables, was used to determine feature importance.

Figure 3 describes the important features in the dataset; to find the important feature, we use the filter selection method. The Pregnancies score is 0.110849, the score for Glucose is 0.220992. For blood pressure, the score is 0.100431 and Skin Thickness score is 0.081882. Insuline score is 0.079309, BMI score is 0.140666, Diabetes Pedigree Function score is 0.117545 and Age has a score of 0.148326.

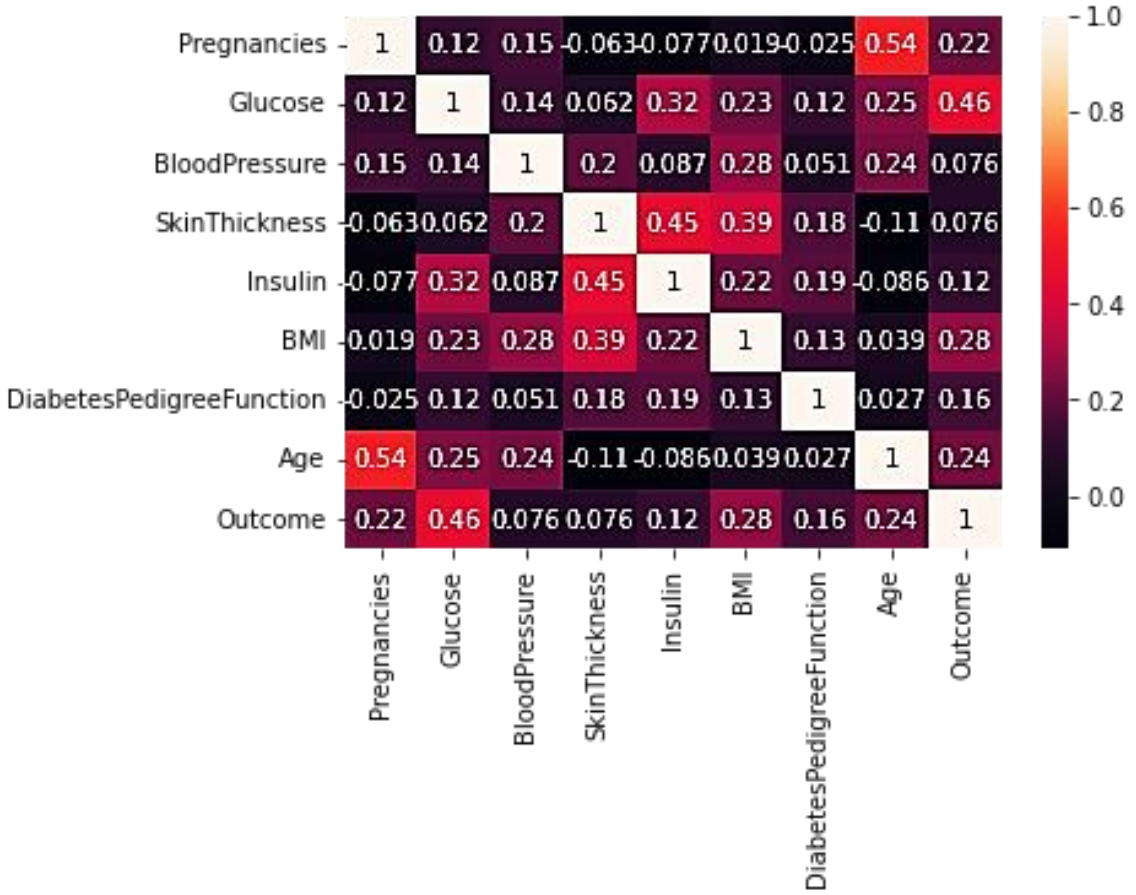

Figure 2. Correlation matrix.

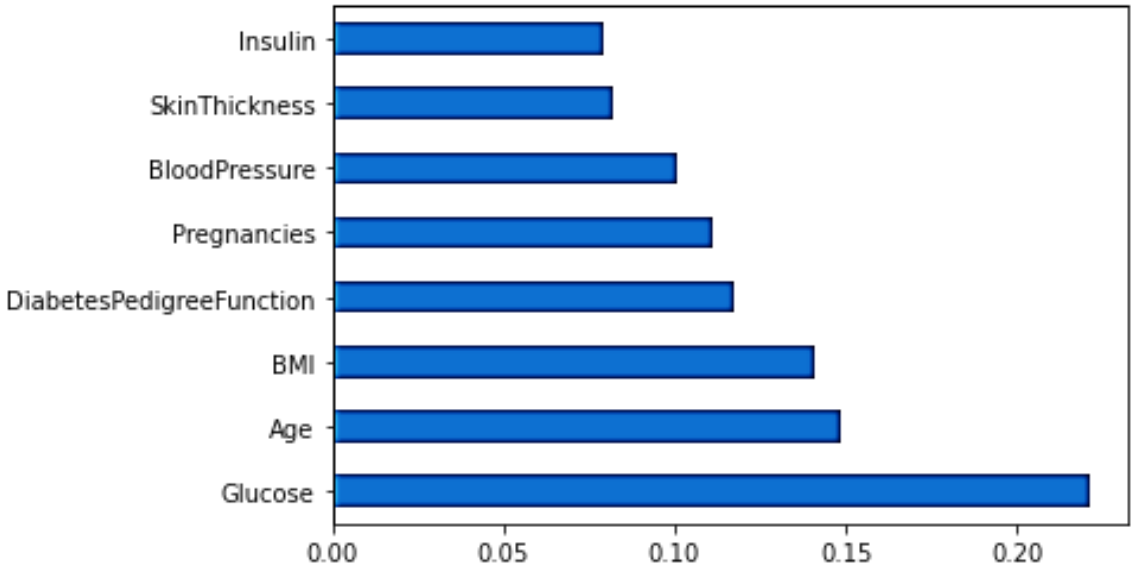

Figure 3. Important feature in the dataset.

MACHINE LEARNING MODELS

The Support Vector Machine (SVM) [14] algorithm is very powerful for classification and regression. It is a supervised learning model that aims to find the optimal hyperplane. Like SVM, a Decision Tree (DT) [15] is also a supervised learning model used for classification as well as for regression. The DT can create overfitting problem if proper pruning is not done. The Random Forest (RF) [16] improves accuracy and reduces overfitting by building multiple decision tress and combining their outputs. For regression, the final prediction is made by averaging. For classification, the final decision is made by majority voting. It is an ensemble learning method.

The Naive Bayes (NB) is a well-known algorithm to address classification problems by employing the Bayes' theorem [17]. The K-Nearest Neighbour (KNN) [18] algorithm is primarily used for classification, although it can also be applied to regression tasks. It determines the similarity between a new instance and existing instances to make predictions. The Logistic Regression (LR) model [19] forecasts the categorical dependent variable. The LR is a set of independent variables and can predict the output of the categorical dependant variables. The Adaptive Boosting (AdaBoost) [20] is a ML model that employs decision trees as the base learner for boosting.

The CatBoost algorithm is a powerful technique for supervised machine learning, built on the principles of Gradient Descent [21]. It is especially well-suited for problems that involve categorical data. LightGBM, which is also built on decision tree principles, is versatile and can be applied to a range of machine learning tasks, such as classification and ranking. It has gained popularity among Kaggle participants, who have used it to achieve success in data science competitions. LightGBM enhances model performance while consuming less memory. To overcome the limitations of the histogram-based approach, it introduces several improvements, making it more efficient for large datasets [22].

Web Application Development

The Flask framework is used and integrated with the RF to develop the web application. The users are prompt to provide necessary inputs to predict diabetes. Then the ML model is applied, and the result is shown accordingly.

RESULTS AND ANALYSIS

This template is based on Version V2. Most of the formatting guidelines provided in this document have been adapted by Causal Productions from the IEEE (Doc) style files.

The Jupyter Notebook and Python 3.3 are used in this work. In addition, many libraries have been created from sklearn, a free Python library for machine learning [20]. The F1 measure and sensitivity, precision, and recall are used as the evaluation standard. The test size is 0.20, and the train split is 0.80 for the classifier algorithms.

The ML models were analysed in terms of confusions metrics. Since, RF provided the best accuracy, it was selected for the web application. The following equations are used to calculate the metrices [21].

$$\text{Accuracy} = \frac{AP+AN}{AP + PN + PP + AN} \tag{1}$$

$$\text{Recall} = \frac{AP}{AP + PN} \tag{2}$$

$$\text{Precision} = \frac{AP}{AP + PP} \tag{3}$$

$$\text{F1 score} = 2 * \frac{precision * recall}{precision + recall} \tag{4}$$

Table 1 contains confusion matrix scores. The diabetes models' actual positive (AP) value shows how accurately they were understood. Predict-negative (PN) results suggest that the diabetes sample

analyses were done improperly. Predict-positive (PP) denotes proper calculation of the observed samples. The actual negative (AN) signifies that the samples with no diabetes were appropriately identified. In this study, employing the parameters of glucose, diabetes pedigree function, BMI, age, and insulin, all models exhibit outstanding performance in identifying diabetes with an accuracy of up to 99.75%. In Table 2, we achieved a Random Forest accuracy of 99.75% utilizing these features. The accuracy of the Catboost and LGBM is 99.25%. With DT, the accuracy is 98.25% and with AdaBoost, the score is 81.5%. The accuracy of KNN, Support Vector, Logistic Regression, and Naive Bayes is 81.25%, 81%, 77.25%, and 76.75%, respectively. The Naive Bayes classifier in this investigation has the lowest accuracy, 76.75%. The random forest has the highest accuracy of the model, at 99.75%.

Table 1. Data reduction table of ML models.

Classifier	Confusion Matrix	Accuracy		Precision	Recall	F1-Score	Support
SVM	[252, 20]	81	0	0.82	0.93	0.87	272
			1	0.78	0.56	0.65	128
	[56, 72]		Macro Avg	0.8	0.74	0.76	400
			Weighted avg	0.81	0.81	0.8	400
Decision Tree	[267, 5]	98.75	0	1	0.98	0.99	272
			1	0.96	1	0.98	128
	[0, 128]		Macro Avg	0.98	0.99	0.99	400
			Weighted avg	0.99	0.99	0.99	400
Random Forest	[271, 1]	99.75	0	1	1	1	272
			1	0.99	1	1	128
	[0, 128]		Macro Avg	1	1	1	400
			Weighted avg	1	1	1	400
Native Bayes	[242, 30]	76.75	0	0.79	0.89	0.84	272
			1	0.68	0.51	0.58	128
	[63, 65]		Macro Avg	0.74	0.7	0.71	400
			Weighted avg	0.76	0.77	0.76	400
KNN	[236, 36]	81.25	0	0.86	0.87	0.86	272
			1	0.71	0.7	0.7	128
	[39, 89]		Macro Avg	0.79	0.78	0.78	400
			Weighted avg	0.81	0.81	0.81	400
AdaBooster	[240, 32]	81.5	0	0.85	0.88	0.87	272
			1	0.73	0.67	0.7	128
	[42, 86]		Macro Avg	0.79	0.78	0.78	400
			Weighted avg	0.81	0.81	0.81	400
Logistic regression	[239, 33]	77.25	0	0.8	0.88	0.84	272
			1	0.68	0.55	0.61	128
	[58, 70]		Macro Avg	0.74	0.71	0.72	400
			Weighted avg	0.76	0.77	0.77	400
LGBM, CatBoost	[240, 32]	99.25	0	1	0.99	0.99	272
			1	0.98	1	0.99	128
	[42, 86]		Macro Avg	0.99	0.99	0.99	400
			Weighted avg	0.99	0.99	0.99	400

All models have excellent performance against projected diabetes, with a highest accuracy of 99.75%, according to the evaluation results, as shown in Figure 4.

The application is shown in Figure 5. Random Forest is used as the prediction model in the model.py file because it had a maximum accuracy of 99.75% across all the features.

Using this model, the server.py package's Flask APIs calculate the expected value and return it after receiving Diabetes data via a GUI or API query.

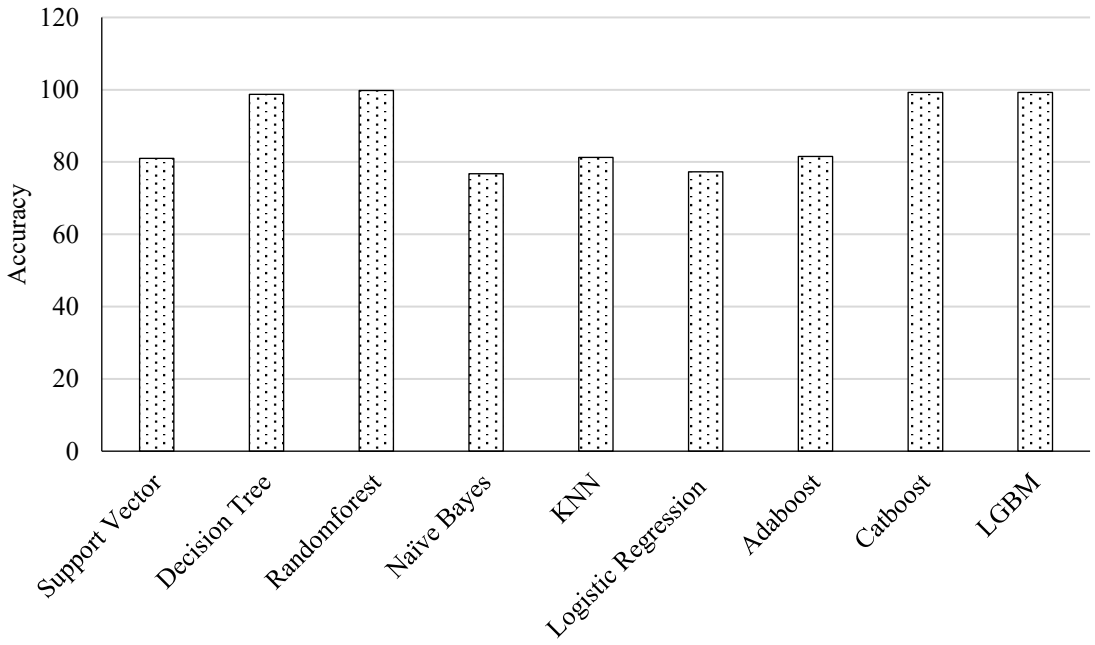

Figure 4. Accuracies obtained by different ML models.

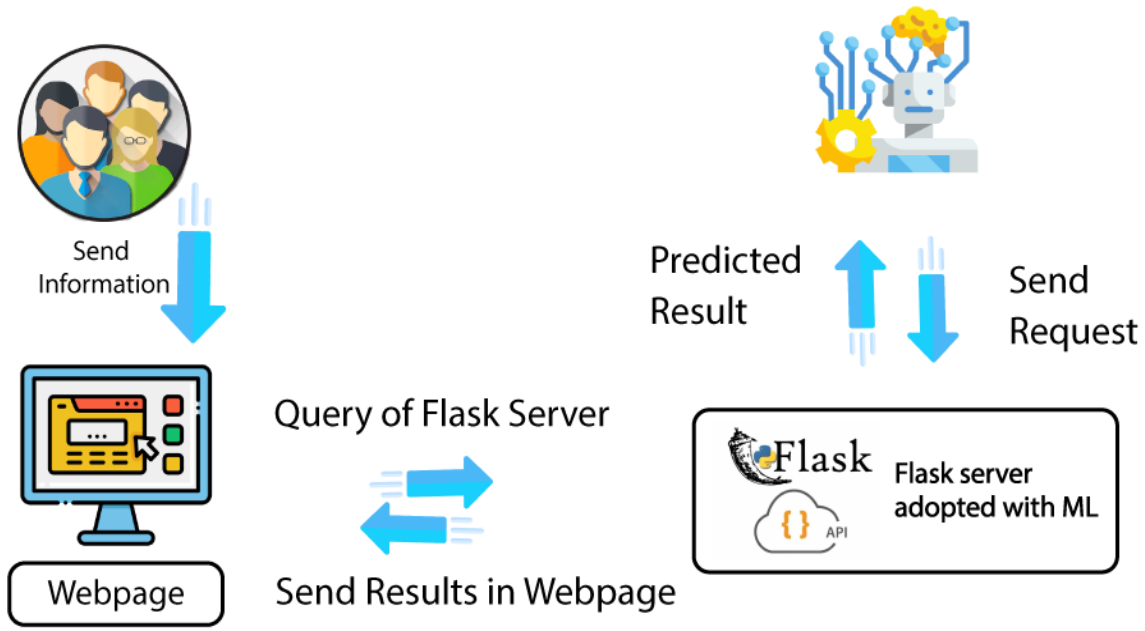

Figure 5. Web application.

The application validates each input. For invalid inputs, a warning message appears. Figure 6 depicts the web application's predicted positive outcomes. The application checks the input field, then predicts possible ML algorithms and provides the predicted result. The web application's expected negative results are shown in Figure 7. This application examines the input field, forecasts potential ML methods,

and then displays the forecasted outcome. The application is also capable of showing relevant videos for the users after the prediction of diabetes.

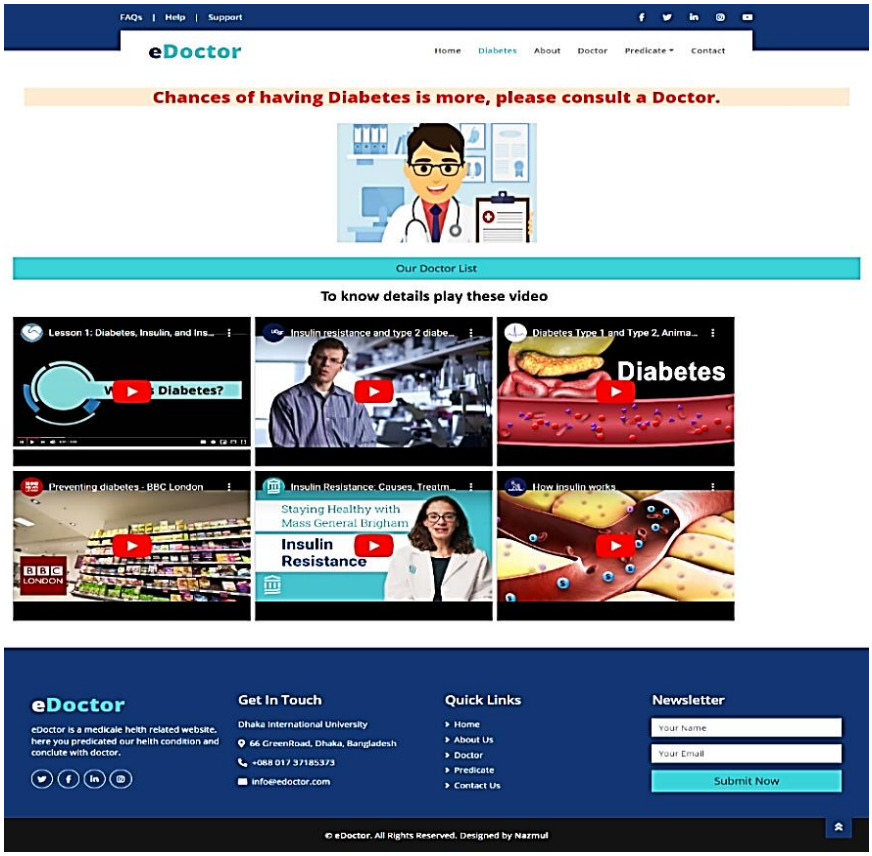

Figure 6. Predict of positive results in web application.

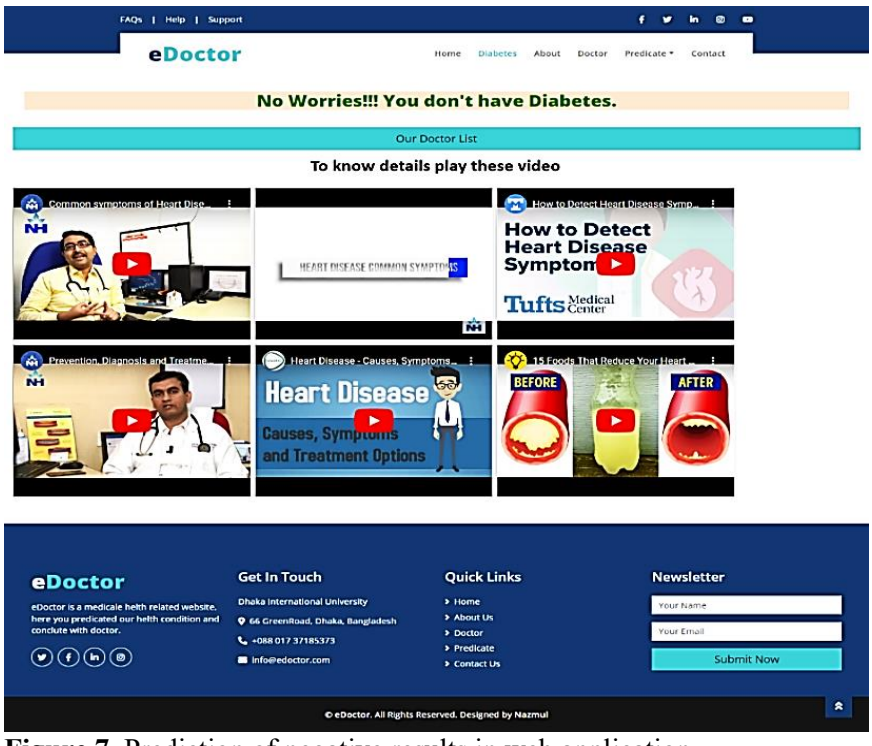

Figure 7. Prediction of negative results in web application.

The comparison of this work with the relevant existing/published works is shown in Table 2.

Table 2. Comparison table.

Existing Work	Dataset	Method	Accuracy
Rishab Bothra [1]	This dataset contains 1405 rows, and 10 columns.	RF, LR, XGBoost, SVM, KNN.	90% (RF)
Kishan Patel et al. [2]	This dataset was collected from MDC (700 patients)	RF, KNN, DT, LR	78% (LR)
Jitranjan et al. [3]	The dataset collected from the National Institute of diabetes	LR, KNN, DT, RF, SVM	79.17% (LR)
Mitushi Soni [4]	This dataset 768 patients gathered from UCI repository.	SVM, KNN, DT, LR, GB	77% (RF)
Faruque et al. [5]	Dataset of 200 patients is collected from (MCC), Bangladesh.	SVM, NB, KNN and C4.5	74% (C4.5)
Xue et al. [6]	This dataset of 500 patients was collected from the UCI.	SVM, NB, LGBM	99.17% (RF)
N. Sneha et al. [7]	This dataset comes from UCI and contains information of 2500 patients.	NB, SVM, DT, RF, KNN	98.20% (DT)
Ritik [8]	This dataset contains 768 samples and 9 attributes.	RF, XGBoost	85.24% (XGBoost)
Shafi et al. [9]	The dataset has 767 female patients with 7 attributes.	NB, SVM, DT	74.28% (NB)
Jyoti et al. [10]	Dataset was collected from kaggale.com.	KNN, LR, DT, RF, SVM	99% (DT)
Nahar et al. [11]	This dataset has 21 attributes and 1000 data.	KNN	98% (KNN)
Llaha et al. [12]	Eight attributes from the dataset are used.	NB, DT, SVM, LR	79% (DT)
The proposed work	The dataset has 2000 patients' information	SVM, DT, RF, NB, KNN, LR, AdaBoost, CatBoost, and LGBM	99.75% (RF)

CONCLUSIONS

This study explores the feasibility of using machine learning algorithms to predict diabetes while minimizing the number of required tests or features. The concept was evaluated through a series of experiments, and a web application was developed for diabetes prediction. Nine machine learning algorithms – SVM, Decision Tree, Random Forest, Naïve Bayes, K-Nearest Neighbor, Logistic Regression, AdaBoost, CatBoost, and LightGBM – were applied to a dataset containing 2,000 records.

Among these models, the Random Forest approach was the best with an accuracy of 99.75%, an F1-score of 100%, precision of 100%, and recall of 100%. This result surpasses previous studies while utilizing fewer features, making the approach more cost-effective. The findings suggest that diabetes can be accurately identified using only five key features.

However, given the relatively small dataset used in this study, future work aims to validate the results using a larger dataset or apply the model to a dataset with identical attributes for further comparison.

REFERENCES

1. Bothra R. Diabetes prediction using machine learning algorithms. Int J Eng Appl Sci Technol. 2021;6(5):151–4. ISSN: 2455-2143.
2. Sahoo J, Dash M, Pati A. Diabetes prediction using machine learning classification algorithms. Int Res J Eng Technol (IRJET). 2020 Aug;7(8):e-ISSN: 2395-0056.
3. Patel KU, Sunyecz IL, McCallinhart PE, Bartlett CW, Trask AJ. Applied predictive modeling of coronary microvascular disease using coronary Doppler and cardiac echocardiography. FASEB J. 2018 Apr;32(S1). doi:10.1096/fasebj.2018.32.1_supplement.784.9.
4. Mitushi S, Sunita V. Diabetes prediction using machine learning techniques. Int J Eng Res Technol (IJERT). 2020;9(1). ISSN: 2278-0181.

5. Faruque MF, Sarker IH. Performance analysis of machine learning techniques to predict diabetes mellitus. In: 2019 Int Conf on Electrical, Computer and Communication Engineering (ECCE); 2019 Feb. p. 1–6.
6. Xue J, Min F, Ma F. Research on diabetes prediction method based on machine learning. J Phys Conf Ser. 2020;1684(1):012062.
7. Sneha N, Gangil T. Analysis of diabetes mellitus for early prediction using optimal features selection. Big Data. 2019;6(1):13. doi:10.1186/s40537-019-0175-6.
8. Baby ST, Karunakaran V. Prediction of diabetics using machine learning classifiers: A review. In: 2021 5th Int Conf on I-SMAC (IoT in Social, Mobile, Analytics and Cloud); 2021 Nov. p. 735–9.
9. Shafi S, Ansari GA. Early prediction of diabetes disease & classification of algorithms using machine learning approach. In: Proc Int Conf on Smart Data Intelligence; 2021 May. p. 453–8.
10. Rani KJ. Diabetes prediction using machine learning. Int J Sci Res Comput Sci Eng Inf Technol. 2020; DOI:10.32628/CSEIT206463.
11. Premamayudu B, Muralikrishna K, Pramodh K. Diabetes prediction using machine learning KNN-algorithm technique. Int J Innov Sci Res Technol. 2022 May;7(5). ISSN: 2456-2165.
12. Llaha O, Rista A. Prediction and detection of diabetes using machine learning. In: Proc 4th Int Conf on Recent Trends and Applications in Computer Science and Information Technology; 2021 May.
13. National Institute of Diabetes and Digestive and Kidney Diseases. Available from: https://www.niddk.nih.gov/. Accessed 10 Jan 2025.
14. Jakkula V. Tutorial on support vector machine (SVM). Pullman, WA: School of EECS, Washington State University; 2006.
15. Charbuty B, Abdulazeez A. Classification based on decision tree algorithm for machine learning. J Appl Sci Technol Trends. 2021 Mar;2(1):20–8. doi:10.38094/jastt20165.
16. Sarica A, Cerasa A, Quattrone A. Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: A systematic review. Front Aging Neurosci. 2017;9:329.
17. Vijayarani S, Dhayanand S. Liver disease prediction using SVM and Naïve Bayes algorithms. Int J Sci Eng Technol Res. 2015 Apr;4(4):816–20.
18. Imandoust SB, Bolandraftar M. Application of k-nearest neighbor (KNN) approach for predicting economic events. Int J Eng Res Appl. 2013 Sep–Oct;3(5):605–10.
19. Shevade SK, Keerthi SS. A simple and efficient algorithm for gene selection using sparse logistic regression. Bioinformatics. 2003 Dec;19(17):2246–53. doi:10.1093/bioinformatics/btg308. PMID: 14630653.
20. Sevinç E. An empowered AdaBoost algorithm implementation: A COVID-19 dataset study. Comput Ind Eng. 2022 Mar;165:107912.
21. Zhou F, Pan H, Gao Z, Huang X, Qian G, Zhu Y, et al. Fire prediction based on CatBoost algorithm. Math Probl Eng. 2021;2021:1929137.
22. Ahamed BS. Prediction of type-2 diabetes using the LGBM classifier methods and techniques. Turk J Comput Math Educ (TURCOMAT). 2021;12(12):2807–13.

