Research Article

Predicting Diabetes Using Machine Learning Approaches

Ratna Rathaur, Shivam Pandey, Ashutosh Mani*

Department of Biotechnology Motilal Nehru National Institute of Technology Allahabad, INDIA

*Corresponding author: amani@mnnit.ac.in

Abstract

Diabetes mellitus is a chronic metabolic disorder and a major public health threat, especially in developing countries like India. Early detection of diabetes may be helpful in preventing long-term complications including cardiovascular disease, nephropathy, neuropathy, and retinopathy in patients. In this study, machine learning (ML) techniques were used to develop a model for predicting early diabetes risk assessment using routinely collected clinical parameters. The widely used Pima Indian Diabetes Dataset obtained from the National Institute of Diabetes and Digestive and Kidney Diseases was used in this work. Several supervised ML algorithms, including Logistic Regression, Decision Tree, Naive Bayes, Support Vector Machine, and Random Forest, were used for prediction and were later evaluated. Model performance was assessed by using accuracy, precision, recall, F1-score, and ROC–AUC. Among the estimated models, the Random Forest classifier demonstrated superior performance with significantly balanced sensitivity and specificity. These findings highlight the potential of ML-based decision support systems for prediction of early diabetes prediction by using clinical data. This approach aligns with the principles of integrative and computational biology for disease diagnosis and treatment.

Keywords: Diabetes mellitus, Machine learning, Random Forest, Risk prediction, Integrative biology

Received on: 12.11.2025
Accepted on: 22.12.2025
Published on 26.12.2025

Introduction

One of the most common non-communicable diseases in the world is diabetes mellitus. It impacts over 537 million adults in the world, and the burden is increasing rapidly in India [1,2]. T2DM causes about 90 out of all cases of diabetes and has strong links with lifestyle-associated factors (obesity, lack of physical activity, and improper diet) [3]. Late detection of T2DM normally causes various complications such as cardiovascular diseases, kidney failure, neuropathy, and vision loss.

The conventional diagnostic approaches are based on biochemical tests like fasting plasma glucose and oral glucose tolerance tests and glycated hemoglobin (HbA1c). On the one hand, these approaches are quite effective, but on the other hand, invasive, necessitate clinical facilities, and are not available to many population groups [4]. This has led to the increased demand of alternative data-driven and cost effective methods that have the potential of supporting the prediction of diabetes risks at an early and scalable stage.

The latest developments in the field of artificial intelligence and machine learning have made it possible to analyze more complex biomedical data and reveal hidden

patterns related to the development of diseases [5,6]. The machine learning algorithms have demonstrated promising results in predicting chronic diseases, such as diabetes, through the combination of several risk factors at once [7]. The scope of the research is to assess and compare several ML models to predict diabetes and select the most useful method to assess the patient at risk in the early stages of the disease.

Materials and Methods

Dataset Description

The dataset used in this study was obtained from the National Institute of Diabetes and Digestive and Kidney Diseases and is publicly available for research purposes [8]. The dataset consists of 768 female patients of Pima Indian heritage aged 21 years or older. Among them, 268 individuals were diagnosed with diabetes, while 500 were non-diabetic. The dataset includes eight clinical features: number of pregnancies, plasma glucose concentration, diastolic blood pressure, triceps skinfold thickness, serum insulin level, body mass index (BMI), diabetes pedigree function, and age.

Data Preprocessing

Data preprocessing was performed to enhance data quality and model performance. Zero values in biologically implausible attributes such as glucose, blood pressure, skin thickness, insulin, and BMI were treated as missing values and replaced using statistical imputation techniques [9]. Feature scaling was performed using Min–Max normalization to ensure uniform contribution of all features during model training.

Machine Learning Models

Five supervised machine learning algorithms were implemented: Logistic Regression, Decision Tree, Naive Bayes, Support Vector Machine, and Random Forest. These algorithms were selected based on their widespread application in biomedical classification tasks [6,7]. The dataset was split into training (80%) and testing (20%) subsets.

Model Evaluation

Model performance was evaluated using accuracy, precision, recall, F1-score, and receiver operating characteristic area under the curve (ROC–AUC). These metrics provide a comprehensive assessment of model reliability, particularly for medical diagnosis where minimizing false negatives is critical [10].

Results

Model Performance Comparison

The performance of all implemented machine learning models was evaluated on the test dataset. The Random Forest classifier achieved the highest overall accuracy and F1-score, followed by Decision Tree and Naive Bayes models. Logistic Regression and Support Vector Machine showed comparatively lower predictive performance.

3.2 Feature Importance Analysis

Feature importance analysis performed using the Random Forest model indicated that plasma glucose concentration was the most significant predictor of diabetes risk, followed by body mass index (BMI), age, and diabetes pedigree function.

3.3 Receiver Operating Characteristic (ROC) Analysis

The ROC curve analysis demonstrated that the Random Forest model achieved the highest area under the curve (AUC), indicating superior discriminative ability compared to other classifiers.

Table 1. Performance comparison of machine learning models for diabetes prediction.

Model	Accuracy (%)	Precision	Recall	F1-score	ROC–AUC
Logistic Regression	74.1	0.72	0.70	0.71	0.78
Decision Tree	76.8	0.75	0.74	0.74	0.81
Naive Bayes	75.9	0.74	0.73	0.73	0.80
Support Vector Machine	73.5	0.71	0.69	0.70	0.77
Random Forest	81.6	0.80	0.79	0.79	0.87

Discussion

These findings highly show the utility of machine learning methods in early diagnosis of diabetic conditions and risk forecasting with available clinical records in the public. The excellent effect of the performance of the Random Forest model can be credited to its learning mechanism, which improves robustness and lowers the overfitting [11]. The results are in line with the findings of other researchers who have carried out studies with regard to the efficacy of ensemble-based techniques in predicting chronic diseases [12,13].

The clinical use of ML-based diabetes predictors models on healthcare systems can assist clinicians in the early diagnosis and individual interventions. Nevertheless, the constraints of population particularity and absence of longitudinal data should be discussed in the future studies. The use of lifestyle, genetic and biochemical markers can also enhance predictive accuracy.

Conclusion

This study highlights the potential of machine learning techniques for early and non-invasive diabetes risk prediction. Among different evaluated models, the Random Forest classifier method demonstrated to have relatively superior and balanced performance. The proposed approach combines clinical data with computational intelligence and may contribute to improved preventive healthcare strategies. Future work should focus on expanding dataset diversity and deploying ML models in real-world clinical and web-based applications.

References:

1. International Diabetes Federation. IDF Diabetes Atlas, 10th ed. Brussels; 2021.
2. Anjana RM, et al. Prevalence of diabetes in India. *Lancet Diabetes Endocrinol*. 2017;5:585–596.
3. DeFronzo RA, et al. Type 2 diabetes mellitus. *Nat Rev Dis Primers*. 2015;1:15019.
4. American Diabetes Association. Standards of medical care in diabetes. *Diabetes Care*. 2022;45:S1–S264.
5. Deo RC. Machine learning in medicine. *Circulation*. 2015;132:1920–1930.
6. Rajkomar A, et al. Machine learning in medicine. *N Engl J Med*. 2019;380:1347–1358.
7. Kavakiotis I, et al. Machine learning and data mining in diabetes research. *Comput Struct Biotechnol J*. 2017;15:104–116.
8. Smith JW, et al. Using the ADAP learning algorithm to forecast diabetes. *Proc Symp Comput Appl Med Care*. 1988:261–265.
9. Little RJA, Rubin DB. Statistical analysis with missing data. Wiley; 2019.
10. Powers DM. Evaluation: From precision, recall and F-measure to ROC. *J Mach Learn Technol*. 2011;2:37–63.

11. Breiman L. Random forests. *Mach Learn*. 2001;45:5–32.
12. Weng SF, et al. Cam Machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One*. 2017;12:e0174944.
13. Dritsas E, Trigka M. Data-driven ML for diabetes prediction. *Sensors*. 2022;22:239.