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 challenge worldwide, particularly in developing countries such as India. Early detection of diabetes is essential to prevent long-term complications including cardiovascular disease, nephropathy, neuropathy, and retinopathy. In this study, machine learning (ML) techniques were employed to develop a predictive model for 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 utilized. Several supervised ML algorithms, including Logistic Regression, Decision Tree, Naive Bayes, Support Vector Machine, and Random Forest, were implemented and evaluated. Model performance was assessed using accuracy, precision, recall, F1-score, and ROC–AUC. Among the evaluated models, the Random Forest classifier demonstrated superior performance with balanced sensitivity and specificity. The findings highlight the potential of ML-driven decision support systems for scalable, non-invasive, and early diabetes prediction, aligning with the principles of integrative and computational biology.

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

Diabetes mellitus is one of the most prevalent non-communicable diseases globally, affecting more than 537 million adults worldwide, with a rapidly rising burden in India [1,2]. Type 2 diabetes mellitus (T2DM) accounts for approximately 90% of all diabetes cases and is strongly associated with lifestyle factors such as obesity, physical inactivity, and poor dietary habits [3]. Delayed diagnosis often results in severe complications including cardiovascular diseases, renal failure, neuropathy, and vision impairment.

Conventional diagnostic methods rely on biochemical tests such as fasting plasma glucose, oral glucose tolerance tests, and glycated hemoglobin (HbA1c). While effective, these methods are invasive, require clinical infrastructure, and may not be accessible to large sections of the population [4]. Consequently, there is a growing need for alternative, data-driven approaches that can facilitate early and scalable diabetes risk prediction.

Recent advancements in artificial intelligence and machine learning have enabled the analysis of complex biomedical datasets to identify hidden patterns associated with disease onset [5,6]. Machine learning algorithms have shown promising performance in chronic disease prediction, including diabetes, by integrating multiple risk factors simultaneously [7]. This study aims to evaluate and compare multiple ML models for diabetes prediction and identify the most reliable approach for early risk assessment.

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.

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

Model	Accur-acy (%)	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

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.

Figure 1. Feature importance ranking of clinical parameters for diabetes prediction using the Random Forest model. *(Bar graph showing relative importance of glucose, BMI, age, pedigree function, insulin, blood pressure, skin thickness, and pregnancies.)*

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.

Figure 2. ROC curves of different machine learning models for diabetes prediction. *(Graph comparing ROC curves of Random Forest, Decision Tree, Naive Bayes, Logistic Regression, and SVM.)*

Discussion

The results demonstrate the effectiveness of machine learning approaches for early diabetes risk prediction using routinely available clinical data. The superior performance of the Random Forest model can be attributed to its ensemble learning mechanism, which enhances robustness and reduces overfitting [11]. These findings are consistent with previous studies that have reported the effectiveness of ensemble-based methods in chronic disease prediction [12,13].

Integrating ML-based diabetes prediction models into healthcare systems can support clinicians in early diagnosis and personalized intervention strategies. However, limitations such as population specificity and lack of longitudinal data must be addressed in future research. Incorporation of lifestyle, genetic, and biochemical markers may further improve predictive accuracy.

Conclusion

This study highlights the potential of machine learning techniques for early and non-invasive diabetes risk prediction. Among the evaluated models, the Random Forest classifier demonstrated superior and balanced performance. The proposed approach aligns with integrative biology by combining 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. Machine-learning cardiovascular risk prediction. *PLoS One*. 2017;12:e0174944.
13. Dritsas E, Trigka M. Data-driven ML for diabetes prediction. *Sensors*. 2022;22:239.