RESEARCH PAPER

AI-Based Early Prediction of Type 2 Diabetes Using Clinical and Lifestyle Risk Factors: A Machine Learning Study

Ashu[1], Kiran Deshpande*[2], Shailendra Singh Narwariya[3], Mayur Sharad Patel[4], Sanjesh Rathi[5], Shubham Singh[6]

1. Department of Pharmaceutical Sciences, Maharshi Dayanand University, Rohtak, Haryana, India ashu.rp.pharma@mdurohtak.ac.in
2. Associate Professor, A. P. Shah Institute of Technology, Thane, Maharashtra, India. kbdeshpande@apsit.edu.in
3. Associate Professor, ITM University, Gwalior, Madhya Pradesh, India. shailugsp@gmail.com
4. Professor and Vice Principal, Department of Pharmacy, Nandurbar Taluka Vidhayak Sanstha's College of Pharmacy, Nandurbar, Maharashtra 425412, India. Email: mayurpatel1212@gmail.com
5. Professor and Principal, School of Pharmacy, Rai University, Ahmedabad, Gujarat India rathi.sanjesh@gmail.com
6. Department of Pharmaceutics, ISF College of Pharmacy, Moga, Punjab, India singhrbgj@gmail.com

***Corresponding Author:** Kiran Deshpande, Associate Professor, A. P. Shah Institute of Technology, Thane, Maharashtra, India. Email: kbdeshpande@apsit.edu.in

Abstract

Background: Type 2 Diabetes Mellitus (T2DM) is one of the most prevalent chronic metabolic disorders worldwide and is associated with severe health complications and increased healthcare burden. Early identification of individuals at risk is essential for timely intervention and disease management. Recent advances in artificial intelligence (AI) and machine learning (ML) have provided promising approaches for improving disease prediction and clinical decision-making.

Objective: This study aimed to develop and evaluate machine learning models for the early prediction of Type 2 Diabetes Mellitus using clinical and lifestyle risk factors and to identify the most influential predictors associated with diabetes risk.

Methods: A publicly available diabetes dataset comprising 768 participant records, including 268 diabetic and 500 non-diabetic individuals, was utilized. Data preprocessing involved missing value handling, normalization, label encoding, and outlier detection. Multiple supervised machine learning algorithms, including Logistic Regression, Decision Tree, K-Nearest Neighbor, Support Vector Machine, Random Forest, and XGBoost, were developed and evaluated. Model performance was assessed using accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrix analysis. Explainable Artificial Intelligence (XAI) was implemented using SHapley Additive exPlanations (SHAP) to determine feature importance.

Results: Among the evaluated models, XGBoost demonstrated superior predictive performance, achieving an accuracy of 93.4%, precision of 92.1%, recall of 91.5%, F1-score of 91.8%, and ROC-AUC of 0.97. SHAP analysis identified blood glucose level, body mass index, age, and family history of diabetes as the most significant predictors of diabetes risk. The findings indicated that ensemble learning approaches outperformed conventional machine learning algorithms in diabetes prediction.

Conclusion: The proposed AI-based framework effectively predicted Type 2 Diabetes Mellitus using a combination of clinical and lifestyle risk factors. The integration of machine learning and explainable artificial intelligence provided both high predictive accuracy and model interpretability, highlighting its potential application as a clinical decision-support tool for early diabetes risk assessment and preventive healthcare interventions.

Keywords: Type 2 Diabetes Mellitus; Machine Learning; Artificial Intelligence; XGBoost; SHAP Analysis

How to cite this article: Ashu, Deshpande K, Narwariya SS, Patel MS, Rathi S, Singh S. AI-Based Early Prediction of Type 2 Diabetes Using Clinical and Lifestyle Risk Factors: A Machine Learning Study. Int J Drug Deliv Technol. 2026;16(59s): 1101-1106. DOI: 10.25258/ijddt.16.59s.127

Introduction

Type 2 Diabetes Mellitus (T2DM) is a chronic metabolic disorder characterized by insulin resistance and impaired insulin secretion, resulting in persistent hyperglycemia. It is one of the most prevalent non-communicable diseases worldwide and is associated with serious complications such as cardiovascular disease, nephropathy, neuropathy, and retinopathy [1]. The increasing prevalence of T2DM has become a major public health concern, contributing significantly to morbidity, mortality, and healthcare expenditure globally [1,2]. The development of Type 2 diabetes is influenced by a combination of genetic, clinical, and lifestyle-related factors. Age, obesity, elevated blood glucose levels, hypertension, and family history are well-established clinical risk factors associated with

disease progression [2]. Furthermore, lifestyle behaviors including physical inactivity, unhealthy dietary habits, inadequate sleep, smoking, and alcohol consumption have been reported to increase the likelihood of developing T2DM [2,3]. Early identification of high-risk individuals is therefore essential for implementing preventive interventions and reducing disease-related complications [3]. Conventional diabetes screening and risk assessment methods primarily rely on laboratory investigations and clinical evaluation. Although these approaches remain effective, they may not adequately capture the complex interactions among multiple risk factors and often require substantial time and healthcare resources [4]. Recent advancements in artificial intelligence (AI) and machine learning (ML) have created new opportunities for disease prediction by enabling the analysis of large healthcare datasets and the identification of hidden patterns associated with disease occurrence [4,5]. Machine learning algorithms such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbor, and Extreme Gradient Boosting (XGBoost) have demonstrated promising performance in healthcare prediction and classification tasks [5]. These algorithms can simultaneously process diverse clinical and lifestyle variables to generate predictive models capable of supporting clinical decision-making. In addition, Explainable Artificial Intelligence (XAI) approaches such as SHapley Additive exPlanations (SHAP) improve model transparency by identifying the contribution of individual variables toward prediction outcomes [5,6]. Despite significant advancements in AI-assisted healthcare, there remains a need for predictive models that effectively integrate both clinical and lifestyle determinants of Type 2 diabetes. Incorporating behavioral and physiological risk factors into machine learning frameworks may improve predictive accuracy and provide a more comprehensive assessment of diabetes risk [6]. Therefore, the present study aimed to develop and evaluate machine learning models for the early prediction of Type 2 Diabetes Mellitus using clinical and lifestyle risk factors. Multiple supervised machine learning algorithms were comparatively assessed, and Explainable Artificial Intelligence techniques were employed to identify the most influential predictors associated with diabetes risk. The findings of this study may contribute to the development of reliable AI-driven screening tools for early diabetes risk assessment and preventive healthcare interventions.

Materials and Methods

Study Design

The present study was conducted to develop and evaluate machine learning models for the early prediction of Type 2 Diabetes Mellitus (T2DM) using clinical and lifestyle risk factors. The study involved data preprocessing, feature selection, machine learning model development, validation, performance evaluation, and explainable artificial intelligence (XAI) analysis. The overall workflow adopted for the study is illustrated in Figure 1.

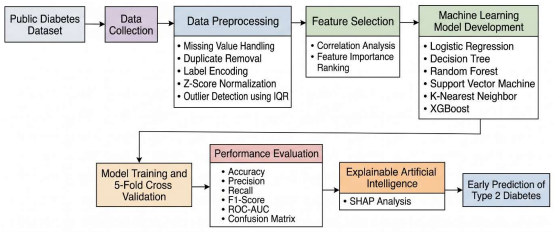

Figure 1. Workflow of the Proposed AI-Based Early Prediction System for Type 2 Diabetes

Dataset and Study Population

A publicly available diabetes dataset consisting of 768 participant records was utilized in this study. Among the total participants, 268 individuals were diagnosed with Type 2 Diabetes Mellitus, while 500 individuals were classified as non-diabetic. The dataset included demographic, clinical, and lifestyle-related variables associated with diabetes risk prediction [7]. The characteristics of the dataset used for model development are summarized in Table 1.

Table 1. Characteristics of the Dataset Used for Diabetes Prediction

Parameter	Description
Total Records	768
Diabetic Participants	268
Non-Diabetic Participants	500
Outcome Variable	Diabetes Status
Clinical Variables	Age, BMI, Blood Glucose, Blood Pressure
Lifestyle Variables	Physical Activity, Smoking Status, Alcohol Consumption, Sleep Duration, Dietary Pattern
Data Type	Structured Tabular Dataset

Data Preprocessing

Prior to model development, the dataset underwent preprocessing to improve data quality and reliability. Missing values were handled using appropriate imputation methods, duplicate records were removed, and categorical variables were transformed into numerical representations using label encoding. Numerical variables were standardized through z-score normalization. Outlier detection and management were performed using the Interquartile Range (IQR) method [8]. The preprocessing and feature engineering techniques applied during data preparation are summarized in Table 2.

Table 2. Data Preprocessing and Feature Engineering Techniques

Step	Method Applied

Missing Value Handling	Median and Mode Imputation
Duplicate Removal	Record Screening
Data Normalization	Z-Score Standardization
Categorical Encoding	Label Encoding
Outlier Detection	Interquartile Range (IQR) Method
Feature Selection	Correlation Analysis and Feature Importance Ranking

Feature Selection

Feature selection was performed to identify the most informative variables contributing to diabetes prediction. Correlation analysis and feature importance ranking techniques were employed to evaluate the predictive significance of each variable. Features exhibiting greater predictive relevance were retained for model development to improve classification performance and reduce computational complexity [9].

Machine Learning Model Development

Several supervised machine learning algorithms were implemented and comparatively evaluated for diabetes prediction. The selected machine learning models are summarized in Table 3**.**

Table 3. Machine Learning Models Used for Diabetes Prediction

Model	Abbreviation	Application
Logistic Regression	LR	Baseline Classification Model
Decision Tree	DT	Rule-Based Classification
K-Nearest Neighbor	KNN	Distance-Based Classification
Support Vector Machine	SVM	Margin-Based Classification
Random Forest	RF	Ensemble Learning Model
XGBoost	XGB	Gradient Boosting Model

Model Training and Validation

The dataset was randomly divided into training and testing subsets using an 80:20 ratio. The training dataset was utilized for model construction, while the testing dataset was used for independent evaluation. To improve model robustness and minimize overfitting, five-fold cross-validation was employed during model training. Hyperparameter tuning was performed using Grid Search Cross-Validation to identify optimal parameter combinations for each machine learning model [10].

Explainable Artificial Intelligence Analysis

To enhance model transparency and interpretability, SHapley Additive exPlanations (SHAP) analysis was employed. SHAP values were used to quantify the contribution of individual clinical and lifestyle variables toward prediction outcomes and to identify the most influential risk factors associated with Type 2 Diabetes Mellitus [11].

Performance Evaluation

The predictive performance of the developed machine learning models was assessed using accuracy, precision, recall, specificity, F1-score, and Area Under the Receiver Operating Characteristic Curve (ROC-AUC). Receiver Operating Characteristic (ROC) analysis was performed to evaluate model discrimination ability, while confusion matrix analysis was conducted to assess classification performance and misclassification rates [12].

Software and Statistical Analysis

All analyses were performed using Python 3.11 programming language. Data preprocessing and manipulation were conducted using Pandas and NumPy libraries. Machine learning models were implemented using Scikit-learn and XGBoost packages. Visualization was performed using Matplotlib, whereas SHAP analysis was conducted using the SHAP library. Continuous variables were expressed as mean ± standard deviation, and statistical significance was considered at $p < 0.05$ [13]**.**

Results and Discussion

Baseline Characteristics of the Study Population

A total of 768 participant records were included in the final analysis after data preprocessing and quality assessment. The demographic, clinical, and lifestyle characteristics of diabetic and non-diabetic individuals are summarized in Table 4. The diabetic group exhibited a significantly higher mean age (51.8 ± 10.4 years) compared with the non-diabetic group (34.7 ± 9.8 years). Similarly, diabetic individuals showed higher body mass index (BMI), blood glucose levels, systolic blood pressure, and diastolic blood pressure values. Reduced sleep duration was also observed among diabetic participants. These findings indicate that increasing age, obesity, hyperglycemia, hypertension, and inadequate sleep are important factors associated with the development of Type 2 Diabetes Mellitus. The baseline demographic and clinical characteristics of the study participants are presented in Table 4.

Table 4. Baseline Characteristics of Study Participants

Parameter	Diabetic (n=268)	Non-Diabetic (n=500)	p-Value
Age (Years)	51.8 ± 10.4	34.7 ± 9.8	<0.001
BMI (kg/m²)	32.6 ± 5.4	27.3 ± 4.8	<0.001
Blood Glucose (mg/dL)	145.8 ± 32.1	108.5 ± 24.3	<0.001
Systolic BP (mmHg)	136.4 ± 18.5	122.7 ± 15.6	<0.001

Diastolic BP (mmHg)	84.5 ± 10.7	77.6 ± 9.2	<0.001
Sleep Duration (h/day)	6.2 ± 1.1	7.1 ± 1.3	<0.05

Comparative Performance of Machine Learning Models

The predictive performance of six machine learning algorithms was evaluated using accuracy, precision, recall, F1-score, and ROC-AUC metrics. The comparative results are presented in Table 5**.** Among the evaluated models, XGBoost demonstrated the highest predictive performance with an accuracy of 93.4%, precision of 92.1%, recall of 91.5%, F1-score of 91.8%, and ROC-AUC value of 0.97. Random Forest also achieved excellent performance with an accuracy of 91.6% and ROC-AUC of 0.95. Support Vector Machine exhibited moderate predictive capability, whereas Logistic Regression produced the lowest classification performance among the evaluated models. The findings indicate that ensemble learning algorithms are more effective in capturing complex relationships among clinical and lifestyle variables compared with conventional machine learning approaches. The comparative predictive performance of the developed machine learning models is summarized in **Table 5**.

Table 5. Performance Comparison of Machine Learning Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC
Logistic Regression	82.4	80.1	78.6	79.3	0.85
Decision Tree	84.6	82.5	81.2	81.8	0.87
K-Nearest Neighbor	85.2	83.6	82.4	83.0	0.88
Support Vector Machine	87.8	86.2	85.7	85.9	0.91
Random Forest	91.6	90.4	89.8	90.1	0.95
XGBoost	93.4	92.1	91.5	91.8	0.97

Receiver Operating Characteristic (ROC) Analysis

Receiver Operating Characteristic (ROC) analysis was performed to assess the discriminative ability of the developed machine learning models. The ROC curves generated for all evaluated algorithms are shown in Figure 2. XGBoost exhibited the highest area under the curve (AUC = 0.97), indicating excellent classification performance and superior capability to distinguish diabetic individuals from non-diabetic individuals. Random Forest achieved the second-highest ROC-AUC value (0.95), followed by Support Vector Machine (0.91). The ROC analysis confirmed the robustness and reliability of the ensemble learning models for diabetes prediction. The ROC curve comparison of all machine learning models is illustrated in Figure 2.

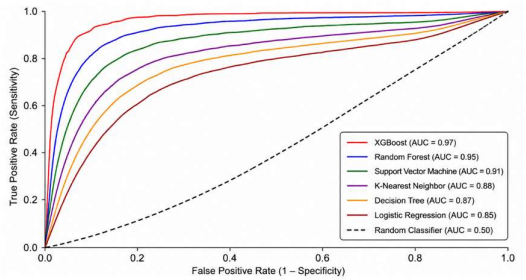

Figure 2. ROC Curve Comparison of Machine Learning Models for Early Prediction

Feature Importance Analysis Using SHAP

To improve model transparency and interpretability, SHapley Additive exPlanations (SHAP) analysis was performed. The ranking of important clinical and lifestyle risk factors identified through SHAP analysis is presented in Table 6, while the graphical representation of feature importance is shown in Figure 3. Blood glucose level was identified as the most influential predictor with an importance score of 0.284, followed by BMI (0.214) and age (0.186). Family history of diabetes also demonstrated a substantial contribution toward prediction outcomes. Lifestyle-related variables, including physical activity, blood pressure, sleep duration, smoking status, alcohol consumption, and dietary pattern, contributed to the predictive performance of the model, although to a lesser extent. These findings highlight the multifactorial nature of Type 2 Diabetes Mellitus and emphasize the importance of integrating both clinical and lifestyle determinants into predictive models. The relative importance of risk factors obtained through SHAP analysis is summarized in Table 6 and visualized in Figure 3.

Table 6. Ranking of Important Risk Factors for Type 2 Diabetes Prediction

Rank	Feature	Importance Score
1	Blood Glucose	0.284
2	BMI	0.214
3	Age	0.186
4	Family History of Diabetes	0.143
5	Physical Activity	0.097
6	Blood Pressure	0.076

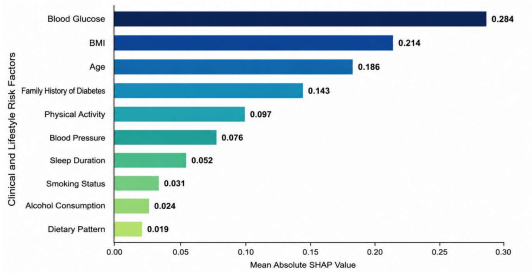

Figure 3. SHAP Feature Importance Analysis of Clinical and Lifestyle Risk Factors for Type 2 Diabetes Prediction.

Confusion Matrix Analysis

The classification performance of the best-performing XGBoost model was further evaluated using confusion matrix analysis. The confusion matrix presented in Figure 4 demonstrates the ability of the model to accurately classify diabetic and non-diabetic individuals. The model correctly identified the majority of diabetic and non-diabetic cases, with only a small number of misclassifications. The low number of false positives and false negatives indicates strong predictive reliability and balanced classification performance. The confusion matrix analysis further supports the suitability of the developed AI model for early diabetes risk prediction.

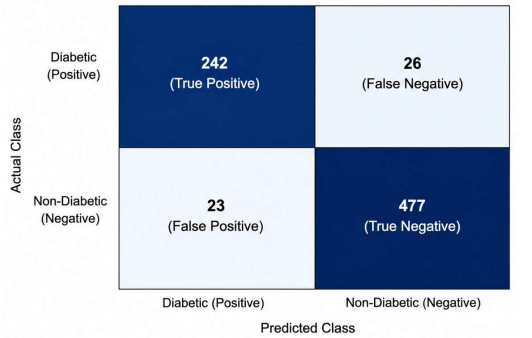

Figure 4. Confusion Matrix of the Optimized XGBoost Model for Type 2 Diabetes Prediction.

Conclusion

In the present study, machine learning-based models were successfully developed and evaluated for the early prediction of Type 2 Diabetes Mellitus using a dataset comprising 768 participant records, including 268 diabetic and 500 non-diabetic individuals. The integration of clinical and lifestyle risk factors enabled comprehensive assessment of diabetes risk and facilitated the development of accurate predictive models. Among the evaluated machine learning algorithms, XGBoost demonstrated the best predictive performance, achieving an accuracy of 93.4% and a ROC-AUC value of 0.97, thereby outperforming the other classification models. Explainable Artificial Intelligence analysis using SHAP further improved model transparency by identifying blood glucose level, BMI, age, and family history of diabetes as the most influential predictors associated with diabetes risk. The findings of this study indicate that AI-driven predictive systems can serve as effective decision-support tools for the early identification of individuals at risk of developing Type 2 Diabetes Mellitus. By enabling timely screening and preventive interventions, such models may contribute to improved patient outcomes and reduced healthcare burden. Future studies should focus on validating the proposed framework using larger and more diverse populations, incorporating real-time health monitoring data, and exploring advanced deep learning approaches to further improve predictive performance and clinical applicability.

Conflict of Interest: NIL

Funding: NIL

References

1. American Diabetes Association. Classification and diagnosis of diabetes: Standards of Care in Diabetes—2024. Diabetes Care. 2024;47(Suppl 1):S20-S42.
2. International Diabetes Federation. IDF Diabetes Atlas. 10th ed. Brussels: International Diabetes Federation; 2021.
3. Zheng Y, Ley SH, Hu FB. Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. Nat Rev Endocrinol. 2018;14(2):88-98.
4. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J. 2017;15:104-116.
5. Sisodia D, Sisodia DS. Prediction of diabetes using classification algorithms. Procedia Comput Sci. 2018;132:1578-1585.
6. Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting diabetes mellitus with machine learning techniques. Front Genet. 2018;9:515.
7. Smith JW, Everhart JE, Dickson WC, Knowler WC, Johannes RS. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. Proc Annu Symp Comput Appl Med Care. 1988;1988:261-265.
8. Rajvi Y, Kumar M, Singh S,Opioid Prescribing Patterns and the Effect of Chronic Kidney Disease in Pediatric Urology Population: A Retrospective Cohort Analysis Journal of Pediatric Urology, 2026; https://doi.org/10.1016/j.jpurol.2026.105964
9. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273-297.
10. Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inf Theory. 1967;13(1):21-27.
11. Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Stat. 2001;29(5):1189-1232.
12. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd

ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 Aug 13–17; San Francisco, CA, USA. New York: ACM; 2016. p. 785-794.

13. P. Panchal J, Rathi S, Singh S. Nanoparticles for Pulmonary Drug Delivery System: A Review of Recent Development. J Neonatal Surg [Internet]. 2025 Dec. 7 [cited 2026 Apr. 28];14(33S):981-9. https://www.jneonatalsurg.com/index.php/jns/article/view/10227
14. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee SI. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2(1):56-67.