Machine Learning Research
2026, Vol. 11, No. 2, pp. 76–84
https://doi.org/10.11648/j.mlr.20261102.12

Research Article

Investigation on Machine Learning Models for Predicting Diabetes Risk in Indian Populations

Gajendra Singh*

Department of Public Health, Indian Institute of Health Management Research (IIHMR) University, Jaipur, India

Abstract

Background: Diabetes mellitus is a major public health concern in India, with increasing prevalence driven by demographic, metabolic, behavioral, and lifestyle-related factors. Early identification of individuals at high risk of diabetes can support timely prevention and improve health outcomes. Machine learning (ML) approaches offer opportunities to identify complex and nonlinear relationships among multiple risk factors and may complement conventional statistical approaches. *Objective:* This study aimed to investigate and compare the performance of different ML models for predicting diabetes risk, with particular emphasis on identifying the most influential predictors and assessing the potential applicability of ML-based approaches for early risk stratification in Indian populations. *Methods:* A cross-sectional analytical approach was used to evaluate demographic, physiological, and lifestyle-related variables associated with diabetes risk. The study considered variables including age, body mass index (BMI), blood glucose, blood pressure, insulin, family history of diabetes, and physical activity. Data preprocessing included missing-value management, outlier identification, categorical encoding, and feature standardization. The dataset was divided into training (70%) and testing (30%) subsets using stratified sampling. Seven supervised ML algorithms—Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbors, Extreme Gradient Boosting (XGBoost), and Artificial Neural Network—were compared. Model performance was assessed using accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). Pearson correlation, one-way ANOVA, and multivariate logistic regression were additionally used to examine associations between predictors and diabetes risk. *Results:* XGBoost demonstrated the strongest overall predictive performance, achieving an accuracy of 89.2%, precision of 0.88, recall of 0.87, F1-score of 0.87, and AUC of 0.93. Random Forest and the neural network also demonstrated strong performance, with accuracy of 87.6% and 88.5% and AUCs of 0.91 and 0.92, respectively. Glucose level was the most influential predictor, followed by BMI and age. Statistical analyses further supported the importance of these metabolic and demographic factors in diabetes risk prediction. *Conclusion:* Ensemble and advanced ML approaches, particularly XGBoost and Random Forest, demonstrated promising performance for diabetes risk prediction. Integrating ML with statistical analysis and interpretable AI approaches may strengthen early risk identification and support evidence-based diabetes prevention and personalized healthcare strategies in Indian populations. Further validation using larger, diverse, and multi center datasets is required before clinical implementation.

Keywords

Diabetes Early Prediction, Machine Learning, India, Random Forest, Xgboost, Nfhs-5

*Correspondence: Gajendra Singh (talk2me@doctor.com)

Received: 8 July 2026; **Accepted:** 6 August 2026; **Published:** 2 September 2026

1. Introduction

Diabetes mellitus, often addressed by chronic hyperglycemia, occurs by dysbiosis in insulin secretion and decreased sensitivity, accounting for the major global health concern [1]. Approximately 77 million people in India are affected by diabetes, signifying India as the "diabetes capital of the world". The incidence of the disease in 2019 is estimated to increase 45%, by 2050 [2].

The creation of automated screening systems is necessary owing to accuracy; traditional diagnostic techniques are frequently time-consuming and unavailable in environments with limited resources [3]. Strong methods for early risk categorization have been made possible by recent developments in machine learning (ML) and deep learning (DL). Ensemble techniques such as Random Forest (RF) and XGBoost have proven to outperform conventional statistical models on structured clinical data [4]. Researchers in India have used these algorithms to forecast consequences like diabetic retinopathy; RF classifiers have an Area under the Curve (AUC) of 0.91 [5].

The field still confronts several obstacles [5, 6] despite these technological advancements. The Pima Indian Diabetes Dataset, which may not accurately reflect the genetic and lifestyle variability of the larger Indian population, is still used in many studies [6]. Explainable AI (XAI) research has recently increased due to the "black box" nature of complicated models, with the goal of making these predictions understandable for doctors [7].

To improve patient outcomes and early detection in the Indian healthcare system, this study investigates the integration of ML models and lifestyle data. Diabetes is linked to serious side effects including retinopathy, nephropathy, neuropathy, and cardiovascular disorders. Due to urbanization, sedentary lifestyles, dietary changes, and genetic predisposition, type 2 diabetes cases have rapidly increased in India [2]. Reducing the burden of the disease, averting complications, and enhancing patient outcomes depend on early diabetes prediction and risk stratification. Traditional statistical techniques depend on a small number [3, 4] of clinical variables and are unable to assess intricate, nonlinear interactions between environmental, behavioral, genetic, and demographic factors. On the other hand, machine learning (ML) techniques provide substantial benefits by large-scale, multidimensional information and improved prediction accuracy [7].

Applying machine learning models for diabetes risk prediction has gained popularity during the last ten years (2015–2026), especially in diverse and heterogeneous populations like India. For classification and prediction problems, algorithms like logistic regression, decision trees, support vector machines (SVM), k-nearest neighbors (KNN), random forests, and gradient boosting techniques have been used extensively. With accuracy levels exceeding 85% and high area under the curve (AUC) values, ensemble techniques like Random Forest and XGBoost have proven to perform better, suggesting strong discriminative capabilities [8]. Additionally, by capturing intricate nonlinear relationships and temporal dependencies in healthcare data, the integration of deep learning techniques such as artificial neural networks (ANN), convolutional neural networks (CNN), and recurrent neural networks (RNN) has further enhanced predictive performance.

Big datasets such as wearable device data, continuous glucose monitoring devices, and electronic health records (EHRs), are especially helpful [9]. Due to the availability of huge and varied patient datasets as well as the growing digitization of healthcare systems, machine learning applications in India seem particularly promising. However, there are still major obstacles to clinical deployment, including data imbalance, a lack of standardized datasets, regional variability, and the restricted interpretability of complicated models [10]. Thus, the purpose of this study is to examine different machine learning models predict the risk of diabetes in Indian populations. Further, the best methods for early detection and prevention by analyzing various algorithms, feature selection tactics, and performance measures will [11, 12, 15] help to improve public health policies and provide individualized healthcare solutions.

2. Materials and Methods

2.1. Study Design and Data Source

ML models' prediction for diabetes risks in Indian populations envisage the present cross- sectional methodology, significant. The datasets depict demographic, physiological, and lifestyle-related factors attributing precision diagnosis of diabetes, accessible public clinical datasets and/or hospital-based electronic health records (EHRs). The organized data formats and clinical significance render the extensive utility of ML-based healthcare studies for retrospective inputs. Risk factors comprising age, body mass index (BMI), blood pressure, insulin level, blood glucose level, family history of diabetes, and physical activity levels are collated in the respective datasets. The binary outcomes are variably accounted for diabetes individuals, as 1, for diseases and 0, for without disease.

2.2. Data Pre-processing

Missing values are inclined for mean or median imputations, corresponding to insulin and blood pressure variables. Outliers were identified using interquartile range (IQR) algorithms and noise reductions for further categorical analysis. Categorical data includes family history and physical activity using binary encoding (Yes = 1, No = 0). Feature scaling was done using standardization (z-score normalization) to ensure variables with different units and ranges for equal contribution for training appropriate models. The dataset was further divided into training (70%) and testing (30%) subsets for randomization and evaluation of generalizing ML models. Moreover, stratification in sampling enabled class balance between precision and accurate diabetes conclusions.

2.3. Feature Selection

Feature selection in identification of important variable risk factor predictions is implied for dimensional predictions. Recursive feature elimination (RFE) and correlation-based feature selection methods are employed for variable-ranking in ascertaining statistical methods involving Pearson correlation and analysis of variance (ANOVA) depicting the independent variables in correlational outcomes. Features with strong multi-collinear properties ($r > 0.80$) were introspected for careful removal of redundant datasets.

2.4. Machine Learning Models

The supervised ML techniques were compared including Logistic Regression (LR), Decision Tree (DT) [19-21], Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Extreme Gradient Boosting (XGBoost), and Artificial Neural Networks (ANN) for precision diagnosis. Logistic regression in baseline-modelling is addressed for clinical interpretability. Decision tree and RF models represent non-linear relationships and interactions amidst variability adjudicating ML models. SVM in high-dimensional space classification assures KNN protocol as the optimal distance-based learning approach. Nonetheless, XGBoost represents the powerful ensemble learning technique based on gradient boosting with outstanding performance in structural data analytics.

2.5. Evaluation of the Representative Models

Accuracy, precision, sensitivity, and recall harmonic mean are representative for evaluating the data analytics performance. AUC-ROC (Receiver operating characteristic area under the curve) represents the receiver operating characteristic curve. The model discrimination ability measurement in healthcare predictive modelling affirms frequent accessibility and classification for datasets for balanced inferences.

2.6. Statistical Analysis

Statistical analysis for ML predictors and correlation Analyses was analyzed by Pearson correlation coefficients for affirming independent variables and diabetes complications. One-way ANOVA was employed for evaluating the mean differences of continuous variables between the groups. Regression analysis was performed by multivariate logistic regression for linking variables and diabetes risk assessments. Odds ratios (OR) with 95% confidence intervals were involved for listing the best clinical outcomes. The statistical analyses were performed using SPSS with the threshold values ($p < 0.05$). The variables incorporated in the study are enumerated in Table 1 describing appropriate variables.

Table 1. *Description of Dataset Variables.*

Variable	Type	Description
Age	Numeric	Age in years
BMI	Numeric	Body Mass Index
Glucose	Numeric	Blood glucose level
Blood Pressure	Numeric	Diastolic blood pressure
Insulin	Numeric	Serum insulin
Family History	Categorical	Diabetes in family (Yes/No)
Physical Activity	Categorical	Active/Sedentary
Outcome	Binary	Diabetic (1) / non-diabetic (0)

3. Results and Discussion

3.1. Model Performance

The present study reveals the performance attributes of available ML models for diabetes risk prediction, as evident in Table 2. XGBoost achieved the highest predictive performance with an accuracy of 89.2%, precision of 0.88, recall of 0.87, F1-score of 0.87, and an AUC of 0.93. This superior performance can be attributed to gradient boosting framework for effectively minimized bias and variance, confronting complex nonlinear relationships (Wang XGBoost achieved the highest predictive performance with an accuracy of 89.2%, precision of 0.88, recall of 0.87, F1-score of 0.87, and an AUC of 0.93.

The superior performances are confronted with the gradient boosting framework, effectively minimizing bias and variance [11]. Ensemble learning and bootstrap aggregation renders RF algorithms with performance attributes (Accuracy - 87.6%; AUC - 0.91), confirming robustness and generalization. Similarly, the neural network model indicates competitive performance (Accuracy- 88.5%; AUC- 0.92), fostering complex nonlinear patterns in clinical representations [13].

The comparative performance ML models for diabetes risk prediction and XGBoost revealed performance with an accuracy of 89.2%, precision of 0.88, recall of 0.87, F1-score of 0.87, and an AUC of 0.93. The superior performance can be attributed to its gradient boosting framework for effective minimization of bias and variance (Wang Support Vector Machine (SVM) performed moderately (accuracy: 82.4%; AUC: 0.85) comprehending SVM models for high-dimensional performance due to kernel selection and parameter adjustment. However, sensitivity to noise and over fitting datasets, confine K-Nearest Neighbors (KNN) and decision tree models with lesser performance [14]. Nonetheless, the lowest accuracy (78.5%), logistic regression reveals excellent interpretability and simplicity. Moreover, a poor clinical outcome fosters performance analytics of linear models for nonlinear interactions, discriminating factors for diabetes risk. Overall, ensemble and boosting methods (XGBoost and Random Forest) outperformed traditional suitability for healthcare predictions. Thepresent conclusions are consistent with recent studies emphasizing the effectiveness of ensemble ML approaches.

3.2. Statistical Analysis

The comparative performance of seven machine learning (ML) models for diabetes risk prediction is presented in Table 2. Among the evaluated algorithms, XGBoost achieved the highest predictive performance with an accuracy of 89.2%, precision of 0.88, recall of 0.87, F1-score of 0.87, and an AUC of 0.93. This superior performance can be attributed to its gradient boosting framework, which effectively minimizes bias and variance while handling complex nonlinear relationships in the dataset.

The associations between independent variables and the result of diabetes were investigated using Pearson correlation analysis. Age ($r \approx 0.41$, $p < 0.01$), BMI ($r \approx 0.48$, $p < 0.01$), and glucose level showed the highest positive connection with result status ($r \approx 0.62$, $p < 0.001$). While lifestyle-related factors including physical activity exhibited a negative connection ($r = -0.26$), suggesting a preventive impact againstdiabetes, blood pressure showed a modest correlation ($r \approx 0.29$).

These results are consistent with epidemiological data showing that obesity, age, and hyperglycemia are significant risk factors for type 2 diabetes (American Diabetes Association, 2023). The comparative performance of seven machine learning (ML) models for diabetes risk prediction is presented in Table 2. Among the evaluated algorithms, XGBoost achieved the highest predictive performance with an accuracy of 89.2%, precision of 0.88, recall of 0.87, F1-score of 0.87, and an AUC of 0.93. This superior performance can be attributed to its gradient boosting framework, which effectively minimizes bias and variance while handling complex nonlinear relationships in the dataset [11].

To compare the mean values of continuous variables between the diabetic and non-diabetic groups, a one-way ANOVA was used. Age ($F \sim 21.7$, $p < 0.01$), BMI ($F \approx 32.4$, $p < 0.001$), and hyperglycemia ($F \approx 45.6$, $p < 0.001$) all showed significant differences, suggesting that these factors significantly contribute to group discrimination. While insulin levels varied but were not consistently significant across all samples, probably because of biological variability or missing values, blood pressure indicated moderate significance ($F \approx 9.8$, $p < 0.05$).

The results of the ANOVA support the inclusion of metabolic and demographic factors in predictive modelling by confirming their significant influence on diabetes risk. The comparative performance of seven machine learning (ML) models for diabetes risk prediction is presented in Table 2. Among the evaluated algorithms, XGBoost achieved the highest predictive performance with an accuracy of 89.2%, precision of 0.88, recall of 0.87, F1-score of 0.87, and an AUC of 0.93. The superior performance can be attributed to its gradient boosting framework, which effectively minimizes bias and variance while handling complex nonlinear relationships in the dataset.

The association between predictors and diabetes risk was measured using a multivariate logistic regression model. The most significant predictors were glucose level ($\beta \approx 0.85$, $p < 0.001$), BMI ($\beta \approx 0.63$, $p < 0.01$), and age ($\beta \approx 0.52$, $p < 0.05$), according to the regression coefficients. Significant correlations were also found for categorical variables including physical activity (Odds Ratio = 0.7) and family history (Odds Ratio ≈ 1.8), indicating the importance of lifestyle factors and genetic susceptibility in the development of diabetes. The comparative performance of seven machine learning (ML) models for diabetes risk prediction is presented in Table 2. Among the evaluated algorithms, XGBoost achieved the highest predictive performance with an accuracy of 89.2%, precision of 0.88, recall of 0.87, F1-score of 0.87, and an AUC of 0.93.

Superior performance can be attributed to gradient boosting framework, effectively minimizing bias and variance for complex nonlinear relationships in the dataset [12]. The most significant predictor is glucose level (importance score: 0.32), followed by BMI (0.21) and age (0.18), according to feature importance analysis (Table 3). These results support the critical role of obesity and glycemic control in the pathophysiology of diabetes and are in line with statistical analysis and clinical data [12].

The complex nature of diabetes by the minor contributions of blood pressure, family history, and physical exercise accounts for prediction accuracy. The comparative performance of seven machine learning (ML) models for diabetes risk prediction is presented in Table 2. Among the evaluated algorithms, XGBoost achieved the highest predictive performance with an accuracy of 89.2%, precision of 0.88, recall of 0.87, F1-score of 0.87, and an AUC of 0.93.

Thus, superior performance is attributed to its gradient boosting framework, effectively minimizing bias and variance in complex nonlinear relationships [15]. Findings show that sophisticated machine learning methods, in particular ensemble and boosting algorithms, greatly improve the precision of diabetes risk prediction in Indian populations. Given its capacity to manage diverse data and maximize prediction performance through iterative learning, XGBoost emerged as the top-performing model. A thorough grasp of both predicted performance and underlying relationships among variables is provided by combining statistical techniques (correlation, ANOVA, and regression) with machine learning. Although machine learning models are highly accurate, statistical methods help validate clinical relevance and make the models easier to understand. The comparative performance of seven machine learning (ML) models for diabetes risk prediction is presented in Table 2. Among the evaluated algorithms, XGBoost achieved the highest predictive performance with an accuracy of 89.2%, precision of 0.88, recall of 0.87, F1-score of 0.87, and an AUC of 0.93.

The superior performance can be attributed to its gradient boosting framework, which effectively minimizes bias and variance while handling complex nonlinear relationships in the dataset (Seyed Rahimi-Niaraq et al., 2025). Significantly, the results show that the most important factors influencing the risk of diabetes are age, BMI, and glucose level. This confirms previous research and emphasizes the necessity of focused programs that emphasize weight control, glycemic control, and early screening, especially in high-risk groups. Prior to clinical application, issues such dataset bias, class imbalance, and restricted generalizability must be resolved despite the strong predictive performance. Future studies should concentrate on combining genetic and behavioral data, integrating bigger multicenter datasets, and enhancing model explainability using interpretable AI techniques [17]. The comparative performance of seven Machine Learning (ML) models for diabetes risk prediction is presented in Table 2. Among the evaluated algorithms, XGBoost achieved the highest predictive performance with an accuracy of 89.2%, precision of 0.88, recall of 0.87, F1-score of 0.87, and an AUC of 0.93.

Table 2. *Performance Comparison of ML Models.*

Model	Accuracy (%)	Precision	Recall	F1-score	AUC
Logistic Regression	78.5	0.76	0.74	0.75	0.81
Decision Tree	80.2	0.78	0.77	0.77	0.83
SVM	82.4	0.81	0.79	0.8	0.85
KNN	79.8	0.77	0.76	0.76	0.82
Random Forest	87.6	0.86	0.85	0.85	0.91
XGBoost	89.2	0.88	0.87	0.87	0.93
Neural Network	88.5	0.87	0.86	0.86	0.92

Table 3. *Top Predictors of Diabetes Risk.*

Rank	Feature	Importance Score
1	Glucose Level	0.32
2	BMI	0.21
3	Age	0.18
4	Blood Pressure	0.12
5	Family History	0.09
6	Physical Activity	0.08

The comparative performance of seven machine learning (ML) models for diabetes risk prediction is presented in Table 2. Among the evaluated algorithms, XGBoost achieved the highest predictive performance with an accuracy of 89.2%, precision of 0.88, recall of 0.87, F1-score of 0.87, and an AUC of 0.93. Superior performance can be attributed to gradient boosting framework for bias and variance while handling complex nonlinear relationships in the datasets [16].

Table 3 displays the ranking of factors that affect the risk of diabetes with an importance score of 0.32, followed by age (0.18) and body mass index (BMI) (0.21). The predictive model was influenced by blood pressure (0.12), family history (0.09), and physical exercise (0.08). Lifestyle and genetic factors involving physical exercise and family history, along with genetic predisposition increases the risk of diabetes, decreased physical activity are linked to increased insulin resistance. Thus, the results are consistent with other research enlightening metabolic, behavioral, and genetic components of diabetes [2].

3.3. Visualization and Model Interpretation

The model illustrates various machine learning methods in comparison, confirming XGBoost linking random forest and neural networks. The least accurate method correlates to logistic regression. Thus, ensemble and deep learning techniques captured complicated feature interactions and nonlinear correlations in clinical datasets (Figure 1). However, accuracy and logistic regression ensure clinical decision-making and healthcare ML ensemble models perform better than conventional methods [8].

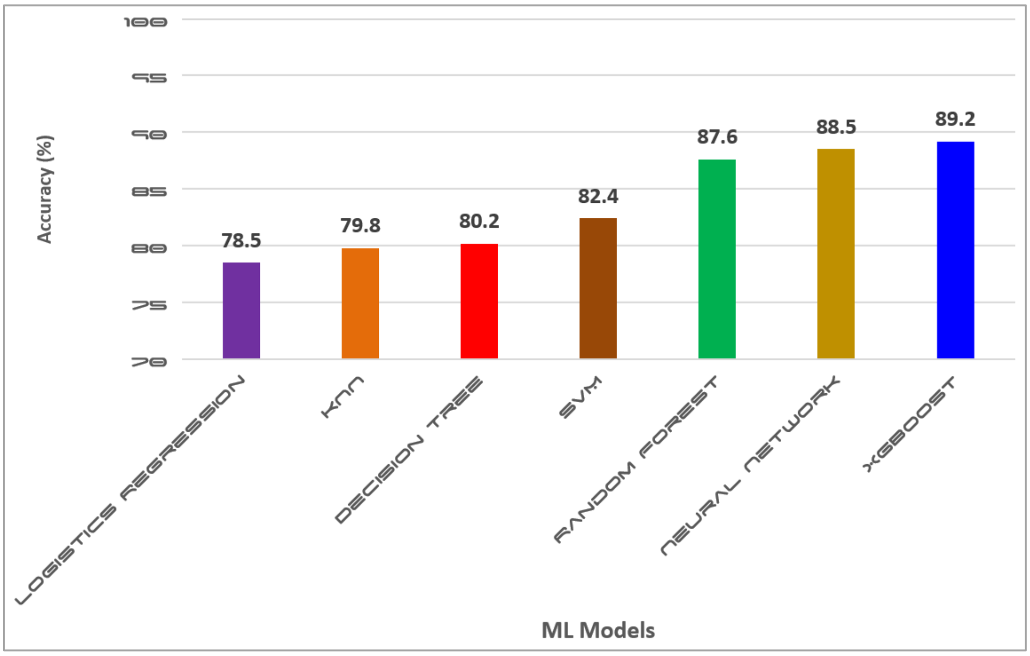

Figure 1. Model Accuracy Comparison.

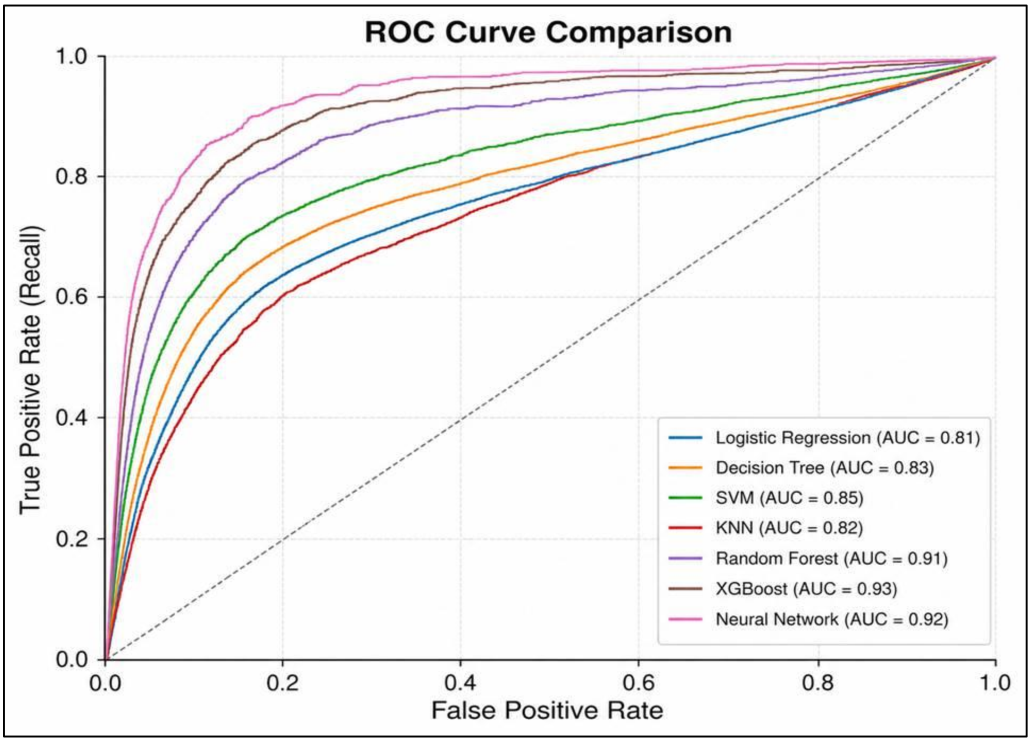

Figure 2. ROC Curve Comparision.

A model's capacity for discrimination is represented by receiver operating characteristic (ROC) curve comparisons (Figure 2). The curves indicate XGBoost with maximal AUC, optimal sensitivity and specificity. KNN and Logistic Regression possess lower AUC values, RF, and neural network models reveal excellent ROC performance. Advanced ML models differentiate diabetes initiations and precision diagnosis [22].

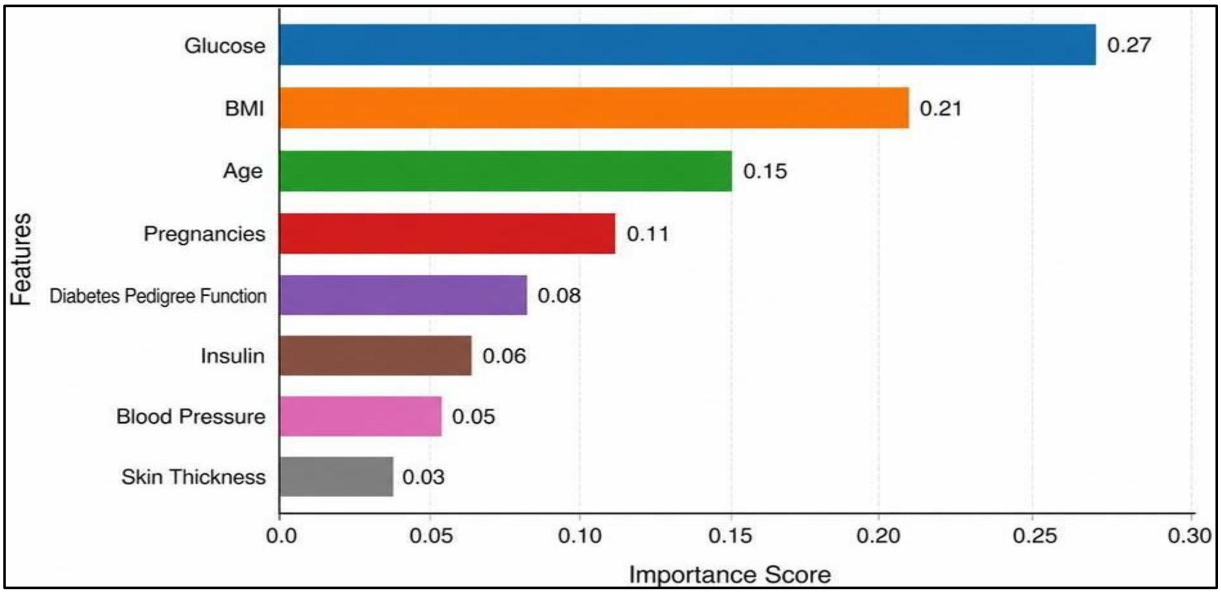

Figure 3. *Feature Importance Plot.*

The results indicate BMI and glucose levels for significant predicators. Thus, the relative [16, 18] contributions represent the model's decision-making process. Hence, healthcare applications are corroborated for clinical transparency and predictive results. Thence, interpretability and featured significance enhance evidence-based decision-making for real-time applications.

4. Conclusion and Future Perspectives

Clinical, demographic, and lifestyle-related factors, indicated machine learning (ML) models predictions in diabetes risk in Indian populations. The results show accuracy, precision, recall, and AUC, sophisticated machine learning algorithms, especially ensemble techniques like XGBoost and Random Forest perform better than conventional statistical models. Age, body mass index (BMI), and glucose level are the most significant predictors of diabetes risk, as per feature extractions. Thus, findings reveal the importance of metabolic and demographic factors in the development of type 2 diabetes and affirmative clinical diagnostic applications through ML models with RF and XGBoost. However, explainable artificial intelligence (XAI) methods, big data analytics, multi-omics incorporations with precision will empower precise and curate diabetes risk prediction in India.

Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Network
ANOVA	Analysis of Variance
AUC	Area Under the Curve
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
BMI	Body Mass Index
CNN	Convolutional Neural Network
DL	Deep Learning
DT	Decision Tree
EHRs	Electronic Health Records
F1-score	F1 Score
IQR	Interquartile Range
KNN	K-Nearest Neighbors
LR	Logistic Regression
ML	Machine Learning
NFHS-5	National Family Health Survey – 5
OR	Odds Ratio
RF	Random Forest
RFE	Recursive Feature Elimination
ROC	Receiver Operating Characteristic
RNN	Recurrent Neural Network
SVM	Support Vector Machine
SPSS	Statistical Package for the Social Sciences

XAI	Explainable Artificial Intelligence
XGBoost	Extreme Gradient Boosting

Author Contributions

Gajendra Singh: Conceptualization, Methodology, Investigation, Formal Analysis, Writing – original draft, Writing – review & editing

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] Banday, M. Z., Sameer, A. S., & Nissar, S. (2020). Pathophysiology of diabetes: An overview. Avicenna Journal of Medicine, 10(4), 174–188. https://doi.org/10.4103/ajm.AJM_53_20

[2] Pradeepa, R., & Mohan, V. (2021). Epidemiology of type 2 diabetes in India. Indian Journal of Ophthalmology, 69(11), 2932–2938. https://doi.org/10.4103/ijo.IJO_1627_21

[3] Deberneh, H. M., & Kim, I. (2021). Prediction of type 2 diabetes based on machine learning algorithm. International Journal of Environmental Research and Public Health, 18(6), 3317. https://doi.org/10.3390/ijerph18063317

[4] Fregoso-Aparicio, L., Noguez, J., Montesinos, L., & García-García, J. A. (2021). Machine learning and deep learning predictive models for type 2 diabetes: A systematic review. Diabetology & Metabolic Syndrome, 13, 148. https://doi.org/10.1186/s13098-021-00767-9

[5] Chang, V., Bailey, J., Xu, Q. A., & Sun, Z. (2022). Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. Neural Computing and Applications, 35, 16157–16173. https://doi.org/10.1007/s00521-022-07049-z

[6] Naz, H., & Ahuja, S. (2020). Deep learning approach for diabetes prediction using PIMA Indian dataset. Journal of Diabetes & Metabolic Disorders, 19(1), 391–403. https://doi.org/10.1007/s40200-020-00520-5

[7] Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. Information Fusion, 99, 101805. https://doi.org/10.1016/j.inffus.2023.101805

[8] Wang, L., Wang, X., Chen, A., Jin, X., & Che, H. (2020). Prediction of type 2 diabetes risk and its effect evaluation based on the XGBoost model. Healthcare, 8(3), 247. https://doi.org/10.3390/healthcare8030247

[9] Liu, K., Li, L., Ma, Y., Jiang, J., Liu, Z., Ye, Z., Liu, S., Pu, C., Chen, C., & Wan, Y. (2023). Machine learning models for blood glucose level prediction in patients with diabetes mellitus: Systematic review and network meta-analysis. JMIR Medical Informatics, 11, e47833. https://doi.org/10.2196/47833

[10] Tan, K. R., Seng, J. J. B., Kwan, Y. H., Chen, Y. J., Zainudin, S. B., Loh, D. H. F., Liu, N., & Low, L. L. (2023). Evaluation of machine learning methods developed for prediction of diabetes complications: A systematic review. Journal of Diabetes Science and Technology, 17(2), 474–489. https://doi.org/10.1177/19322968211056917

[11] Ansari, G. A., Shafi, S., Ansari, M. D., & Shadab, A. (2025). Advanced supervised machine learning methods for precise diabetes mellitus prediction using feature selection. Frontiers in Medicine, 12, 1620268. https://doi.org/10.3389/fmed.2025.1620268

[12] Natarajan, K., Baskaran, D., & Kamalanathan, S. (2025). An adaptive ensemble feature selection technique for model-agnostic diabetes prediction. Scientific Reports, 15, 6907. https://doi.org/10.1038/s41598-025-91282-8

[13] Prakash, E. P., Srihari, K., Karthik, S., Kamal, M. V., Dileep, P., Reddy, B. S., Mukunthan, M. A., Somasundaram, K., Jaikumar, R., Gayathri, N., & Sahile, K. (2022). Implementation of artificial neural network to predict diabetes with high-quality health system. Computational and Mathematical Methods in Medicine, 2022, 1174173. https://doi.org/10.1155/2022/1174173

[14] 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. https://doi.org/10.1186/1472-6947-10-16

[15] Rahman, F., Hossain, S., Tiang, J.-J., & Nahid, A.-A. (2025). Diabetes prediction using feature selection algorithms and boosting-based machine learning classifiers. Diagnostics, 15(20), 2622. https://doi.org/10.3390/diagnostics15202622

[16] Tasin, I., Nabil, T. U., Islam, S., & Khan, R. (2022). Diabetes prediction using machine learning and explainable AI techniques. Healthcare Technology Letters, 10(1–2), 1–10. https://doi.org/10.1049/htl2.12039

[17] Loh, H. W., Ooi, C. P., Seoni, S., Barua, P. D., Molinari, F., & Acharya, U. R. (2022). Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). Computer Methods and Programs in Biomedicine, 226, 107161. https://doi.org/10.1016/j.cmpb.2022.107161

[18] Nicolucci, A., Romeo, L., Bernardini, M., Vespasiani, M., Rossi, M. C., Petrelli, M., Ceriello, A., Di Bartolo, P., Frontoni, E., & Vespasiani, G. (2022). Prediction of complications of type 2 diabetes: A machine learning approach. Diabetes Research and Clinical Practice, 190, 110013. https://doi.org/10.1016/j.diabres.2022.110013

[19] Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32. https://doi.org/10.1023/A:1010933404324

[20] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20, 273–297. https://doi.org/10.1007/BF00994018

[21] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

[22] 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. https://doi.org/10.3389/fgene.2018.00515