International Journal for Research in Health Science and Nursing
(ISSN: 2208- 2670)
Vol. 10, Issue 02, June 2024
DOI: 10.65327/hsn.v10i3.2495

MACHINE LEARNING-BASED PREDICTION OF DIABETES RISK USING ROUTINE CLINICAL INDICATORS FOR EVIDENCE-BASED PRIMARY HEALTHCARE DECISION-MAKING

Dr. Leila Ismail[1*], Dr. Huned Materwala[2], Dr. Jalal Taneera[3], Dr. Musaab Sedig Ahmed Omer[4],

[1*]Department of Computer Science and Software Engineering, College of Information Technology, United Arab Emirates University, Al Ain, Abu Dhabi, United Arab Emirates, Email: leila@uaeu.ac.ae
[2]Department of Computer Science and Software Engineering, College of Information Technology, United Arab Emirates University, Al Ain, Abu Dhabi, United Arab Emirates, Email: huned.m@uaeu.ac.ae
[3]Department of Basic Medical Sciences, College of Medicine, University of Sharjah, Sharjah, United Arab Emirates, Sharjah Institute for Medical Research, University of Sharjah, Sharjah, United Arab Emirates, Email: jtaneera@sharjah.ac.ae
[4]College of Medicine, Ajman University, Ajman, United Arab Emirates, Email: m.omer@ajman.ac.ae

Received: 22/05/2024 **Revised:** 03/08/2024 **Acceptance:** 12/09/2024 **Published:** 25/09/2024

ABSTRACT-

Diabetes mellitus is a substantial public health challenge and there is a need for valid methods to help identify people at risk at an early stage, based on information that is routinely collected, such as clinical data. This study aimed to create and compare machine learning models predicting diabetes risk based on eight clinical parameters: age, body mass index, blood pressure, insulin, glucose, skin thickness, Diabetes Pedigree Function and pregnancy. The secondary dataset was made up of 1168 patient records, 397 of which were diabetic cases and 771 were non-diabetic cases. The glucose, blood pressure, skin thickness, insulin and BMI data were treated clinically as missing-equivalent data and imputed before model development. The classifiers tested were: logistic regression, decision tree, random forest, XGBoost, support vector machine, and k-nearest neighbors, with the metrics of accuracy, precision, sensitivity, specificity, F1-score and ROC-AUC. The best-performing algorithm was XGBoost with an ROC-AUC of 0.853 while logistic regression had the best accuracy of 79.5%. BMI and age were also strong predictors of diabetes, with glucose being the strongest. The results suggest that automated tools utilizing machine learning algorithms using routine clinical markers may offer valuable assistance in diabetes risk stratification in primary care. Further validation is necessary, however, in larger and more varied patient populations before it is used routinely in the clinic.

Keywords: Diabetes mellitus; Machine learning; Risk prediction; Clinical indicators; Primary healthcare

International Journal for Research in Health Science and Nursing
(ISSN: 2208- 2670)
Vol. 10, Issue 02, June 2024
DOI: 10.65327/hsn.v10i3.2495

1.Introduction

Diabetes mellitus is a common chronic metabolic disease defined by hyperglycaemia due to a failure of insulin production, action or both. It is becoming a significant public health problem because uncontrolled diabetes is linked to cardiovascular disease, kidney dysfunction, neuropathy, retinopathy, and early death. The emphasis in current clinical practice is on the early detection of those who have abnormal glucose regulation and early initiation of preventive and therapeutic measures to avoid the onset of severe complications (Committee, 2023; ElSayed et al., 2025). The early risk stratification is a key component of modern healthcare practice, especially for people who have diabetes-associated metabolic and lifestyle risk factors that can be modified (Committee, 2023).

Primary care providers often measure clinical factors like blood glucose, blood pressure, BMI, age, insulin level, pregnancy history, and diabetes-related family history. These are regularly available and give valuable information for determining who needs a more in-depth diagnostic evaluation or prevention action. Diabetes screening through such indicators can then be practically supported with the use of predictive approaches without highly specialized or expensive resources (Firdous et al., 2022). Community-based studies have also shown the increasing potential of using routinely collected health information to predict type 2 diabetes risk (DesilvaK, 2020).

The classification of diabetes and prediction of the disease risk has been increasingly explored using machine learning methods, due to the ability to capture the complexity and potential nonlinearity of multiple clinical variables. Logistic regression, decision trees, random forests, support vector machines, k-nearest neighbors, boosting techniques and ensemble classifiers are effective classifiers in diabetes-related applications (Butt et al., 2021; Hasan et al., 2020). Research with the Pima Indians diabetes dataset has also shown that popular metrics can be used to facilitate effective classification if suitable machine learning methods and pre-processing are employed (Chang et al., 2023).

According to comparative evaluations, the sensitivity, specificity, discrimination, and overall predictive power of different algorithms differ, highlighting the need to test multiple algorithms and not just a single classifier (Kaur & Kumari, 2022). Recent advances in diabetes prediction are also documented, with a focus on sophisticated machine learning and artificial intelligence methods (Jaiswal et al., 2021; Khokhar et al., 2025). The combination of information from many classifiers further enhances the robustness of prediction using ensemble approaches (Kibria et al., 2022).

Interpretability is also an important feature of predictive accuracy in health care applications. For machine learning to meaningfully assist in clinical decisions, clinicians must have an understanding of which patient factors are most predictive of risk. The predictive reliability and transparency in diabetes risk assessment are thus key to recent interpretable machine learning

frameworks (Khan et al., 2026). When using AI or machine learning in clinical prediction, the reporting guidelines like TRIPOD+AI also emphasize model transparency, meaning how the model is developed, its validation, performance, and reporting (Collins et al., 2024).

Although there has been significant advancement, there are a number of limitations to diabetes prediction studies. Most studies emphasize maximization of correct classifications and neglect to address clinical interpretation, data quality issues, model comparison and clinical implementation in primary healthcare. The unreliability of prediction is affected by the absence or clinical improbability of measurements of variables like glucose, blood pressure, insulin, BMI and skin thickness if not managed appropriately. At the same time, there is a need for relatively simple models which are based on indicators that are commonly available and have adequate predictive power for the healthcare decision support arena. Combining accurate prediction with clinically explainable information, Explanable ensemble and artificial intelligence approaches offer a solution to these limitations (Khokhar et al., 2025; Kibria et al., 2022).

This research is set to create and test machine learning models to forecast the risk of diabetes based on standard clinical data as evidence for diabetes primary healthcare decision making. Specifically, the study investigates the clinical features associated with diabetes, preprocesses clinically implausible values, builds and compares various classification algorithms with standard performance measures, identifies the most important features that influence the prediction of diabetes risk, and assesses the clinical value of diabetes risk prediction by machine learning as a supportive tool for primary healthcare decision making.

2. Materials and Methods

2.1 Study Design

A machine learning (ML) approach for prediction of diabetes risk based on routine clinical indicators was performed using a retrospective secondary-data design. The aim was to compare several classification models and to find the most informative predictors that are pertinent to primary healthcare decision support.

2.2 Dataset and Data Source

There were 1,168 patient records in the dataset, with nine variables, eight of which were predictors and one binary outcome (Hernesto, 2026). The dependent variable was the diabetes status (0: non-diabetes, 1: diabetes). The non-diabetic cases accounted for about 66.0% and the diabetic cases accounted for about 34.0%.

2.3 Study Variables

Predictors were pregnancy, glucose, blood pressure, skin thickness, insulin, body mass index, Diabetes Pedigree Function and age. These variables were demographic, metabolic,

anthropometric, and hereditary risk factors of diabetes. The target variable used for the development of the model is Diabetes status.

2.4 Data Preprocessing

Duplication, missing values, abnormal values, and class imbalance were investigated. There were no duplicate records or commonly coded missing values. But clinically implausible zero values were found for glucose, blood pressure, skin thickness, insulin and BMI and these were regarded as missing-equivalent values. Zero values for pregnancies were kept, as these values were clinically valid. The missing-equivalent values were imputed using median imputation using only the training data. Continuous variables were scaled as necessary for scale-sensitive algorithms.

2.5 Exploratory Data Analysis

The clinical variables and the diabetes outcome distribution were summarized using descriptive statistics. Appropriate measures of central tendency and dispersion were used to describe continuous variables. Clinical characteristics were compared for the diabetic and non-diabetic groups and the relationship between clinical characteristics was assessed by correlation analysis for potential multicollinearity.

2.6 Machine Learning Models

The six supervised classification algorithms that were tested include logistic regression, decision tree, random forest, extreme gradient boosting, support vector machine, and k-nearest neighbors. Logistic regression was used as the baseline model and tree-based, boosting, margin-based and distance-based algorithms were added to model different patterns in the clinical data.

2.7 Model Training and Validation

Using stratification, the data were split into training and testing data sets with the same proportions of diabetic and non-diabetic individuals. To avoid data leakage, preprocessing, imputation, scaling and hyperparameter tuning were only done on the training data. During model development, stratified cross-validation was used and the independent test set was used for the assessment of the final model performance.

2.8 Model Performance Evaluation

The model performance was assessed based on Accuracy, Precision, Sensitivity, Specificity, F1 score and ROC-AUC. True-positives, true negatives, false positives, and false-negatives were summarized in confusion matrices. ROC curves were also constructed to compare the discriminatory power of the models.

2.9 Feature Importance and Interpretation

International Journal for Research in Health Science and Nursing
(ISSN: 2208- 2670)
Vol. 10, Issue 02, June 2024
DOI: 10.65327/hsn.v10i3.2495

Random forest and extreme gradient boosting models were used to assess feature importance. SHAP analysis was performed where appropriate to explore the direction and size of the contribution of individual predictors. This analysis identified the clinical indicators most strongly associated with predicted diabetes risk and was used to enhance the interpretability of the final model for primary care applications.

3. Results

3.1 Dataset Characteristics

A total of 1168 patient records were included, of which 771 (66.0%) were non-diabetic and 397 (34.0%) diabetic. There were eight routine clinical indicators that were available as predictors: pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, Diabetes Pedigree Function, and age. The overall mean glucose level was 121.68, mean BMI 32.33 kg/m² and mean age 33.51 years after clinically implausible zero values were preprocessed and median imputed. Table 1 presents the overall characteristics of the analytical data set. As seen in Table 1, all the variables with the exception of sex, had significant inter-individual variability, and thus held promise as predictors for diabetes status.

Table 1. Overall characteristics of the analytical dataset

Variable	Mean ± SD	Minimum	Maximum
Pregnancies	3.89 ± 3.30	0	17
Glucose	121.68 ± 30.23	44	199
Blood pressure	71.46 ± 14.31	20	122
Skin thickness	29.58 ± 8.50	7	99
Insulin	141.34 ± 95.94	14	846
BMI	32.33 ± 7.22	8.70	67.10
Diabetes Pedigree Function	0.45 ± 0.30	0.08	2.42
Age	33.51 ± 11.31	21	81
Non-diabetic	771 (66.0%)	—	—
Diabetic	397 (34.0%)	—	—

3.2 Data Preprocessing Results

International Journal for Research in Health Science and Nursing
(ISSN: 2208- 2670)
Vol. 10, Issue 02, June 2024
DOI: 10.65327/hsn.v10i3.2495

There were no duplicate records or conventionally coded missing values. Some however, had physiologically unrealistic zero values in some of the variables. These were 24 glucose measurements, 57 blood-pressure measurements, 351 skin-thickness measurements, 549 insulin measurements, and 18 BMI measurements. These zero values were considered as missing-equivalent and replaced with training-data medians. The values of pregnancy of zero were retained since they were clinically valid. The distribution of the zero-coded values that need to be preprocessed is shown in Figure 1.

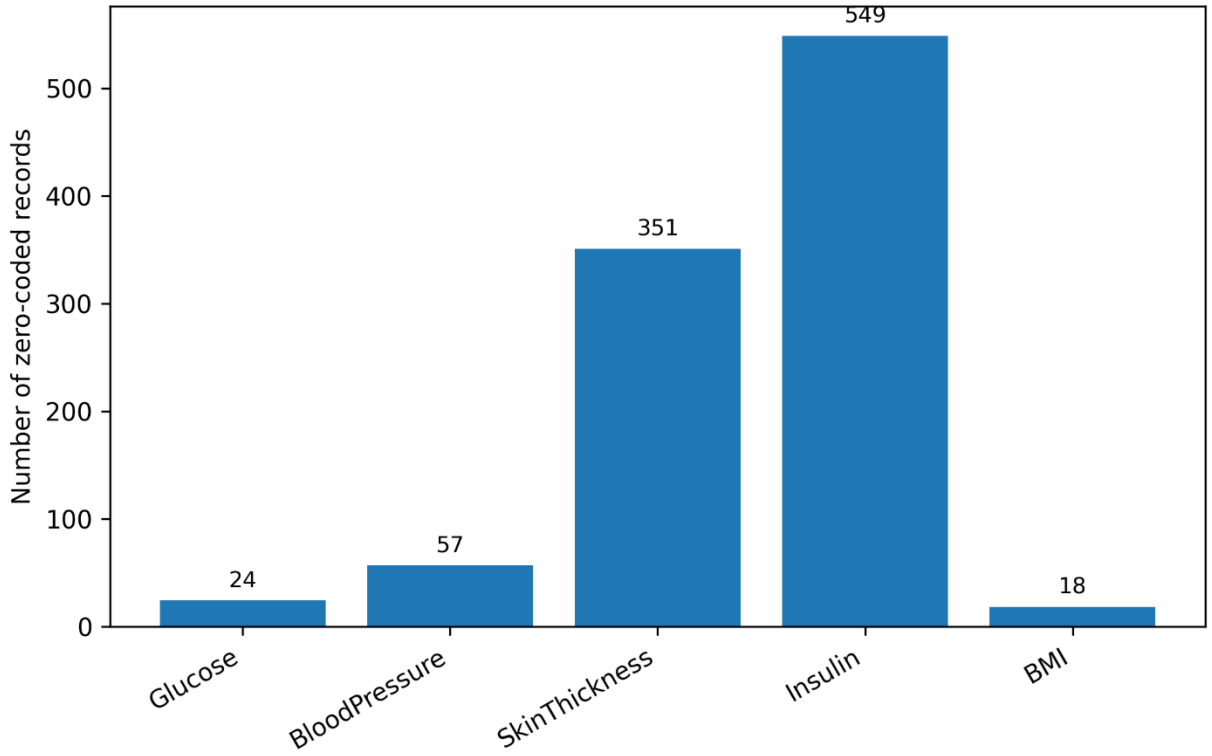


Figure 1. Distribution of clinically implausible zero-coded values across selected predictors

3.3 Descriptive and Exploratory Findings

There were evident disparities between diabetic and non-diabetic groups on a number of predictors. The mean glucose, BMI, age, pregnancy count, insulin level, and Diabetes Pedigree Function were higher in those with diabetes compared to controls. The biggest absolute difference between groups was observed with glucose with a mean of 141.44 in diabetic participants and 111.51 in non-diabetic participants. The diabetic group also had a higher BMI of 35.25 as compared to 30.82 kg/m 2. The greatest clinical differences between the outcome groups were observed in glucose, BMI, age, and pregnancy history, as indicated in Table 2, with blood pressure showing a small difference.

International Journal for Research in Health Science and Nursing
(ISSN: 2208- 2670)
Vol. 10, Issue 02, June 2024
DOI: 10.65327/hsn.v10i3.2495

Table 2. Clinical characteristics according to diabetes status

Variable	Non-diabetic Mean ± SD	Diabetic Mean ± SD
Pregnancies	3.30 ± 2.94	5.04 ± 3.64
Glucose	111.51 ± 24.72	141.44 ± 30.21
Blood pressure	70.75 ± 13.93	72.83 ± 14.94
Skin thickness	28.77 ± 8.45	31.13 ± 8.38
Insulin	133.78 ± 91.26	156.03 ± 102.97
BMI	30.82 ± 7.00	35.25 ± 6.75
Diabetes Pedigree Function	0.43 ± 0.27	0.51 ± 0.34
Age	31.48 ± 10.91	37.47 ± 11.03

3.4 Correlation Analysis

Glucose correlated with diabetes status the most (Pearson correlation coefficient ~ 0.47). BMI had the next highest correlation with outcome ($r \approx 0.29$), and age ($r \approx 0.25$) and pregnancies ($r \approx 0.25$) were also significantly correlated. Weaker positive correlations were found with Diabetes Pedigree Function, skin thickness, insulin, and blood pressure. The overall correlation structure is shown in Figure 2.

International Journal for Research in Health Science and Nursing
(ISSN: 2208- 2670)
Vol. 10, Issue 02, June 2024
DOI: 10.65327/hsn.v10i3.2495

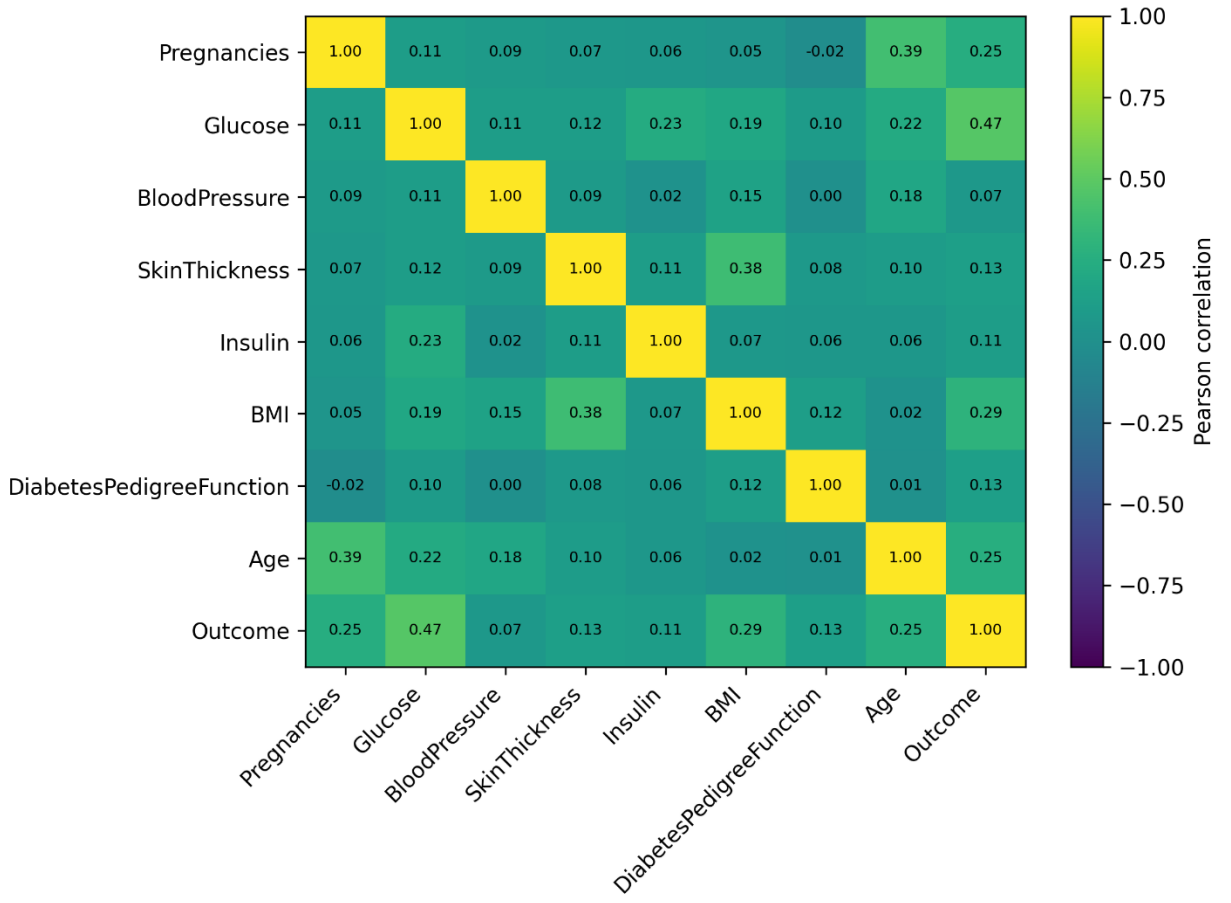


Figure 2. Correlation heatmap of clinical predictors and diabetes outcome

As shown in Figure 2, no two predictors had extremely high correlation, which meant that severe multicollinearity was not likely to significantly impact the development of the model. The strongest association with the binary diabetes outcome was still with glucose.

3.5 Machine Learning Model Performance

Machine-learning models showed discriminatory ability greater than random classification for all six models. XGBoost had the highest ROC-AUC score of 0.853 followed by Random Forest at 0.850, SVM at 0.847 and Logistic Regression at 0.846. Logistic Regression achieved the best test accuracy (79.5%) and the XGBoost model obtained the best sensitivity (66.3%). Both Random Forest and SVM showed high specificity values of > 90%. Table 3 shows that the overall performance of the models was similar for the strong classifiers. When looking at the overall discrimination, XGBoost models performed the best, while Logistic Regression models had the highest accuracy, and XGBoost models had the highest sensitivity for the detection of diabetic cases.

International Journal for Research in Health Science and Nursing
(ISSN: 2208- 2670)
Vol. 10, Issue 02, June 2024
DOI: 10.65327/hsn.v10i3.2495

Table 3. Comparative performance of machine-learning models

Model	Accuracy	Precision	Recall	Specificity	F1-score	ROC-AUC
Logistic Regression	0.795	0.750	0.600	0.896	0.667	0.846
Decision Tree	0.739	0.744	0.363	0.935	0.487	0.786
Random Forest	0.786	0.750	0.563	0.903	0.643	0.850
XGBoost	0.778	0.679	0.663	0.838	0.671	0.853
SVM	0.791	0.754	0.575	0.903	0.652	0.847
KNN	0.756	0.695	0.513	0.883	0.590	0.832

3.6 ROC Curve Analysis

All the classification algorithms were compared by receiver operating characteristic analysis. XGBoost had the highest AUC among the other classifiers, and Random Forest, SVM and Logistic Regression had very similar performance. KNN obtained an intermediate value of 0.832 whereas Decision Tree obtained the lowest value of 0.786. The ROC curves for both XGBoost and Random Forest were grouped tightly, both above the no-discrimination curve, as seen in Figure 3.The ROC curves for XGBoost, Random Forest, SVM and Logistic Regression were grouped close together and were significantly above the diagonal - the one representing no discrimination, as seen in Figure 3. The results suggest fairly, but not entirely satisfactory, discrimination between diabetics and non-diabetics.

International Journal for Research in Health Science and Nursing
(ISSN: 2208- 2670)
Vol. 10, Issue 02, June 2024
DOI: 10.65327/hsn.v10i3.2495

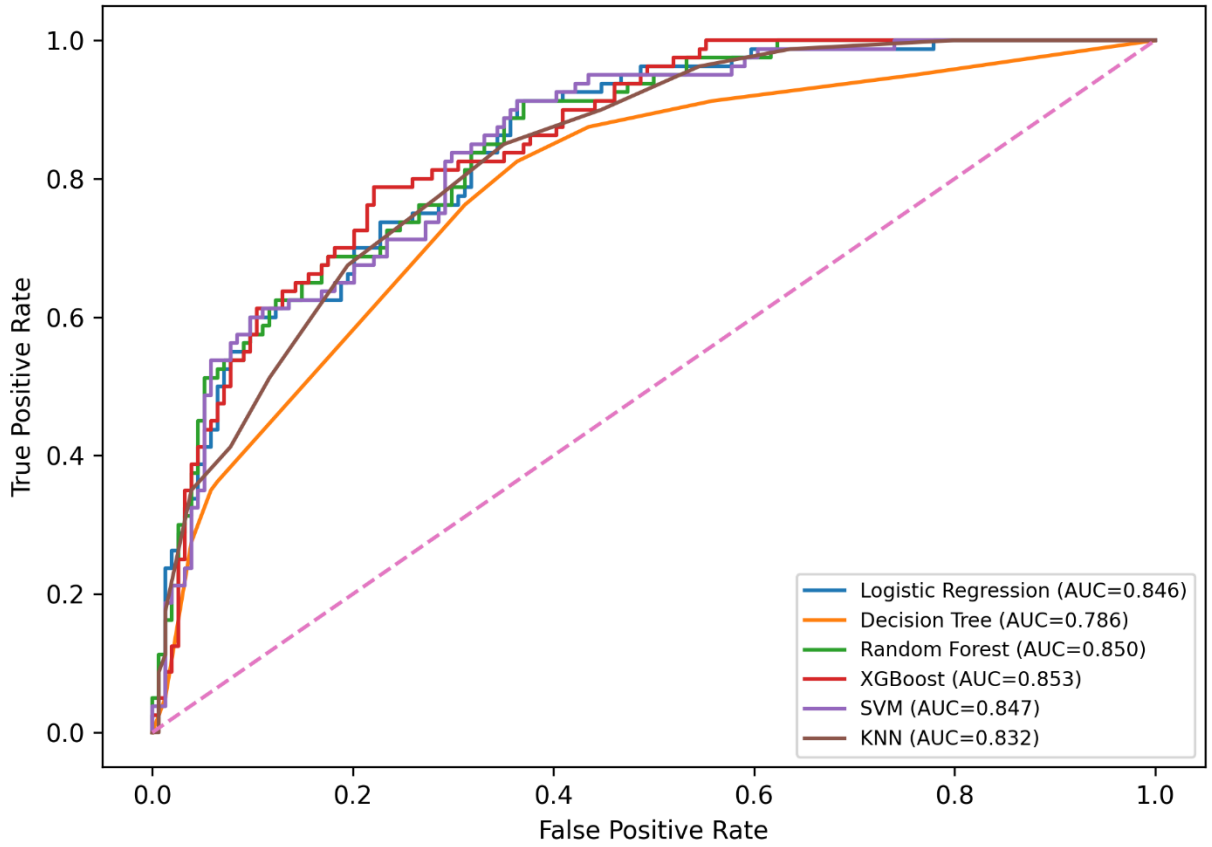


Figure 3. Receiver operating characteristic curves for the six machine-learning models

3.7 Confusion Matrix Analysis

There were also differences in sensitivity and specificity as shown by the confusion matrices. On the test set, XGBoost made the most correct classifications of the diabetic cases (53 out of 80). Decision Tree was the method with the fewest true-positive classifications, but the highest number of true-negative classifications. As seen in Table 4, XGBoost decreased false-negative classifications to 27, while Decision Tree had 51 false negatives. This is significant for diabetes risk screening as it leads to failure to identify high-risk individuals during initial assessment.

Table 4. Confusion-matrix results on the independent test set

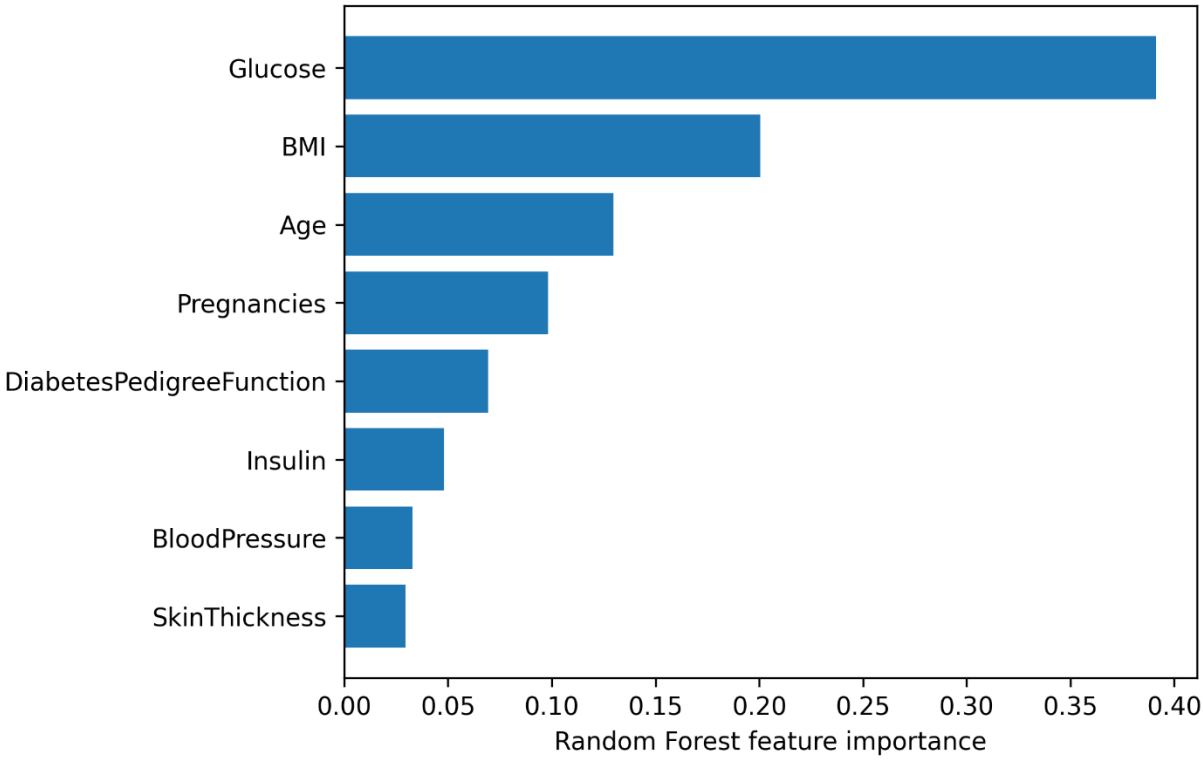
Model	True Negative	False Positive	False Negative	True Positive
Logistic Regression	138	16	32	48
Decision Tree	144	10	51	29
Random Forest	139	15	35	45

International Journal for Research in Health Science and Nursing
(ISSN: 2208- 2670)
Vol. 10, Issue 02, June 2024
DOI: 10.65327/hsn.v10i3.2495

XGBoost	129	25	27	53
SVM	139	15	34	46
KNN	136	18	39	41

3.8 Important Predictors of Diabetes

Random Forest feature-importance analysis showed that the most important predictor was glucose with an importance of ~0.391. BMI was ranked second with 0.201, age was ranked third with 0.130, and pregnancies were ranked 4th with 0.098. The remaining individual factors, Diabetes Pedigree Function, insulin, blood pressure and skin thickness made smaller contributions to the importance of prediction. The results for XGBoost were largely similar, with glucose, BMI and age being the most important. As seen in Figure 4, the discriminatory information for the classification of diabetes was glucose as the dominant predictor and BMI and age with additional information. These predictors were consistent between exploratory analysis and machine-learning interpretation, suggesting that they may be relevant for routine risk assessment in primary healthcare.



International Journal for Research in Health Science and Nursing
(ISSN: 2208- 2670)
Vol. 10, Issue 02, June 2024
DOI: 10.65327/hsn.v10i3.2495

Figure 4. Random Forest feature importance for diabetes risk prediction

4. Discussion

The present study showed that the clinical parameters that are routinely available in a clinical setting facilitate meaningful machine learning-based prediction of diabetes risk. XGBoost had the best ROC-AUC among the six algorithms evaluated and logistic regression had the best overall accuracy. Random forest and support vector machine also yielded promising discrimination, suggesting that both linear and nonlinear methods deliver good discrimination if suitable preprocessing and validation techniques are employed. These results are in line with earlier work on the use of relatively straightforward clinical parameters to improve diabetes risk stratification by machine learning (Kopitar et al., 2020; Sisodia & Sisodia, 2018). The results also confirm the importance of using multiple algorithms, rather than one model alone, given that sensitivity and specificity are used in different clinical contexts.

In the current analysis, glucose was the most important variable, with BMI and age as the second and third most important variables, respectively. The Diabetes Pedigree Function and pregnancy history were also important and comparatively blood pressure and insulin had less predictive value. These patterns are clinically plausible because the factors of hyperglycemia, excess body weight, age and familial predisposition are all known components of the risk of diabetes. Other studies have indicated that glucose-related indicators and anthropometric factors consistently add substantially to the prediction of diabetes (Olisah et al., 2022; Talebi Moghaddam et al., 2024). Additionally, the need for careful preprocessing was obvious as there were some clinical variables with impossible zero values that have been reported to affect diabetes risk modeling (Maniruzzaman et al., 2018).

The overall performance of the present models is in line with previous machine learning studies that reported that structured clinical data yielded useful discrimination for predicting diabetes (Kopitar et al., 2020; Nibareke & Laassiri, 2020). Later studies focused on external validation and on independent larger cohorts to increase the level of confidence in predictive models (Lee et al., 2025). Especially when models are going to be used in healthcare, explainability has gained importance too. Feature importance analysis and SHAP are similar to the current trends of connecting accurate prediction with understandable clinical interpretation (Lundberg et al., 2020; Majyambere et al., 2026). Among the published machine learning approaches in this area, one was later retracted, and should not be considered reliable supporting evidence for the effectiveness of the model.

The results indicate that machine learning models using routinely collected clinical markers are potential tools for primary healthcare to facilitate early diabetes risk screening. These tools used to identify people who need: Clinical decision-support methods have also been suggested to enhance diabetes-related risk recognition and to assist healthcare providers in their daily practice

(Islam et al., 2023). In general, machine learning applications in healthcare are most beneficial when they augment, not supplant, clinical decision making (Rajkomar et al., 2019).

One of the strengths of this study was that it incorporates easily obtained and clinically meaningful predictors. Six machine learning algorithms were also compared and tested for various metrics such as sensitivity, specificity, F1-score, and ROC-AUC. Moreover, model development was preceded by preprocessing to eliminate bias due to implausible values. These provide greater methodological transparency and help to interpret results in a more clinically meaningful way (Olisah et al., 2022).

This study involved a dataset with some predictors and a medium level of class imbalance. There were also significant numbers of missing-equivalent zero values in several variables, such as insulin and skin thickness. The use of imputation has minimized the amount of data lost but may have added uncertainty. The data are also population-specific, which reduces generalizability. Furthermore, the models were not validated externally whereas recent multi-cohort studies emphasize the need for independent clinical validation prior to clinical uptake.

These models need to be tested in larger and more diverse populations and with the inclusion of further lifestyle, laboratory and socio-economic factors. Prospective evaluation in primary health care settings also be helpful in establishing clinical utility. Enhanced transparency, clinician trust, and patient-specific interpretation of diabetes risk estimates can be achieved through further integration of explainable artificial intelligence.

5. Conclusion

This study demonstrated that machine learning effectively supports diabetes risk prediction using routinely available clinical indicators. XGBoost had the best ROC-AUC and logistic regression had the best overall accuracy with random forest and support vector machine also performing highly on predictive performance. The most significant predictor was glucose, and the next most significant ones were BMI and age, which indicates that these factors are significant in the risk stratification of diabetes. The results also revealed that clinically implausible zero values especially in insulin and skin thickness need preprocessing care so that reliable model development can be achieved. The study offers a viable model for detecting patients with high risk of diabetes in primary care by integrating daily clinical measurements and comparative machine learning analysis to identify patients at high risk. These models may help healthcare providers to prioritize their patients and refer them to further examination, confirmatory tests, preventive counseling and closer observation. The results, however, need to be viewed in the context of the secondary character of the data, moderate imbalance of classes, small set of predictors and the lack of external validation. The models should be further validated under larger and more heterogeneous populations in future studies, include more clinical and lifestyle variables, and determine their

practical functioning in the context of the primary care processes before they can be routinely used in clinical practice.

References

1. Butt, U. M., Letchmunan, S., Ali, M., Hassan, F. H., Baqir, A., & Sherazi, H. H. R. (2021). Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications. Journal of Healthcare Engineering, 2021, 1–17. https://doi.org/10.1155/2021/9930985

2. Chang, V., Bailey, J., Xu, Q. A., & Sun, Z. (2023). Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. Neural Computing and Applications, 35(22), 16157–16173. https://doi.org/10.1007/s00521-022-07049-z

3. Collins, G. S., Moons, K. G., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., & Van Smeden, M. (2024). TRIPOD+ AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. Bmj, 385. https://www.bmj.com/content/385/bmj-2023-078378.short

4. Committee, A. D. A. P. P. (2023). 2. Diagnosis and classification of diabetes: Standards of care in diabetes—2024. Diabetes Care, 47(Suppl 1), S20.

5. DesilvaK, L. K. (2020). Useandperformanceof machine learning models for type 2 diabetes prediction in community settings: A systematic review and meta analysis. Int J Med Inform, 143(104268), 10–1016.

6. ElSayed, N. A., Professional Practice Comm, R. G., Mccoy, R. G., Aleppo, G., Balapattabi, K., Beverly, E. A., Briggs Early, K., Bruemmer, D., Ebekozien, O., & Echouffo-Tcheugui, J. B. (2025). 2. Diagnosis and classification of diabetes: Standards of care in diabetes-2025. Diabetes Care, 48, S27–S49.

7. Firdous, S., Wagai, G. A., & Sharma, K. (2022). A survey on diabetes risk prediction using machine learning approaches. Journal of Family Medicine and Primary Care, 11(11), 6929.

8. Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access, 8, 76516–76531.

9. Hernesto, C. (2026). Predicting Diabetes From Tracking Medical Records. 3. https://doi.org/10.17632/nxnty5g7y6.3

10. Islam, R., Sultana, A., Tuhin, Md. N., Saikat, Md. S. H., & Islam, M. R. (2023). Clinical Decision Support System for Diabetic Patients by Predicting Type 2 Diabetes Using Machine Learning Algorithms. Journal of Healthcare Engineering, 2023(1), 6992441. https://doi.org/10.1155/2023/6992441

11. Jaiswal, V., Negi, A., & Pal, T. (2021). A review on current advances in machine learning based diabetes prediction. Primary Care Diabetes, 15(3), 435–443.

12. Kaur, H., & Kumari, V. (2022). Predictive modelling and analytics for diabetes using a machine learning approach. Applied Computing and Informatics, 18(1–2), 90–100.

13. Khan, T. F., Saeed, M., Hussain, M., Khan, L., Khan, M. Z., Nadeem, A., & Alghamdi, T. (2026). A comparative and interpretable machine learning framework for reliable diabetes risk prediction. Scientific Reports. https://www.nature.com/articles/s41598-026-61038-z

14. Khokhar, P. B., Gravino, C., & Palomba, F. (2025). Advances in artificial intelligence for diabetes prediction: Insights from a systematic literature review. Artificial Intelligence in Medicine, 164, 103132.

15. Kibria, H. B., Nahiduzzaman, M., Goni, M. O. F., Ahsan, M., & Haider, J. (2022). An ensemble approach for the prediction of diabetes mellitus using a soft voting classifier with an explainable AI. Sensors, 22(19), 7268.

16. Kopitar, L., Kocbek, P., Cilar, L., Sheikh, A., & Stiglic, G. (2020). Early detection of type 2 diabetes mellitus using machine learning-based prediction models. Scientific Reports, 10(1), 11981.

17. Lee, H., Hwang, S. H., Park, S., Choi, Y., Lee, S., Park, J., Son, Y., Kim, H. J., Kim, S., & Oh, J. (2025). Prediction model for type 2 diabetes mellitus and its association with mortality using machine learning in three independent cohorts from South Korea, Japan, and the UK: A model development and validation study. EClinicalMedicine, 80. https://www.thelancet.com/journals/eclinm/article/PIIS2589-5370(25)00001-X/fulltext

18. Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56–67.

19. Majyambere, S., Lindgren, T., Twizere, C., & Ntakirutimana, I. (2026). Early Type 2 diabetes risk prediction using explainable machine learning in a two-stage approach. Frontiers in Digital Health, 8, 1743619.

20. Maniruzzaman, Md., Rahman, Md. J., Al-MehediHasan, Md., Suri, H. S., Abedin, Md. M., El-Baz, A., & Suri, J. S. (2018). Accurate Diabetes Risk Stratification Using Machine

Learning: Role of Missing Value and Outliers. Journal of Medical Systems, 42(5), 92. https://doi.org/10.1007/s10916-018-0940-7

21. Nibareke, T., & Laassiri, J. (2020). Using Big Data-machine learning models for diabetes prediction and flight delays analytics. Journal of Big Data, 7(1), 78. https://doi.org/10.1186/s40537-020-00355-0

22. Olisah, C. C., Smith, L., & Smith, M. (2022). Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. Computer Methods and Programs in Biomedicine, 220, 106773.

23. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine Learning in Medicine. New England Journal of Medicine, 380(14), 1347–1358. https://doi.org/10.1056/NEJMra1814259

24. Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. Procedia Computer Science, 132, 1578–1585.

25. Talebi Moghaddam, M., Jahani, Y., Arefzadeh, Z., Dehghan, A., Khaleghi, M., Sharafi, M., & Nikfar, G. (2024). Predicting diabetes in adults: Identifying important features in unbalanced data over a 5-year cohort study using machine learning algorithm. BMC Medical Research Methodology, 24(1), 220. https://doi.org/10.1186/s12874-024-02341-z