Malaysian Journal of Computing, 11 (2): 2561-2577, 2026
eISSN: 2600-8238

ASSESSMENT OF AGE-WISE DIABETES RISK USING BIOCHEMICAL INDICATORS AND MACHINE LEARNING CLASSIFICATION MODELS

Farwa Afzal[1*], Mohibullah Aziz[2] and Shariffah Suhaila Syed Jamaludin[3]

[1]National University of Computer and Emerging Sciences (NUCES), Lahore, Pakistan*
[2]Service Institute of Medical Science, Lahore, Pakistan
[3]Department of Mathematical Sciences, Universiti Teknologi Malaysia, Johor Bahru, Malaysia
[1*]farwa.afzal008@gmail.com, [2]mohibullah.azizkhan@gmail.com, [3]suhailasj@utm.my
(indicates the corresponding author)*

ABSTRACT

Diabetes mellitus is a major metabolic disease across the world, leading to serious complications and healthcare challenges. Prediction of diabetes at an early stage with the help of biochemical as well as demographic factors can lead to better diagnosis and preventive measures for the same. This research paper aims to predict diabetes using biochemical parameters in combination with various machine learning classification techniques, stratified by age group. This study uses a dataset of diabetes data that includes demographic and clinical attributes like age, gender, body mass index (BMI), urea, creatinine, HbA1c, cholesterol, TG, HDL, LDL, and VLDL. Machine Learning classifiers, including logistic regression, K-nearest neighbours (KNN), naive Bayes, SVM, random forest, and XGBoost, have been used in this study. The initial dataset consisted of 1,000 records, but 53 of them were removed because they were prediabetic. Thus, 947 records of diabetic and non-diabetic individuals were used for binary classification. The subjects were categorized into six age groups based on ages from 20-29 up to 70-80. Stratified 80:20 train-test split was utilized, and 5-fold stratified cross-validation was conducted in the training data set. Preprocessing and SMOTE were performed only in the training data set. The best performances were obtained by Random Forest and XGBoost; they reached 100% accuracy with the inclusion of HbA1c while the mean cross-validation accuracies for them were 98.5% and 99.1%. In sensitivity analysis where HbA1c was excluded, these two models kept their test accuracy of about 98.9%, showing that they performed well even without the use of HbA1c. HbA1c, BMI, and age were found to be some of the most informative variables for diabetes status. This research showed strong classification performance at the dataset level. Nevertheless, external validation is needed using different large clinical datasets.

Keywords: *Diabetes Mellitus, Machine Learning, XGBoost, Random Forest, Biochemical Markers, Diabetes Classification.*

Received for review: 22-08-2026; Revised: 29-09-2026; Accepted: 30-09-2026; Published: 01-10-2026
DOI: 10.24191/mjoc.v11i2.12095

This is an open access article under the CC BY-SA license (https://creativecommons.org/licenses/by-sa/3.0/).

1. Introduction

Diabetes mellitus is a condition associated with hyperglycemia, which is defined as elevated blood sugar due to the failure of either proper insulin production or the body's resistance to the effects of insulin. It has become increasingly common throughout the world within the past few decades, becoming a health problem with no barriers regarding location, social status, or ethnicity. As noted by the International Diabetes Federation, there are currently about 537 million adults who are suffering from the disease around the globe in 2021, while by 2030 this number is expected to rise above 643 million (Magliano *et al*., 2021). However, apart from statistics, a disturbing phenomenon relates to the age at which diabetes is being diagnosed, especially type 2, which was typically an adult disease but is now increasingly found among teenagers.

The physiological consequences of poorly controlled diabetes extend far beyond the issue of impaired glucose homeostasis. Chronic exposure to high levels of glucose creates an increased predisposition towards several conditions like heart disease, peripheral neuropathy, diabetic retinopathy, and nephropathy, which increase morbidity and lower the quality of life of patients while creating huge burdens for healthcare services (Tasin *et al*., 2023). Biochemical parameters have a fundamental importance in the diagnosis of diabetes and associated complications. The gold standard biochemical parameter to assess long-term glycemic control continues to be glycated hemoglobin, as it represents the average blood sugar level of two to three months (Feng *et al*., 2023). The evaluation of renal function parameters, which include serum urea and serum creatinine levels, is crucial because diabetic nephropathy represents a common complication of diabetes mellitus. Furthermore, abnormal lipid levels in the plasma indicate dysmetabolism in diabetic patients and contribute to the risk of cardiovascular disease (Abnoosian *et al*., 2023). Moreover, anthropometrical data are very important since BMI is an established marker of metabolic risk, and adiposity leads to insulin resistance and type 2 diabetes (Wee *et al*., 2024).

Despite the presence of well-known markers from clinical practice, the current methods of diagnosis and risk assessment still cannot address the complexity and nonlinearity inherent in the disease pathology. The typical assumption of statistical models is that the relationship between the variables in the model is linear, and these models may be inadequate in terms of processing large-scale, heterogeneous clinical datasets. This issue becomes especially acute when one tries to describe the risks associated with diabetes for different age groups, where interaction between markers, demographic factors, and physiological parameters is highly non-trivial (Ahmed *et al*., 2025).

Machine learning represents a promising approach to overcome such obstacles faced when analyzing biomedical data. Compared to classical statistical methods, machine learning algorithms have an advantage in their ability to capture more complicated, non-linear relations; they also work well with higher dimensional feature spaces and are capable of discovering hidden patterns without any a priori assumption regarding the distribution of data. The above-mentioned advantages make ML widely studied in connection with the application of machine learning techniques to the problem of diabetes prediction and diagnosis (Sampath *et al*., 2024). Several supervised learning models, such as logistic regression, random forests, SVM, and gradient boosting have shown significant discriminatory power in predicting the presence of diabetes based on sets of variables describing patients' clinical and biochemical profiles (Parimala *et al*., 2023). Multilayer artificial neural networks and other deep learning models have made another leap in the area, allowing researchers to learn more complicated interaction between features (Patidar *et al*., 2025). Another promising approach is based on assembling, which consists in averaging predictions of several base learners to minimize the variance of results (Xiao *et al*., 2025).

Several scientific evidences validate the clinical efficacy of using different categories of biomarkers in an ML-based prediction system. Research studies have confirmed that the

combination of biomarkers from lipid profile, kidney function, inflammation, and anthropometric categories along with the use of glycemic biomarkers is more effective than using only one type of biomarkers in the process (Lee *et al*., 2025). This was observed when using HbA1c, lipid profile variables, and BMI together in an ML pipeline results in more accurate outcomes compared to those obtained by using only one variable at a time (Yesupatham *et al*., 2025). ML prediction models based on the use of metabolic markers are better than clinical criteria in predicting prediabetes and early-stage diabetes (Ayoade *et al*., 2025). Furthermore, it was established by Rahman *et al*. that lipid biomarkers provide greater accuracy for ML-based diabetes prediction system (Rahman *et al*., 2025). In addition, Shetty *et al*. have found that the incorporation of demographic and biochemical markers in the process of gradient boosting models, including XGBoost, can provide a classification accuracy greater than 90% (Shetty *et al*., 2026).

Despite these developments, however, there still exist some crucial deficiencies in the current literature. First, little attention has been paid to studying the differences in the risk profiles and the distribution of biomarkers among different age cohorts, thus preventing an understanding of age-related metabolic weaknesses. Secondly, the vast majority of past works have been devoted exclusively to binary classification problems, without specifying those biochemical markers that are the most significant during specific life periods. Lastly, it remains unclear whether the machine learning models built based on homogenous data could be generalized to a heterogeneous clinical population (Masood *et al*., 2026).

This current study is intended to overcome these deficiencies through stratification on the basis of age to analyze the risk associated with diabetes using a clinical database consisting of several important biochemical parameters like urea, creatinine, HbA1c, cholesterol, triglycerides, high-density lipoprotein (HDL), low-density lipoprotein (LDL), very-low-density lipoprotein (VLDL), and body mass index (BMI). Several machine learning classification methods are applied for comparison purposes to determine the best performing model. Using biomarker analysis along with age stratification and machine learning techniques will help us to develop a better understanding of diabetes risk prediction.

2. Methods and Materials

2.1 Dataset Description

The data set used in this experiment comes from the Multiclass Diabetes Data Set on the open-source data repository at Kaggle (Yasir Hussein Shakir) (*Multiclass Diabetes Dataset*, n.d.) . This data set consists of 1,000 patient records and has demographic, anthropometric, and biochemical variables associated with diabetes. The variables used in the data set are gender, age, serum urea, creatinine, glycated haemoglobin (HbA1c), total cholesterol, triglycerides (TG), (HDL), (LDL), (VLDL), (BMI), and diabetes class.

The initial dataset had three class labels, namely: Y (Diabetic), N (Non-Diabetic), and P (Predict-Diabetic). After checking and standardization of the class labels, the dataset had 844 Diabetics, 103 Non-Diabetics, and 53 Predict-Diabetics. However, since the current study is concerned with binary classification (between diabetics and non-diabetics), the 53 Predict-Diabetics were excluded before the development of the models to be left with 947 observations for binary classification. Hence, the modelling dataset contained 844 diabetics and 103 non-diabetics (See Table 1).

There were no missing values in the dataset used in the current study. Ages of the participants varied from 20 to 79 years. In order to carry out age-wise analysis, the participants were grouped into six age groups: 20-29, 30-39, 40-49, 50-59, 60-69, and 70-79 years.

Table 1 Distribution of the original diabetes classes

Class	Description	Original Records	Included in Binary Model
N	Non-Diabetic	103	Yes
P	Predict-Diabetic	53	No
Y	Diabetic	844	Yes
Total		1000	
Final Binary Sample		947	Yes

2.2 Age Group Stratification

In order to examine the link between age and diabetic risks, the study sample was divided into six age categories: 20-29 years, 30-39 years, 40-49 years, 50-59 years, 60-69 years, and 70-80 years. The rationale behind this stratification approach lay in the desire to observe the different biochemical marker distributions and prevalence rates of diabetes in different life stages, hence gaining insights about the evolution of metabolic risks with age.

2.3 Data Preprocessing

Missing values, formatting problems, and inconsistent categorical labels were checked in the dataset before building the models. In order to reduce information leakage, the binary dataset was initially partitioned into train and holdout test sets via stratified splitting in an 80:20 ratio. Parameter estimation of all preprocessing methods was done on the training data alone. When performing five-fold cross-validation, each training fold underwent independent preprocessing and SMOTE.

In summary, the sequence of pre-processing steps was: label cleaning and removal of predict-diabetic class label → stratified train-test splitting → categorical encoding → scaling of features → SMOTE on training data alone → model training and cross-validation → test set evaluation.

2.3.1 Missing Value Treatment

Prior to model building, the data set was inspected for missing values and anomalies. Missing values were not found, hence there was no need for imputation. All categorical data and class labels were also inspected and standardized prior to further processing.

2.3.2 Categorical Encoding

The categorical variables were transformed into numerical formats that could be fed into the machine learning algorithms. The method for encoding the data was chosen based on the training data and was consistently applied to the test set which was out of sight of the algorithm during training.

2.3.3 Feature Scaling

The continuous predictor variables were standardized with the help of StandardScaler in Scikit-learn. The transformer fitted on the training data and the obtained parameters of the transformation were applied to the test set to prevent leakage of information from the test observations.

2.3.4 Class Imbalance Handling

As the final binary dataset was imbalanced, SMOTE, Synthetic Minority Over-sampling Technique, was used to over-represent the minority class when building the models. SMOTE was only applied to the training data and was not applied to the test data. In case of 5-fold cross-validation, the SMOTE and the preprocessing transformations were performed separately for each training fold, whereas the validation folds were not touched during the preprocessing operations and generating synthetic samples.

2.3.5 Data Partitioning

After the exclusion of 53 predict-diabetics cases, the 947 observations that remained were split in an 80:20 proportion to generate the training set and test set. The resulting dataset consisted of 844 diabetic and 103 non-diabetic observations, so by using stratified 80:20 split this produced 757 observations for model training and 190 for final testing. Stratification was adopted to achieve balanced proportions for both Diabetic and Non-Diabetic classes within the two sets in order to obtain a test set mirroring the entire data distribution.

2.4 Classification Models

In this research, six supervised classification machine learning models were used: Logistic Regression model, k-Nearest Neighbours (KNN), Naïve Bayes, Support Vector Machine (SVM), Random Forest and Extreme Gradient Boosting (XGBoost). The models were implemented using Scikit-learn and XGBoost in Python. The same training, cross-validation, preprocessing, and held-out testing framework was applied consistently to all classifiers to ensure a comparable evaluation.

2.4.1 Logistic Regression

Logistic Regression is a parametric and linear classification algorithm which computes the probability of class membership using the application of the logistic (sigmoid) function to the linear combination of predictor variables. The Logistic Regression model was trained using the Scikit-learn library, the L2 penalty parameter was applied to avoid overfitting, and the lbfgs solver was set for up to 1000 iterations.

2.4.2 K-Nearest Neighbours

The k-Nearest Neighbors algorithm classifies the diabetes category of an observation depending on the categories of its nearest observation in the feature space. Since the KNN algorithm is scale-sensitive, continuous variables used for prediction were scaled down prior to model building. The model was used to classify diabetic and non-diabetic people depending on their similarities in the chosen demographics and biochemical features.

2.4.3 Naive Bayes

Naive Bayes classifier is a probabilistic method that uses Bayes' theorem and also conditional independence between the predictor variables given the class. In this study, Naive Bayes was used to estimate the likelihood of being diabetic and non-diabetic depending on the distribution of demographic and biochemical predictors.

2.4.4 Support Vector Machine

Support Vector Machine (SVM) classification was used to develop a decision rule that maximizes the separation between diabetic and non-diabetic samples. Since SVM classification is sensitive to the scale of features, standardized predictor variables were used for the development of the SVM classifier. The evaluation of SVM was done using the same process used for evaluating other classifiers.

2.4.5 Random Forest

Random Forest is a machine learning algorithm that builds multiple decision trees from the sample data using bootstrapped samples and random subset of predictor variables. Final prediction is made by combining the predictions of individual decision trees. This approach takes into account the non-linear interaction and relationship between biochemical and demographic predictors

2.4.6 Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is an ensemble boosting method where decision trees are grown one after another and each subsequent tree tries to minimize the prediction error made by the previously grown trees. Regularization and gradient boosting techniques are used in XGBoost. This method is compared to the other classifiers on the same datasets used for classifying diabetes.Model Training and Validation

During each cross-validation iteration, all preprocessing transformations and SMOTE were fitted exclusively on the training portion of that fold and were not fitted using the corresponding validation portion. To further assess the robustness of the models, to minimize the chance of overfitting to the train-test split used in validating the model, and to ensure that model parameters are not tuned in such a way as to perform well on a single train-test split but not generalize to unseen data, stratified k-fold cross validation was performed using k=5 during the development process. The performance of each classifier was then estimated as the average score over all folds of the k-fold cross-validation procedure.

2.5 Sensitivity Analysis for Potential Target Leakage

As HbA1c has been identified as an established marker for diabetes prediction, it was deemed necessary to conduct a sensitivity analysis in order to check if the results were overly reliant on this particular variable. The following sensitivity analysis was performed by employing all six classifiers using two predictor sets: (i) full predictor set that included HbA1c and (ii) predictor set where HbA1c had been removed.

In both cases, the same method of data partitioning and evaluation was utilized. A binary modeling data set was split in an 80/20 manner based on stratification. Preprocessing and balancing of classes were performed using only training data, while five-fold stratified cross-validation was employed during modeling. Subsequent performance was evaluated on the untouched test set.

The comparison of the two predictor sets would be used in order to measure the impact of HbA1c on performance and to explore possible concerns related to target leakage.

2.6 Performance Evaluation Metrics

The effectiveness of model prediction was evaluated with four classic classification metrics, including Accuracy, Precision, Recall (sensitivity), specificity, negative predictive value (NPV), F1-score, receiver operating characteristic–area under the curve (ROC-AUC), precision-recall area under the curve (PR-AUC), and the confusion matrix. Each of the four classification measures was obtained based on the confusion matrix of test set predictions and is formally described below:

Accuracy reflects the percentage of correct classification for the entire population regardless of class membership:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

Precision refers to the percentage of the predicted positive cases which turn out to be truly positive and shows the reliability of positive predictions made by the model:

$$\text{Precision} = TP / (TP + FP)$$

Recall (Sensitivity) characterizes the ratio of actual diabetic cases correctly identified by classifier and can be seen as the ability of the algorithm to detect diabetics:

$$\text{Recall} = TP / (TP + FN)$$

Specificity represents the proportion of actual non-diabetic observations that are correctly identified:

$$\text{Specificity} = TN / (TN+FP)$$

Negative Predictive Value (NPV) represents the proportion of observations predicted as non-diabetic that are actually non-diabetic:

$$\text{NPV} = TN / (TN+FN)$$

F1-score represents the harmonic mean of Precision and Recall, thus giving a well-balanced measure useful especially under class imbalance:

$$\text{F1-score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

ROC-AUC measures the ability of a classifier to distinguish between diabetic and non-diabetic observations across different classification thresholds.

PR-AUC summarizes the relationship between precision and recall across classification thresholds and provides additional information when the outcome classes are imbalanced.

Here, TP, TN, FP, and FN stand for True Positive, True Negative, False Positive, and False Negative. Each of the four metrics was calculated separately for different age group strata as well as for the overall data.

2.7 Implementation Environment

All the computations were done using Python programming language. Machine Learning activities involved usage of Scikit-learn package, whereas Pandas and NumPy packages respectively provided tools for data manipulation and numeric computing. Data visualization was achieved using Matplotlib and Seaborn, to present graphs describing biomarkers, their distribution, proportions in each class and results of models applied. Implementation of SMOTE was done through imbalanced-learn package, providing Scikit-learn wrapper for sampling techniques.

3. Results and Discussion

3.1 Descriptive Analysis of Biochemical Indicators by Diabetes Status

3.1.1 BMI Distribution

Distribution of BMI relative to diabetes in diabetic vs. non-diabetic groups is shown graphically below in Figure 1. Non-diabetic patients (CLASS_NO=0) demonstrated a very tight range of BMI values with median close to 22 kg/m², and BMI values mainly varying between 19 and 25 kg/m², which signifies that patients under consideration represent a metabolically thin group lacking excess fat mass. Patients with diagnosed diabetes mellitus (CLASS_NO=1), on the other hand, demonstrated much higher BMI values, with the median value close to 30 kg/m², and IQR ranging roughly from 29-33 kg/m². Distribution of diabetes patients was also much broader, implying great variability in body constitution in relation to diabetes in the patient group. It should be noted that several outliers, namely those with BMI>43 kg/m², were found only in the diabetic group but not in the control sample.

The high elevation and broad distribution of the BMI measures in the case of diabetics aligns well with existing scientific knowledge regarding the role played by increased adiposity, especially accumulation of visceral fat, in causing insulin resistance through disturbances in adipokine regulation and chronic inflammation. The low level of overlap of the two distributions for high BMI readings also emphasizes the discriminatory power of the variable, indicating that BMI is likely to be an important predictor in the machine learning models used in this study.

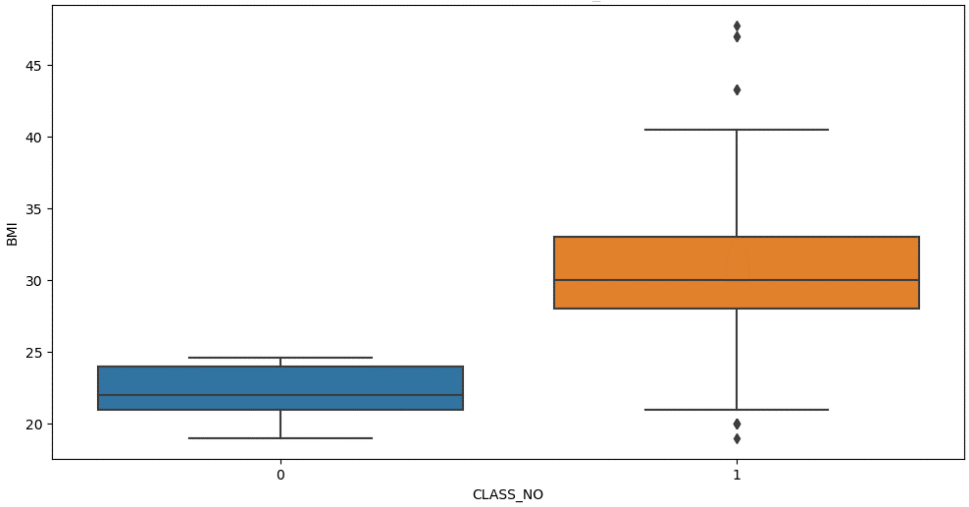


Figure 1 Box Plot of BMI by CLASS_NO

3.1.2 Serum Urea Distribution

The distribution of serum urea values classified according to diabetic status is shown in the form of a violin plot in Figure 2. There were no significant differences between the two groups concerning the central tendency of serum urea values, which could be estimated at around 4-5 mmol/L for each of the groups. On the other hand, a notable difference between the two groups was apparent in the higher range of the distribution of interest. The highest value obtained in the non-diabetic group was around 23 mmol/L, while the diabetic patients had much higher urea values, up to around 40 mmol/L.

The presence of very high levels of urea in certain patients suffering from diabetes can be explained scientifically in relation to diabetic nephropathy, which involves chronic hyperfiltration of the glomeruli, resulting in deterioration of renal function and poor urea excretion, due to chronic hyperglycaemia. While the two distributions have considerable overlap on the lower end of the urea scale, the fact that only extremely high values of urea are observed in the diabetic population suggests that urea is indeed discriminatory on the higher end of the scale.

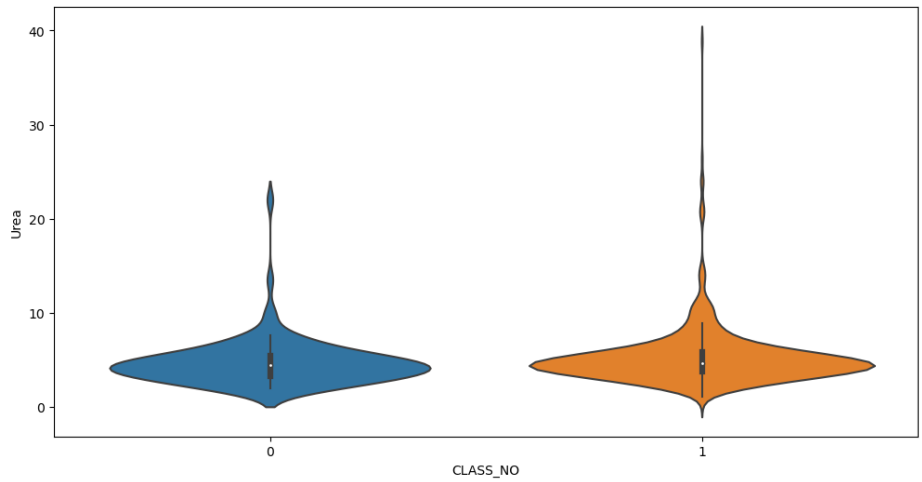


Figure 2 Boxplot of UREA and CLASS_NO

3.1.3 Age-Stratified Diabetes Distribution

Figure 3 portrays the number of diabetic and non-diabetic individuals in the six age categories defined. An extreme imbalance between the number of individuals belonging to each category was noted in almost all age categories. Age Category 4 had the maximum number of diabetic patients of all age groups, amounting to more than n=525 individuals, which is far greater than the number of diabetic individuals of any other age group combined. Age Category 5 had the second-highest number of diabetics, which was about n=210 individuals. Both of these age categories together constitute the vast majority of diabetic individuals in the dataset as stated in Table 2. On the other hand, Age Categories 1 and 2 have a negligible contribution towards the diabetic category, with counts of less than n=10 and n=40, whereas the non-diabetic category maintains a fairly uniform count across all age categories not exceeding n=50 in Age Categories 2 and 3.

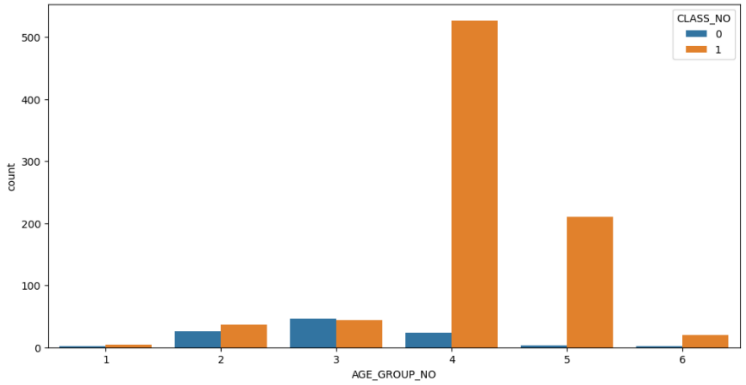


Figure 3 Count plot of AGE_GROUP_NO by CLASS_NO

This stratification according to age is in line with the documented relationship between an increase in age and risk of developing type 2 diabetes, which results from declining levels of insulin sensitivity and decreased secretory capacity of pancreatic beta cells coupled with prolonged exposure to metabolic risk factors. The clustering of cases of diabetes in the higher-age categories suggests that being in any particular age category significantly improves the predictive power of the classification algorithm based on diabetes, so that the probability of a positive diagnosis will increase monotonically as a function of age. Moreover, the fact that age-based prevalence patterns mirror real-world distribution of epidemiological data gives the analysis ecological validity.

Table 2 Distribution of diabetes classes across age groups

Age group (years)	**Non-Diabetic (N)**	**Predict-Diabetic (P)**	**Diabetic (Y)**	**Total**
20–29	2	0	5	7
30–39	26	18	37	81
40–49	46	17	44	107
50–59	24	18	527	569
60–69	3	0	211	214

70–79	2	0	20	22
Total	**103**	**53**	**844**	**1,000**

3.1.4 HbA1c Distribution

Figure 4 shows a swarm plot that displays the HbA1c value distribution of the two diabetes types at the individual level. The HbA1c values for individuals without diabetes were narrowly distributed around their normal clinical values ranging from about 3.8% to 5.9%, with a high frequency of readings clustered around the interval between 4.2% and 5.2%. There was a very small number of outliers in the HbA1c values of individuals without diabetes, which ranged between less than 2.0% and could be attributed to measurement or data entry errors. For diabetic individuals, HbA1c readings showed a much wider and high-frequency distribution ranging between 3.0% and 16.0%, clustered around 7% to 12% greater than 6.5%.

It should be noted that almost complete spatial discrimination of the two class distributions in the HbA1c dimension, with practically no superimposition beyond the threshold of 6.5%, clearly indicates that it is the sole strongest variable in the dataset. This finding is entirely consistent with the outcomes of the correlation analysis performed on the dataset in question, as well as with the recognized clinical significance of HbA1c as a measure of the average level of blood sugar experienced during the last two to three months. It is notable that there is a relatively high degree of class-to-class variation in HbA1c readings within the diabetic population, ranging from elevated to severely abnormal.

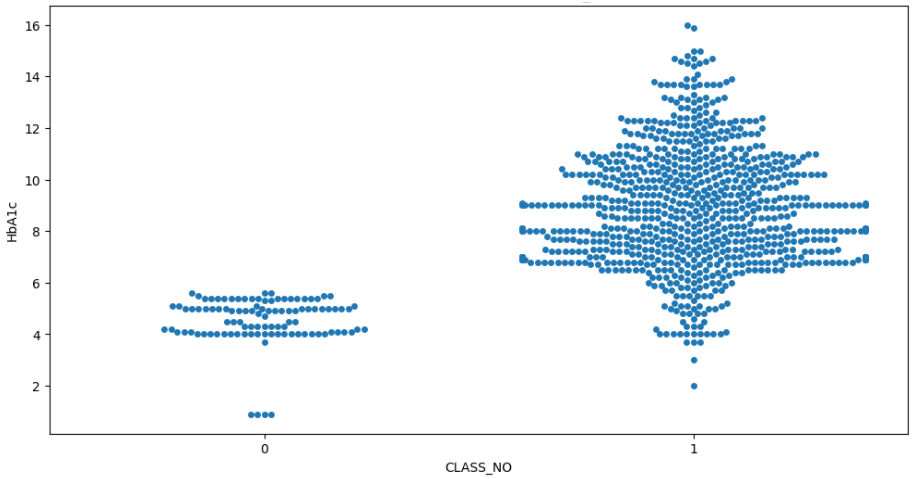


Figure 4 Swarm Plot of HbA1c by CLASS_NO

3.1.5 Total Cholesterol Distribution

The strip plot shown in Figure 5 represents the distribution of total cholesterol measurements by diabetes status. Patients with non-diabetes conditions showed a relatively tight distribution, where most of the values ranged between 2.5 and 6.5 mmol/L, while only some observations exceeded 7.5 mmol/L. On the contrary, the data obtained for diabetic patients produced a much tighter and taller point cloud with an interval ranging from less than 1.0 mmol/L up to more than 10.0 mmol/L, with most of the values concentrated between 3.5 and 7.0 mmol/L.

The high cholesterol variability and wide distribution among patients with diabetes may be explained by the adverse impact of the lack of insulin on metabolic processes, such as the inhibition of lipoprotein lipase activity and the increase in the production of very low-density lipids by the liver. Even though there were low correlations between the values of total

cholesterol and the diabetes class label, it can still provide valuable supplementary information alongside other features like HbA1c and BMI.

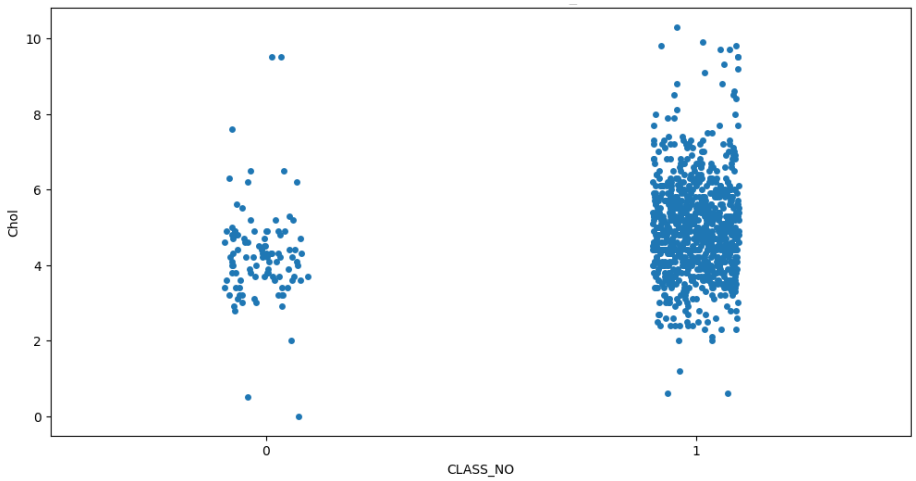


Figure 5 Strip Plot of Chol by CLASS_NO

3.1.6 Inter-Feature Correlation Structure

The Pearson correlation matrix that has been calculated among all clinical, biochemical, and demographic variables is shown in Figure 6 below. Regarding the diabetes class outcome variable (CLASS_NO), the strongest positive relationships have been found for BMI ($r = 0.54$) and HbA1c ($r = 0.53$), which makes them the most informative individual predictors associated with diabetes class in the data sample. A moderate positive correlation with the class label ($r = 0.39$) was obtained for the age group (AGE_GROUP_NO), which, in turn, confirms the results related to the distribution of classes by age groups (see Section 3.1.3). Among other variables, weak positive correlations were detected for triglycerides ($r = 0.18$) and total cholesterol ($r = 0.16$), whereas no relationship with the class label was discovered for HDL ($r = -0.009$), LDL ($r = -0.003$), VLDL ($r = 0.089$), serum urea ($r = 0.056$), and creatinine ($r = 0.036$).

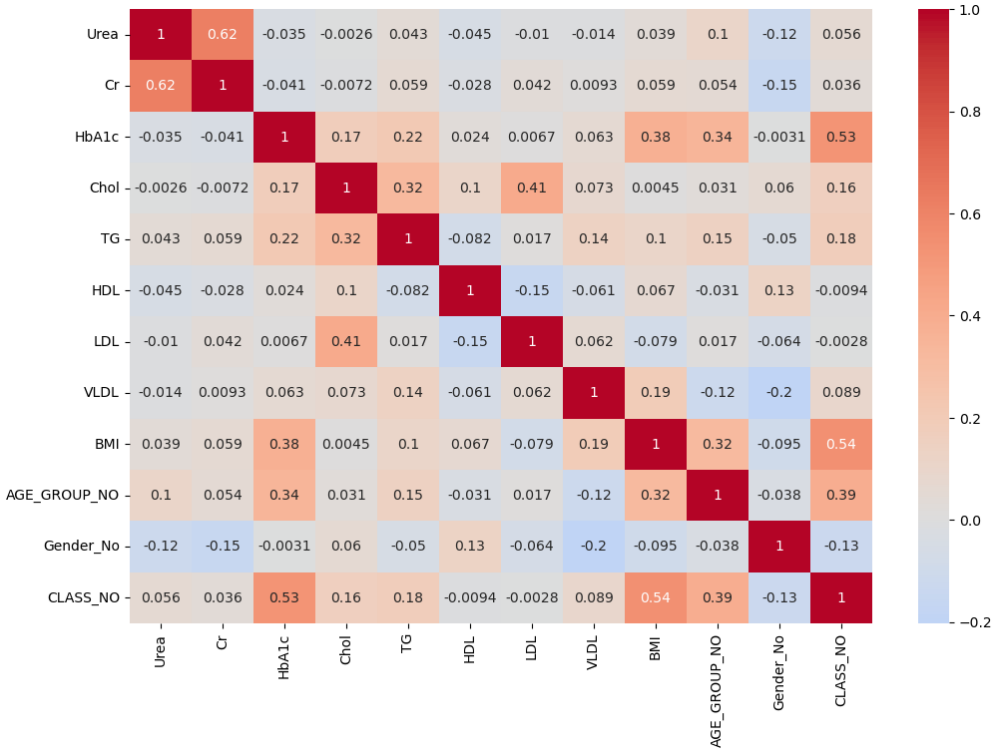


Figure 6 Correlation Matrix of all Variables

Concerning inter-feature relations, the strongest correlation occurred between serum urea and creatinine (r = 0.62). The physiological nature of the relation can be explained by the fact that the two mentioned markers have co-regulatory functions within kidney function assessment, with both values tending to rise simultaneously under nephropathic conditions. Moderate inter-feature correlations could be seen between total cholesterol and LDL (r = 0.41), as well as between total cholesterol and triglycerides (r = 0.32). In addition, HbA1c was moderately correlated with AGE_GROUP_NO (r = 0.34) and BMI (r = 0.38). In terms of HbA1c, such moderate correlations may occur due to the common metabolic stress associated with glycemic dysregulation, increased age, and high body fat percentage. Concerning all features of our dataset, the lack of high-level correlations between pairs suggests an acceptable level of multicollinearity. It makes the analysis easier as the presence of multicollinearity leads to redundant information about features, making learning impossible.

3.2 Machine Learning Classification Performance

3.2.1 Overall Model Performance

After stratified data split, class balancing using SMOTE technique for the training data set, and Z-score standardization for all continuous variables, all six classifiers were trained and later tested on the test data set. Performance results for the machine learning algorithms are provided in Table 3 below. Diabetes (Y) was treated as the positive class. Confusion-matrix values are reported in the order TN, FP, FN, TP.

The relatively higher Recall value of the Random Forest algorithm has specific implications in the diagnosis of diabetes mellitus, where a misdiagnosis leading to an incorrect true positive value – i.e., a false negative – is a significantly more serious mistake than the former because of the severe consequences of undiagnosed hyperglycemia. In this respect, the high F1 score further indicates that the model kept an appropriately balanced Recall and Precision rate, considering the imbalance present in the dataset itself even before being processed through SMOTE. Ultimately, both algorithms prove to be reliable classifiers for distinguishing between diabetic and non-diabetic subjects using the available variables.

Table 3 Overall Classification Performance of the Models

Index	Model Number	Accuracy	Precision	Recall	F1-Score	ROC-AUC	Confusion Matrix (TN, FP, FN, TP)
1	Logistic Regression	0.979	1	0.976	0.988	0.997	21,0,4,165
2	KNN	0.968	10.994	0.964	0.982	0.987	21, 0, 6, 163
3	Naive Bayes	0.968	0.982	0.982	0.982	0.988	18, 3, 3, 166
4	SVM	0.979	0.988	0.988	0.988	0.995	19, 2, 2, 167
5	Random Forest	1	1	1	1	1	21, 0, 0, 169
6	XGBoost	1	1	1	1	1	21, 0, 0, 169

3.2.2 Sensitivity Analysis with and without HbA1c

To investigate whether the strong classification performance was primarily driven by HbA1c, all six classifiers were re-evaluated after excluding HbA1c from the predictor set. Table 4 compares the held-out test performance obtained with and without HbA1c.

Table 4 Sensitivity analysis of model performance with and without HbA1c

Model	Accuracy with HbA1c	Accuracy without HbA1c	ROC-AUC with HbA1c	ROC-AUC without HbA1c
Logistic Regression	0.979	0.953	0.997	0.980
KNN	0.968	0.947	0.987	0.959
Naive Bayes	0.968	0.916	0.988	0.962
SVM	0.979	0.974	0.995	0.985
Random Forest	1.000	0.989	1.000	0.997
XGBoost	1.000	0.989	1.000	0.996

Removal of HbA1c led to decrease in performance of all the classifiers, thus implying that HbA1c played a vital role in class separation. Test accuracies showed biggest drops for Naive Bayes, Logistic Regression, and KNN. On the other hand, Random Forest and XGBoost managed to exhibit extremely high performances even without HbA1c, achieving test accuracies of 98.9%.

From the above observations, it is clear that the high performance seen when HbA1c was used is not due to HbA1c alone. Despite this, the drop in performance implies that HbA1c plays a significant role as a predictor in the dataset. Hence, the perfect performance achieved by using the complete set of features should be handled with care due to the small size of the dataset and separability of biochemical and demographic features.

Stratified five-fold cross-validation further supported the model's stability. The mean accuracy score for Random Forest was 0.985 ± 0.013 with HbA1c, and for XGBoost 0.991 ±

0.009. When HbA1c was not considered, the scores were still relatively high, namely 0.981 ± 0.010 for Random Forest and 0.985 ± 0.010 for XGBoost. So, while both algorithms demonstrated absolutely perfect accuracy on a held-out test set including HbA1c, the performance metrics from cross-validation were somewhat smaller.

3.3 Comparative Model Analysis

A comparative analysis was done to evaluate the performance of six different machine learning algorithms with respect to the task of predicting diabetes from biochemical and demographic attributes. The algorithms considered in this study are Logistic Regression, K-Nearest Neighbors, Naive Bayes, Support Vector Machines, Random Forest, and Extreme Gradient Boosting. Model performance is evaluated based on accuracy, precision, recall, F1 score, ROC-AUC, and confusion matrix.

Of all the models analyzed above, both Random Forest and Extreme Gradient Boosting have shown outstanding performance as evident from their scores of 100% for all evaluation metrics, i.e., accuracy, precision, recall, F1 score, and ROC-AUC. In addition, the two models have shown excellent performance with an optimal confusion matrix without any error rates in classifying the individuals as diabetic and non-diabetic. It can be concluded that ensemble-based learning of trees is quite efficient for dealing with nonlinear relationships among biochemically related diabetes.

Similarly, Logistic Regression and SVM performed extremely well, scoring accuracy of 97.9%, ROC-AUC above 99%, recall rates above 98%, and ROC-AUC of more than 0.99. These models demonstrated minimal misclassifications based on their confusion matrixes, suggesting high stability in their classifications and significant discriminatory capability. From the results obtained from the above-mentioned models, it can be deduced that the dataset had clearly separated patterns that could be classified using either the linear or non-linear techniques.

KNN and Naive Bayes had relatively lower performance, scoring accuracy of 96.8% and an F1-Score of 0.982. However, it demonstrated a very high ROC-AUC of 0.988, which was similar to that of other classifiers. Nevertheless, its confusion matrix suggested relatively more false negative cases. The relatively lower performance of Naive Bayes might be due to the assumption made for this classifier, which assumes that the features such as HbA1c, BMI, cholesterol, and triglycerides are independent.

Of all the classifiers analyzed, Random Forest and XGBoost demonstrated the best accuracy in the current data set. Perfect classification was obtained for both classifiers when HbA1c was used. Yet, the findings should not be taken as proof of universal perfect diabetes classification. Five-fold cross-validation led to slightly lower mean accuracies of 98.5% for Random Forest and 99.1% for XGBoost, implying some variation across the data splits.

In the sensitivity analysis, it was revealed that HbA1c omission decreased the performance of all the models. Yet, Random Forest and XGBoost still had high test accuracy of about 98.9% without HbA1c, which means that the performance of these algorithms was not dependent solely on this variable. This suggests that some demographic and biochemical variables also played a role in the separation of classes.

Since the study is conducted on a relatively small publicly available data set, and there is no independent validation performed, the obtained performance should be considered as specific to the data set only. Validation on bigger clinical data sets is needed to ensure that the classifiers can be used in practice.

In terms of clinical relevance and interpretation of the findings, the results of the current study show that routinely measured demographic and biochemical predictors can serve as valuable sources of information for differentiation of diabetic and non-diabetic observations from the current data set. Among all predictors investigated, HbA1c, BMI, and age demonstrated especially strong associations with the status of patients and served as valuable discriminators.

The good performance of Random Forest and XGBoost in terms of prediction suggests that nonlinear machine learning methods can serve as effective tools for extraction of the patterns from the given data set. However, the results obtained have to be understood within the framework of the current data set. The analysis of the current study was carried out using a relatively small publicly available data set and the validation process using the independent clinical sample was not performed. Thus, the models presented in this study should be understood as exploratory predictive models but not validated clinical diagnostic tools. Additional investigation with more complex data sets is needed to evaluate the clinical applicability of the models presented in this study.

4. Conclusion

In this regard, the present research successfully analyzed the relationship between age-wise prevalence rates of diabetes and several other biochemical indices based on machine learning classification approaches. The results showed that there is a significant association between several clinical variables such as HbA1c, BMI, cholesterol parameters, urea, and age-wise categories and diabetes and contribute to the prediction of diabetes to a significant extent. Based on graphical and analytical analysis, patients diagnosed with diabetes have shown a higher level of HbA1c, high BMI, high metabolic variance, and high concentration in elderly age groups especially above 50 years.

Furthermore, a comparative analysis of the machine learning models provided a validation that machine learning models were highly effective in predicting diabetes. Among the evaluated models, Random Forest and XGBoost achieved the highest held-out test performance, with both models reaching 100% accuracy when HbA1c was included. However, five-fold cross-validation produced slightly lower mean accuracies of approximately 98.5% for Random Forest and 99.1% for XGBoost, indicating some variation across data partitions. In the sensitivity analysis excluding HbA1c, both models retained a test accuracy of approximately 98.9%, suggesting that their strong classification performance was not solely dependent on HbA1c.

The study demonstrates the possible clinical significance of combining routine biochemical parameters along with machine learning models to diagnose early, estimate personal risks, and optimize medical decisions. Predictive models could be highly useful in terms of preventive healthcare because it will be easier to identify people who have the highest risk before any severe diabetic complications occur.

In summary, the present work proves that biochemical testing combined with age-wise data processing and machine learning models represent an efficient method for predicting diabetes. Further investigations could increase the efficiency of the suggested method through introducing multicenter datasets, additional variables related to lifestyles, and other measures.

Acknowledgement

The author acknowledges the use of open-source machine learning tools, publicly available resources, and the Google Colab platform for conducting this research work.

Funding

The author(s) received no specific funding for this work.

Author Contribution

Author 1 prepared the literature review, developed the research methodology, and conducted the results analysis and discussion. Author 2 reviewed the literature review and provided constructive feedback on the manuscript. Author 3 contributed to the methodological review, interpretation of the results, and critical revision of the manuscript for intellectual content. All authors have reviewed and approved the final version of the manuscript.

Declaration of generative AI in the writing process

During the preparation of this work, the authors used an AI-assisted language tool solely to improve the readability and language quality of the manuscript. The authors subsequently reviewed and edited the content and took full responsibility for the accuracy, originality, and integrity of the final manuscript.

Data Availability/Supplementary Materials

The dataset analysed during the current study is publicly available in the Kaggle Multiclass Diabetes Dataset repository at: [https://www.kaggle.com/datasets/yasserhessein/multiclass-diabetes-dataset]. The analysis code used in this study is available from the corresponding author upon reasonable request.

Ethics Statement

Given that the dataset employed for this study was obtained from an open-access database, and did not include any personally identifiable patient data, ethics committee approval was deemed unnecessary, and the analysis was done entirely in line with best data practices without disclosing the identity of any individual patient through these records.

Conflict of Interest

The authors have no conflicts of interest to declare.

References

Abnoosian, K., Farnoosh, R., & Behzadi, M. H. (2023). Prediction of diabetes disease using an ensemble of machine learning multi-classifier models. *BMC Bioinformatics*, *24*(1). https://doi.org/10.1186/s12859-023-05465-z

Ahmed, A., Khan, J., Arsalan, M., Ahmed, K., Shahat, A. A., Alhalmi, A., & Naaz, S. (2025). Machine Learning Algorithm-Based Prediction of Diabetes Among Female Population Using PIMA Dataset. *Healthcare (Switzerland)*, *13*(1). https://doi.org/10.3390/healthcare13010037

Ayoade, O. B., Shahrestani, S., & Ruan, C. (2025). Machine Learning and Deep Learning Approaches for Predicting Diabetes Progression: A Comparative Analysis. *Electronics (Switzerland)*, *14*(13). https://doi.org/10.3390/electronics14132583

Feng, X., Cai, Y., & Xin, R. (2023). Optimizing diabetes classification with a machine learning-based framework. *BMC Bioinformatics*, *24*(1). https://doi.org/10.1186/s12859-023-05467-x

Lee, H., Park, M.-B., & Won, Y.-J. (2025). AI Machine Learning–Based Diabetes Prediction in Older Adults in South Korea: Cross-Sectional Analysis. *JMIR Formative Research*, *9*, e57874–e57874. https://doi.org/10.2196/57874

Magliano, D. J., Boyko, E. J., & committee, I. D. A. 10th edition scientific. (2021). IDF DIABETES ATLAS. *IDF DIABETES ATLAS [Internet]. Brussels*, 1–141. https://www.ncbi.nlm.nih.gov/books/NBK581934/

Masood, S., Al-Bashrawi, M. A., Khan, M. A., & Dwivedi, Y. K. (2026). From data to diagnosis: a systematic review on AI-driven approaches to diabetes prediction. *Artificial Intelligence Review*, *59*(3). https://doi.org/10.1007/s10462-025-11485-3

Multiclass Diabetes Dataset. (n.d.). Retrieved September 27, 2026, from https://www.kaggle.com/datasets/yasserhessein/multiclass-diabetes-dataset

Parimala, G., Kayalvizhi, R., & Nithiya, S. (2023). Diabetes Prediction using Machine Learning. *2023 International Conference on Computer Communication and Informatics, ICCCI 2023*. https://doi.org/10.1109/ICCCI56745.2023.10128216

Patidar, P. K., Shiwani, S., & Garg, S. (2025). Prediction of Diabetes using Machine Learning Classifiers with Polar Bear Optimization. *Procedia Computer Science*, *258*, 1338–1347. https://doi.org/10.1016/j.procs.2025.04.367

Rahman, F., Hossain, S., Tiang, J. J., & Nahid, A. Al. (2025). Diabetes Prediction Using Feature Selection Algorithms and Boosting-Based Machine Learning Classifiers. *Diagnostics*, *15*(20). https://doi.org/10.3390/diagnostics15202622

Sampath, P., Elangovan, G., Ravichandran, K., Shanmuganathan, V., Pasupathi, S., Chakrabarti, T., Chakrabarti, P., & Margala, M. (2024). Robust diabetic prediction using ensemble machine learning models with synthetic minority over-sampling technique. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-78519-8

Shetty, R., Paul, A., Kumar, A. P., & Rai, S. (2026). Early prediction of diabetes using a reduced and interpretable feature set with a scalable machine learning framework. *Discover Artificial Intelligence*, *6*(1). https://doi.org/10.1007/s44163-026-01152-z

Tasin, I., Nabil, T. U., Islam, S., & Khan, R. (2023). Diabetes prediction using machine learning and explainable AI techniques. *Healthcare Technology Letters*, *10*(1–2), 1–10. https://doi.org/10.1049/htl2.12039

Wee, B. F., Sivakumar, S., Lim, K. H., Wong, W. K., & Juwono, F. H. (2024). Diabetes detection based on machine learning and deep learning approaches. *Multimedia Tools and Applications*, *83*(8), 24153–24185. https://doi.org/10.1007/s11042-023-16407-5

Xiao, Z., Wang, M., Zhao, Y., & Wang, H. (2025). A Biomarker-Driven and Interpretable Machine Learning Model for Diagnosing Diabetes Mellitus. *Food Science and Nutrition*, *13*(5). https://doi.org/10.1002/fsn3.70234

Yesupatham, A., Das, R., Bharani, G., Shaikmeeran, M., & Saraswathy, R. (2025). Artificial intelligence model as a tool to predict prediabetes. *Scientific Reports*, *15*(1). https://doi.org/10.1038/s41598-025-23227-0