Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 24 June 2026
Revised 29 July 2026
Accepted 13 Aug 2026

Natural Resources for Human Health 2026, Vol. 6 (16s): 1557–1572

KEYWORDS:
Diabetes Risk Stratification; Explainable Artificial Intelligence (XAI); Ensemble Learning; CatBoost; SHAP Explainability; Multi-class Classification; Feature Importance Analysis; Early Disease Detection.

An Ensemble Machine Learning Framework for Early Prediction and Risk Classification of Diabetes Using Health Indicators

Gurdeepika Kaur[1], Harsh Sadawarti[2]

[1,2]Department of Computer Science and Application, Faculty of Engineering Technology and Computing , Desh Bhagat University, Mandi Gobindgarh, Punjab, India
[2]Department of Computer Science and Engineering, Faculty of Engineering Technology and Computing , Desh Bhagat University, Mandi Gobindgarh, Punjab, India
Email Id: gurdeepikakaur1804@gmail.com1, vc@deshbhagatuniversity.in2

ABSTRACT: Diabetes is a major chronic disease that, if identified early and appropriately risk-assessed, can be prevented from causing serious complications. In this study, an exhaustive machine learning framework to identify the risk of diabetes and stratification by physiological, metabolic and lifestyle parameters was developed. The data set was publicly available and was comprised of demographic, clinical and behavioral attributes which were used to label people as Low Risk, Prediabetes, and High Risk. It comes with data preprocessing, exploratory analysis, feature selection, dimensionality reduction, explainable artificial intelligence (XAI), and several machine learning and ensemble models. The body mass index (BMI) followed by Waist circumference, fasting glucose, HbA1c, triglycerides and physical activity were important features identified by feature importance analysis using the Mutual Information and SHAP methods. Classification accuracy, precision, recall, F1 score, log loss and ROC-AUC were used to assess several classifiers. All ensemble and boosting techniques were better than standard techniques, and they performed very well in comparison. CatBoost outperformed remaining models in the overall accuracy, precision, recall, F1 score, and log loss with the scores of 94.0%, 92%, 93%, 92% and 0.1501, respectively, and achieved the best weighted ROC-AUC of 99.25%. The results show the feasibility of using ensemble learning and XAI in early detection and personalized diabetes risk management.

1. INTRODUCTION

Diabetes mellitus represents the most important chronic metabolic diseases of many millions of people and has emerged as an important public health issue. Diabetes is characterized by elevated blood glucose levels resulting from decreased insulin production, insulin resistance, or both (Banday et al. 2020). The incidence of diabetes has risen significantly over the last several decades worldwide, and a number of factors have contributed to this rise, including: more sedentary lifestyle, poor nutrition, obesity, increased life expectancy and genetic susceptibility. The complications of diabetes can be very serious if the condition goes unrecognized or is not treated properly, such as kidney failure, retinopathy, cardiovascular disease, neuropathy, and death. Thus, early detection is crucial for immediate action, effective disease management, and reducing the burden of care (Hossain et al. 2024; Liu et al. 2022).

Diagnosis of diabetes is mainly determined through clinical evaluation and laboratory tests such as OGTT, FBS, and HbA1c measurements. Although these methods are useful in diagnosis they may detect the disease when much of the physiological changes have taken place. Moreover, healthcare systems are producing vast amounts of patient information such as personal, physiological, behavioral, and lifestyle-related information, which can be used to make use of predictive analytics (Gourlay et al. 2024). Artificial intelligence (AI) and machine learning (ML) are rapidly evolving technologies which offer innovative opportunities to develop intelligent decision-support systems that can detect disease risks earlier (Khosravi et al. 2024).

Machine learning methods have shown exceptional performance in several healthcare-related fields such as disease diagnosis, prognosis, medical image analysis, patient monitoring and providing personalized recommendations for treatment. When it comes to predicting diabetes, ML algorithms can take into account the intricate interactions between a myriad of risk factors and uncover patterns that might not be apparent from traditional statistical models (Afsaneh et al. 2022). Various studies have used machine and deep learning for predicting diabetes occurrence. While these methods have yielded good results, there are remaining issues to be resolved related to increasing predictive accuracy, risk categorization for multiple classes, top health indicators and clinical interpretability of models (Saxena et al. 2023). Ensemble learning is an emerging technique that has gained popularity over the last few years for using multiple base learners to reduce variance, bias, and increase generalization skills of predicting. In health care prediction, Gradient Boosting, XGBoost, LightGBM, CatBoost, Bagging, Voting, and Stacking ensembles have been shown to be the most effective techniques for various applications. Furthermore, XAI techniques has paved the way for understanding the rationale behind model predictions, enhancing transparency, trustworthiness and clinical applicability (Khan et al. 2024; Sagenela et al. 2026).

This has motivated the present study to develop a comprehensive machine learning system for predicting the likelihood of diabetes and diabetes stratification based on some physiological and lifestyle health indicators. The proposed framework does not aim to simply classify individuals as either Low Risk or High Risk as is done in traditional binary classification methods, but rather classifies individuals into three clinically relevant risk groups: Low Risk, Prediabetes, and High Risk. The system combines the following techniques: data preprocessing, exploring data properties, assessing feature importance, dimensionality reduction, explainable AI, and multiple base learning algorithms and ensemble learning methods. In addition, the most important aspects of health factors that raise the risk of diabetes are identified by means of Mutual Information and explainability methods based on SHAP. The main objective of this study is to build a reliable and accurate predictive mathematical model of diabetes, which is also interpretable so that it can be used to early detect and assess the risk of diabetes.

1.1 Contribution

The study's primary contributions are summarized as follows:

- Development of a comprehensive machine learning framework to predict the risk of multiple classes of diabetes and stratification.
- Comparative evaluation of a variety of traditional machine learning, ensemble learning, and boosting-based algorithms.
- Identification of key physiological and lifestyle factors influencing diabetes risk through feature importance and explainability analysis.
- Integration of XAI techniques to improve transparency of model and its interpretability.
- Demonstration of the effectiveness of ensemble learning methods for supporting early disease detection and preventive healthcare decision-making.

2. RELATED WORK

Previous research has investigated various AI methods however, some cautions must be noted about the early risk stratification, the interpretability of the models and the possible incorporation of lifestyle-related factors. Many image-based techniques have been adopted to detect DR and other diabetic related complications. Singh et al. (2025) performed the comparative study of various deep neural networks such as ResNet50, InceptionV3, MobileNet, DenseNet121, NASNetMobile, Xception and VGG19 using the dataset called ODIR. The experimental findings revealed that ResNet50 had the best performance of 99.74% accuracy. Likewise, Akshaya et al. (2026) designed a framework for diabetic retinopathy classification using convolutional neural network (CNN) from retinal fundus images. While these methods were successful in their performance, the methods require special imaging equipment and are primarily applicable once retinal complications have already formed, and not for early prediction of the disease. There are multiple studies that have concentrated on machine learning methods utilizing tabular health information. Zaki and Imaduddin (2026) used Diabetes Health Indicators dataset to compare various learning models, concluding that Logistic Regression was the best screening model. Abdulhadi et al. (2021) [12] tested six traditional classifiers on the PIMA Indian Diabetes Dataset and concluded that Random Forest achieves maximum accuracy of 82%. Similarly, Khaleel, Al-Bakry (2021) compared the classifiers, and found Logistic Regression with 94% accuracy. However, these studies focused mainly on the binary classification of diabetes prediction and only a few of the algorithms were studied. Deep Learning

techniques have also shown good results..In another study, Singh et al. (2025) developed DiaXplain, a hybrid model combining convolutional neural networks (CNNs), XGBoost, and SHAP based explainability to improve model transparency and interpretability. These methods not only improved predictive accuracy and explainability, but also were confined to binary classification and lacked the ability to stratify diabetes risk. To alleviate the small sample size problem, García-Ordás et al. (2021) used sparse autoencoders and variational autoencoders to augment the number of features and synthetic samples, with 92.31% accuracy. In addition, Wee et al. (2023) and Sharma and Shah (2021) conducted comprehensive review papers that revealed the same challenges in the field; overreliance on the PIMA dataset, poor evaluation metrics, limited interpretability and limited consideration of lifestyle related factors.

There have been multiple studies that have shown that multiple learning techniques are effective in predicting diabetes; however there are a number of limitations. In fact, most of the existing methods consider diabetes prediction as a binary classification problem, which neglects the clinically important prediabetes stage, in which early action to prevent diabetes progression is effective. Also, most studies test a small number of algorithms and focus on clinical/imaging metrics, rather than lifestyle factors like physical activity, diet, sleep and stress. Furthermore, model evaluation is frequently limited to accuracy, and interpretability and full model validation and assessment are underemphasized. To overcome these drawbacks, the current study aimed to present a thorough and easily-interpreted multi-class diabetes risk stratification framework. The suggested approach combines physiological, metabolic and lifestyle-related markers to determine Low Risk, Prediabetes and High Risk status. A variety of machine learning and ensemble learning models are explored and tested with several performance metrics to guarantee a strong predictive performance. Furthermore, XAI techniques are incorporated to identify the most significant risk factors, promoting transparency and fostering trust in the early detection of diabetes and personalized prevention strategies.

3. METHODOLOGY

In this section, a comprehensive machine learning based framework for diabetes risk prediction and stratification has been developed. The proposed methodology involves multiple stages as shown in Figure 1, where persons are classified as Low Risk, Prediabetes and High Risk.



Figure 1. Workflow of the Proposed Framework for Diabetes Risk Prediction using Machine Learning

3.1. Dataset Description

In this study, the experiments were conducted with the publicly available diabetes risk prediction dataset. The dataset contains a broad range of demographic, physiological, metabolic and lifestyle-related features related to diabetes risk. Features such as age, body mass index (BMI), gender, fasting glucose level, insulin level, HbA1c level, blood pressure, cholesterol level, waist circumference, triglycerides level, daily calorie intake, sugar intake, sleep duration, stress level, physical activity level, and family history of diabetes are included in the dataset. These factors increase the risk of developing and worsening diabetes; some are modifiable, whereas others are non-modifiable (Vishard 2025).

The dataset is composed of three target classes, representing Low Risk, Prediabetes and High Risk, which allows multi-class diabetes risk stratification and not only binary classification. Lifestyle and behavioral parameters are added to the clinical and metabolic parameters, giving a more comprehensive picture of a person's health state. Furthermore, there are numerical and categorical data in the dataset, so many machine learning and ensemble learning algorithms can be tested on the data. Its

physiological and lifestyle characteristics, which are well distributed, enable the development of strong and meaningful predictive models to enable early detection of diabetes, individual risk calculation, and management of preventive health care (Vishard 2025).

3.2. Data Preprocessing and Visualization

The diabetes risk data was preprocessed before development of the model to ensure consistency and compatibility with machine learning algorithms. The data set was reviewed for missing, duplicate, and inconsistent data. Categorical variables were one-hot encoded, while the diabetes outcome was categorized into three classes: Low Risk, Prediabetes, and High Risk. The numerical attributes have different scales and ranges so feature standardization was carried out using z-score normalization.

One-hot encoding was applied to categorical variables where the target variable was encoded into three classes as: y = 0 as Low Risk, 1 as Prediabetes, 2 as High Risk. The data was then cleaned and processed to ensure its suitability for further analysis and modeling using machine learning techniques.

Later, Exploratory Data Analysis (EDA) was carried out to get insights into the distribution of the diabetes risk categories and the health-related features. As illustrated in Figure 2, the distribution of the data class shows that there are three classes in the data: Low Risk, High Risk and Prediabetes, with more instances of Low Risk and High Risk than Prediabetes. This distribution shows a moderate class imbalance, indicating that there is a need to consider model performance metrics other than accuracy.

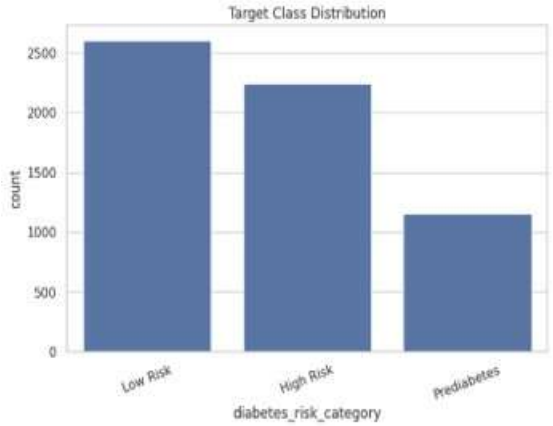


Figure 2. Distribution of Samples Across Diabetes Risk Categories

The histograms of the numerical features in Figure 3 show patterns of the health indicators and their variability. The distribution of age is fairly uniform while the distribution of BMI, blood pressure, cholesterol level and triglycerides level are approximately bell-shaped. The distributions of daily calorie intake insulin level, fasting glucose level, and sugar intake are all positively skewed with a significant number of people having much higher values. HbA1c values tend to cluster at the normal level, and few values appear at extreme levels, while sleep hours are clustered around 7–8 hours. Stress levels are distributed in a discrete manner from low to high and waist circumference is distributed more widely with higher numbers in the middle range. The observations can be used to make important decisions on what features should be included in the dataset and further help in feature reduction and model building.

Natural Resources for Human Health
eISSN: 2583-1194

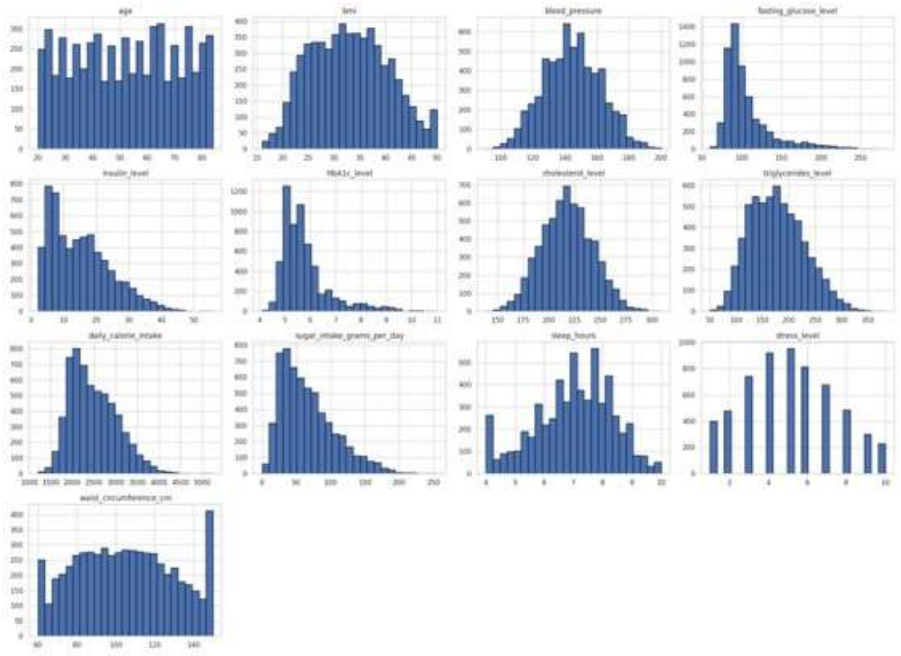


Figure 3. Histograms Illustrating the Distribution of Numerical Health Features in the Diabetes Risk Dataset

Additional exploratory analyses were done to investigate the relationship between categorical variables and diabetes risk categories in Figure 4. The distribution of diabetes risk by gender shows a similar distribution for both sexes, which suggests that sex alone is not a strong determinant of diabetes risk. There is, however, a clear relationship between physical activity level and diabetes risk. People who have lower levels of physical activity are mainly clustered in the High Risk group and people with higher levels of physical activity are mainly clustered in the Low Risk group. Moderate activity levels are evenly spread across the Low Risk and Prediabetes groups, suggesting the protective effect of physical activity against the progression of prediabetes to Type 2 diabetes disease.

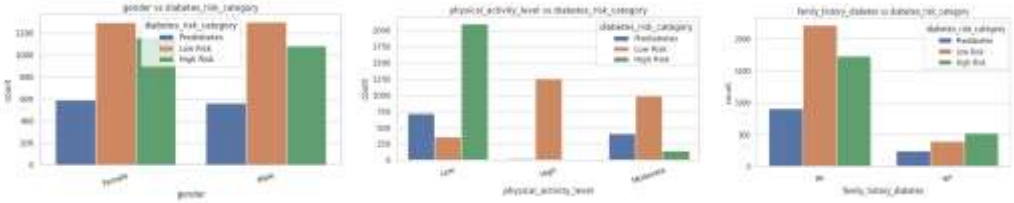


Figure 4. Relationship Between Categorical Variables and Diabetes Risk Categories

Likewise, family history is known to be positively associated with High Risk and the proportion of such cases is relatively higher among those with positive family history than those with a negative family history. Low Risk individuals make up the majority of both groups, but a genetic predisposition seems to make some people more susceptible to diabetes. These observations clearly indicate that the type of lifestyle and genetic predispositions play an essential role in the diabetes risk and therefore are useful for future feature selection and model building.

The correlation heatmap (Figure 5) shows the relationships between the numerical attributes and the diabetes risk category. There are several positive correlations that occur between the physiological and metabolic variables. The level of fasting glucose and the level of HbA1c also have a very high correlation (0.98), meaning that they are strongly related to blood glucose regulation. Likewise, BMI has high positive correlations with waist circumference (0.96) and calorie intake per day (0.89), indicating that body mass and abdominal obesity are correlated with higher calorie intake. Triglycerides level also shows significant correlations with sugar intake (0.85) and waist circumference (0.76), indicating that diet plays a significant role in metabolic health.

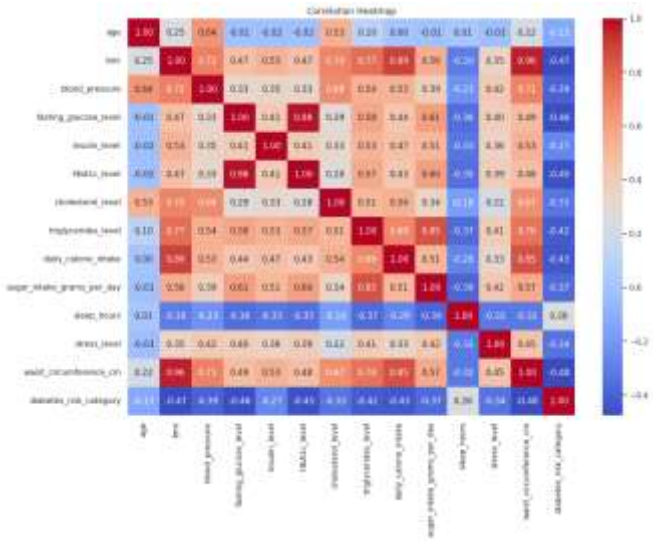


Figure 5. Correlation Heatmap of Physiological, Metabolic, and Lifestyle Features in the Diabetes Risk Dataset

There are moderate negative correlations between the diabetes risk category and BMI (-0.47), fasting glucose level (-0.46), HbA1c level (-0.45), daily calorie intake (-0.43), triglycerides level (-0.42) and waist circumference (-0.48); this means that changes in these factors are linked with considerable changes in diabetes risk. Sleep hours, on the other hand, have a weak positive correlation with the target variable of interest, and have negative correlations with a number of metabolic indicators, indicating that sufficient sleep may have a protective effect. The variables are weakly correlated with age, and moderately positively correlated with stress level, with blood pressure, fasting glucose level and sugar intake. In general, the heatmap shows that obesity-related parameters, glycemic parameters, lipid parameters and lifestyle parameters are well connected and have an effect on diabetes prognosis.

3.3. Feature Selection

Feature selection was performed using the Mutual Information (MI) technique to quantify the dependency between each input feature and the diabetes risk category. Mutual Information measures the amount of information that one variable provides about another (Aich et al. 2026). For a feature (X) and target variable (Y), the mutual information is defined as

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p\,(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \quad (1)$$

where (p(x,y)) defines the joint probability distribution of (X) and (Y). Stronger dependence between the feature along with the target variable is indicated by a with greater mutual information values indicating greater predictive importance.

BMI (0.5867), waist circumference (0.3781), triglycerides level (0.3781), daily calorie intake (0.3587), fasting glucose level (0.3385), and low physical activity level (0.3083) received the highest score. HbA1c level and blood pressure also had relatively high scores highlighting the importance of metabolic and cardiovascular risk factors in diabetes risk assessment. The prediction task was on the other hand less dependent on age, family history of diabetes and gender. These results suggested that obesity related measurements, glycemic markers, dietary behaviors, and physical activity were most informative predictors, and these were then used for model development

The features selected by the Mutual Information method were then clustered by dimensionality reduction to investigate the structure of the data and identify the ground truth separability of the three diabetes risk categories. Those features that had the highest Mutual Information values were kept for further analysis.

Later, the feature space have been projected to a lower dimensional space using Principal Component Analysis (PCA) while maximizing the variance. The two-dimensional PCA projection of Figure 6 indicates a clear separation of Low Risk and High Risk samples mainly in the direction of the first principal component (PC1).

Natural Resources for Human Health
eISSN: 2583-1194

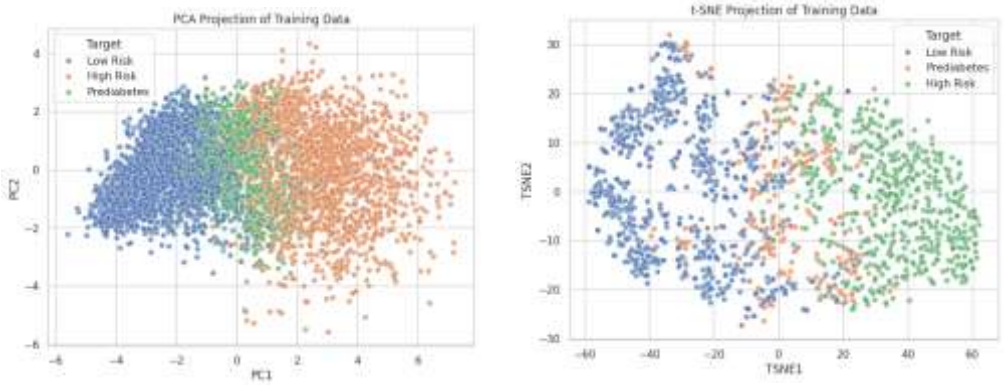


Figure 6. PCA and t-SNE visualization of extracted features

Low Risk examples cluster on the negative side of PC1 and samples of High Risk cluster on the positive side. Prediabetes samples fall into an intermediate zone and overlap with both classes, indicating their intermediate physiological and metabolic status. This observation suggests that the selected features are effective at capturing the key sources of variation of diabetes progression. The selected features were further examined for their nonlinear structure by using t-distributed Stochastic Neighbor Embedding (t-SNE). The Low Risk and High Risk classes are relatively well separated from each other, and the separation between low risk and high risk classes is better than between the classes obtained by PCA; and the Prediabetes samples are spread between low risk and high risk classes, as shown in Figure 6. Prediabetes instances are distributed across neighbouring regions, indicating that this class can have characteristics similar to those of healthy and high-risk individuals, which makes its classification more difficult.

3.4. Applied Classifiers

To comprehensively investigate the predictive capability of different learning paradigms for diabetes risk stratification, a diverse set of classifiers encompassing conventional machine learning, tree-based methods, boosting algorithms, neural networks, and ensemble techniques was employed. All models were trained on the selected features obtained through Mutual Information analysis and were designed to classify individuals into Low Risk, Prediabetes, and High-Risk categories.

Conventional machine learning models, including Logistic Regression (LR), Decision Tree (DT), K-Nearest Neighbors (KNN), Naïve Bayes (NB), and Support Vector Machine (SVM), were first considered as baseline classifiers. Logistic Regression estimates the posterior probability of class membership through the sigmoid function (Abdalrada et al. 2024; Azad et al. 2020; Alnowaiser 2024)

$$P(y = 1|x) = \frac{1}{1+e^{-(w^T x+b)}} \quad (2)$$

where (w) and (b) denote the model parameters. Decision Trees recursively partition the feature space based on impurity measures, whereas KNN assigns a class according to the majority vote of the nearest neighbors. Naïve Bayes utilizes Bayes' theorem (Azad et al. 2020; Alnowaiser 2024)

$$P(Ck \mid X) = P(X)P(X \mid Ck)P(Ck) \quad (3)$$

assuming conditional independence among features. SVM constructs an optimal hyperplane by maximizing the margin between classes and can be expressed as (Bilal et al. 2024)

$$f(x) = w^T x + b \quad (4)$$

These traditional approaches provide baseline performance and facilitate comparison with more sophisticated ensemble techniques.

Tree-based models, namely Random Forest (RF) as well as Extra Trees (ET), were employed to capture nonlinear relationships among physiological, metabolic, and lifestyle variables. Random Forest generates multiple decision trees using bootstrap samples and aggregates their predictions through majority voting, (Noviyanti & Alamsyah 2024; Arya et al. 2022)

$$haty = \text{mode}\big(h_1(x), h_2(x), \ldots, h_n(x)\big) \quad (5)$$

where (h_i(x)) denotes the prediction of the (i^{th})tree. Extra Trees further increase diversity by selecting random split points, thereby reducing variance and improving generalization.

Boosting algorithms iteratively improve weak learners by focusing on previously misclassified samples. Gradient Boosting minimizes the loss function through additive optimization, (Sai et al. 2023)

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \quad (6)$$

where ($h_m(x)$) implies the weak learner and (γ_m) denotes the learning rate. AdaBoost updates sample weights according to

$$w_i^{(t+1)} = w_i^{(t)} e^{\alpha_t I(y_i \neq h_t(x_i))} \quad (7)$$

to emphasize difficult instances. Advanced gradient boosting frameworks including XGBoost, LightGBM, and CatBoost were subsequently employed to efficiently learn complex interactions among health indicators. These models incorporate regularization and optimized tree construction mechanisms, making them particularly suitable for structured tabular datasets such as diabetes risk data (Gündoğdu 2023; Jaiswal & Gupta 2022).

A Multilayer Perceptron (MLP) was employed to model nonlinear relationships among the selected features. Each neuron processes a weighted sum of input signals and then passes the result through an activation function (Mohapatra et al. 2018)

$$a_j = \phi\left(\sum_{i=1}^{n} w_{ij}\, x_i + b_j\right) \quad (8)$$

where ($\phi(\cdot)$) denotes the activation function. Through multiple hidden layers, the MLP learns complex patterns associated with diabetes progression and risk stratification.

To further enhance predictive performance and robustness, Bagging, Voting Ensemble, and Stacking Ensemble models were investigated. The Bagging classifier combines predictions from multiple bootstrap-based learners through majority voting. The Voting Ensemble aggregates predictions from different classifiers as (Kudin et al. 2026; Kumari et al. 2021)

$$\hat{y} = \arg\max_c \sum_{i=1}^{N} w_i\, P_i(c) \quad (9)$$

where ($P_i(c)$) defines the probability assigned by the (i^{th})classifier to class (c). Furthermore, the Stacking Ensemble integrates multiple base learners and employs a meta-classifier to generate the final prediction,

$$\hat{y} = g\left(h_1(x), h_2(x), \ldots, h_n(x)\right) \quad (10)$$

where ($g(\cdot)$) = the meta-learner. By combining complementary strengths of diverse classifiers, ensemble learning aims to improve accuracy and generalization capability.

Collectively, these classifiers provide a comprehensive framework for evaluating various learning paradigms on diabetes risk assessment.

4. RESULTS ANALYSIS

This part of paper analyses multiple classifiers for multiclass diabetes risk classification using physiological and lifestyle health indicators. These variables collectively represent metabolic, behavioral, and demographic factors associated with diabetes progression and provide the basis for evaluating the effectiveness of many methods for machine learning for risk stratification of diabetes in its early stage.

Table 1 shows the comparative performance of the machine learning models developed for diabetes risk classification with multiclass classification using physiological and lifestyle health indicators.

Model	Acc (%)	Log Loss	Pre (%)	Rec (%)	F1 (%)
Logistic Regression	90.67	0.2192	88	88	88
Decision Tree	84.33	5.6468	80	80	80
Random Forest	89.92	0.2530	87	87	87
Extra Trees	89.50	0.2667	87	86	87

Gradient Boosting	90.75	0.2389	88	89	88
KNN	84.83	0.7913	81	80	80
Naïve Bayes	77.17	1.3578	72	70	70
SVM	90.00	0.2362	87	87	87
AdaBoost	87.08	0.9255	84	83	84
Bagging	90.25	0.2664	88	88	88
LightGBM	93.50	0.1847	92	92	92
CatBoost	94.00	0.1501	92	93	92
MLP Neural Network	93.17	0.4054	91	91	91
Voting Ensemble	93.75	0.1647	92	92	92
Stacking Ensemble	93.75	0.1584	92	92	92
PCA + Random Forest	89.42	0.2819	87	85	86
XGBoost	93.08	0.1611	91	91	91

Table 1. Comparing machine learning classifiers' performances

It can be inferred from the findings that the ensemble-based methods generally performed better than the conventional machine learning algorithms, which shows that the relationships between the health indicators are complex and nonlinear. CatBoost model had the highest accuracy with 94.00% and the lowest log loss of 0.1501, and macro precision, recall, and F1-scores of more than 92% among all models. These remarkable results indicate that CatBoost is able to capture the interactions between metabolic and lifestyle factors and various stages of diabetes risk effectively. LightGBM, Voting Ensemble, Stacking Ensemble, MLP Neural Network and XGBoost also showed very good performance in terms of classification, having accuracies of more than 93%. The high macro precision, recall and F1-scores reflect well-balanced performances in all three risk classes. This is especially relevant to the data set since the Prediabetes class is more difficult to separate from the Low-Risk and High-Risk classes, as it has a number of common attributes. The traditional machine learning classifiers like RF, SVM, LR, and Bagging classifiers achieved accuracies in the range of 77–89%, indicating that the chosen health indicators contain adequate discriminatory information for assessing diabetes risk. These models, however, were somewhat less effective in capturing the complex relationship which existed within the dataset, than the boosted approaches. However, the accuracies and macro performance measures were comparatively low for Decision Tree, KNN, and Naïve Bayes. Specifically, Naïve Bayes had the lowest accuracy score of 77.17% and the highest log loss of 1.3578, suggesting that the independence assumptions of the Naïve Bayes model are not appropriate for this dataset as there are several features that are highly correlated, most notably BMI, fasting glucose, HbA1c, triglycerides, and waist circumference. Likewise, the low performance of Decision Tree and KNN indicates that there is insufficient capacity of simple decision boundaries to effectively classify the three diabetes risk categories.

Table 2 shows the class-wise performance of top five models from Table 1 as CatBoost, Voting Ensemble, Stacking Ensemble, LightGBM and MLP Neural Network.

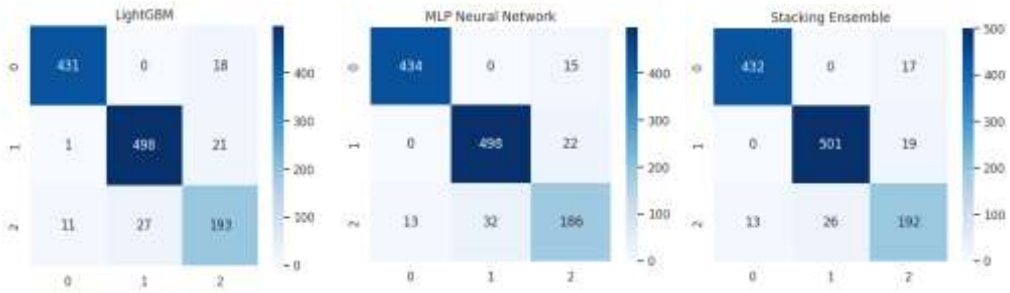
Model	Class	Pre (%)	Rec (%)	F1 (%)
CatBoost	High Risk	98	96	97
	Low Risk	96	96	96
	Prediabetes	83	86	85
Voting Ensemble	High Risk	97	96	97
	Low Risk	95	96	96

	Prediabetes	84	84	84
Stacking Ensemble	High Risk	97	96	97
	Low Risk	95	96	96
	Prediabetes	84	83	84
LightGBM	High Risk	97	96	97
	Low Risk	95	96	95
	Prediabetes	83	84	83
MLP Neural Network	High Risk	97	97	97
	Low Risk	94	96	95
	Prediabetes	83	81	82

Table 2. Class-wise comparison of machine learning classifiers for physiological sensor–based disorder

The overall accuracy of these models was greater than 93% as seen in Table 1, reflecting their ability to predict diabetes risk in multiple classes. A detailed analysis is presented in Table 2, which looks at the performance of each model for each of the individual risk categories (High Risk, Low Risk, and Prediabetes). For the category 'High Risk' all five models performed very well, with precision, recall and F1 score in the range of 96% - 98%. This suggests that there are physiological and metabolic differences among individuals with severe diabetic conditions (fasting glucose, HbA1c, triglycerides, BMI, and waist circumference) that make them more easily identifiable from the other groups. The same is true for the Low-Risk category as all its F1 scores are 95% or 96%, indicating that the feature patterns of healthy individuals are fairly well-distributed across the dataset. Yet, a significant decrease in performance is observed for the Prediabetes class. The best-performing models achieved F1-scores ranging from 82% to 85%, but this is not as high as was seen for the High-Risk and Low-Risk categories. The lower performance for Prediabetes may be due to common features of prediabetic and other two classes. Other factors don't have distinct cut-off points between the low-, medium-, and high-risk categories and therefore are more difficult to classify precisely. The CatBoost model achieved the highest class-wise recall (86%) and F1 score (85%) in the model with best performance among 5 models, especially in the Prediabetes category. This finding is consistent with the overall performance results presented in Table 1. Ensemble voting and Stacking Ensemble also showed a good balance of performance in all three classes, illustrated by the high values of around 92%. The LightGBM and MLP Neural Network models behaved similarly, with the MLP model having a slightly lower recall of the Prediabetes class, but a higher F1-score of 82%. The relatively small scores for Prediabetes also underscore the difficulty of defining the intermediate-risk group, a phenomenon which reinforces the need for more sophisticated ensemble classification methods to effectively stratify individuals for diabetes risk and provide early preventive health care.

The corresponding confusion matrices of the top-performing models, LightGBM, MLP Neural Network, CatBoost, Voting Ensemble, and Stacking Ensemble are analysed in Figure 7 to further understand the distribution of classification errors and the effectiveness of these models.



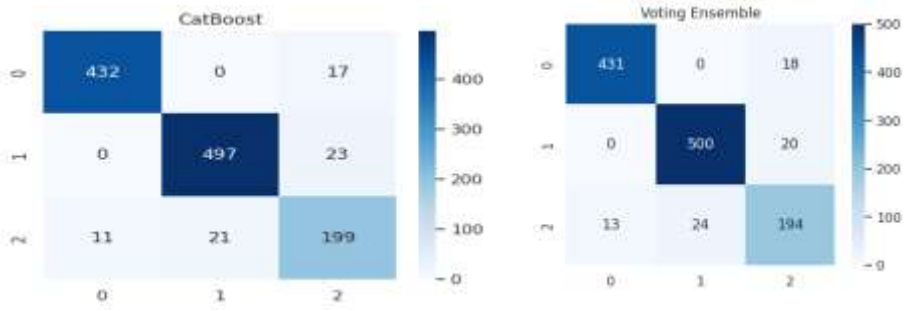
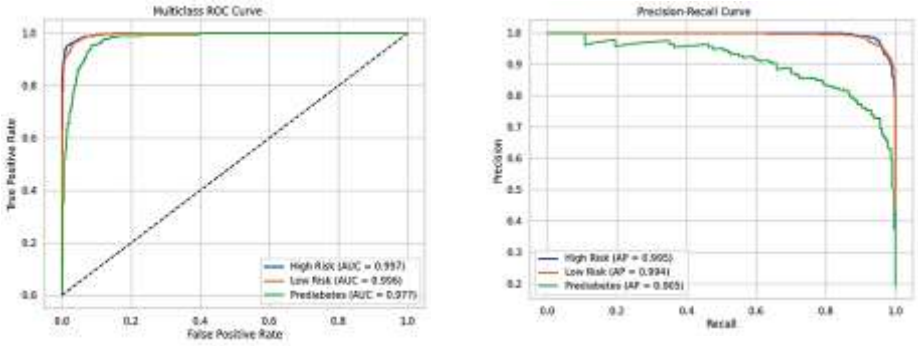


Figure 7. Confusion matrices of the top-performing machine learning models for multiclass diabetes risk classification

In all cases, the highest number of values are along the diagonal, reflecting the high accuracy in classification of most samples. The High-Risk and Low-Risk categories were identified with very high accuracy, indicating that there are distinct metabolic profiles of these groups with differences in few varaibles. The best overall classification capability was achieved by CatBoost model with 432 High-Risk samples, 497 Low-Risk samples, and 199 Prediabetes samples correctly classified. Other algorithms like LightGBM, Voting Ensemble, and Stacking Ensemble also performed well in predicting the classes of Prediabetes and High-Risk, with over 190 and 430 cases respectively correctly classified. The results obtained by the MLP Neural Network were similar, but this model made slightly more misclassifications for Prediabetes class. Another pattern that can be seen in all confusion matrices is that most errors occur with the Prediabetes category. A few Prediabetes data were mislabeled as either High or Low-Risk. The findings are also similar to the class-wise performance results shown in Table 2, revealing that the Prediabetes class has lower values than the other two classes. This reflects the inherent characteristics of the dataset because the people in the Prediabetes group are in between healthy and diabetic people. As a result, some of the features demonstrate overlapping distributions making the discrimination of this class more challenging. In general, the results of the confusion matrices are in line with the observations from the Tables 1 and 2 and show that boosting and ensemble learning methods are good methods to capture complex relationships among physiological and lifestyle health indicators. Their performance in reducing classification errors and still achieving a balanced performance across all three diabetes risk categories make them suitable for early diabetes risk assessment and preventive health-care applications.

Further tests of discriminative ability and probability estimation for the best model were conducted with the multi-class ROC curve, precision-recall curve and calibration curve as shown in Figure 8. The multiclass ROC curves show good separation between the three diabetes risk groups. The High-Risk class's AUC was 0.997, 0.996 for the Low-Risk class and 0.977 for the Prediabetes class. The results suggest that the model was very successful in discriminating between groups of people according to physiological and lifestyle characteristics.



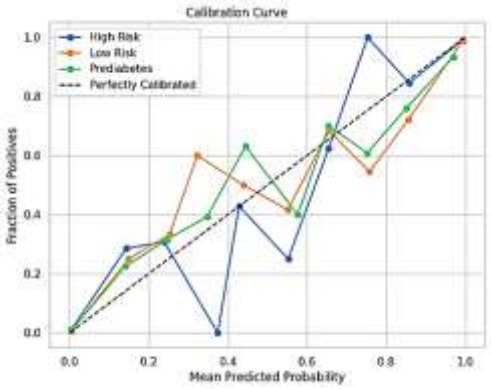


Figure 8. ROC, precision–recall, and calibration curves of the CatBoost model for multiclass diabetes risk classification

The distinction between Prediabetes and the other categories is not as clear because some variables vary gradually during disease progression. The precision-recall curves further demonstrate this. The average precision values of 0.995 and 0.994 obtained for the class High-Risk and Low-Risk respectively, clearly show the excellent predictive capability with very few false positive cases. The average precision of the Prediabetes class was slightly lower at 0.905 but the model had a high precision score across a broader range of recall values. This behavior again highlights the complexity of the identification of subjects with intermediate metabolic characteristics, that is, between healthy and diabetic. The calibration curve is another measure of the reliability of the predicted probabilities besides classification performance. The model's probabilities well approximate the ideal calibration line, especially for the medium and high confidence ranges, suggesting it is a good model for the probability estimates. Minor deviations are seen for the Prediabetes class, indicating some uncertainty in class assignment, as there is some overlap in feature distribution in the dataset. In general, the calibration analysis shows that the estimated probabilities closely match the observed likelihoods, suggesting that the model is calibrated to provide reliable estimates of diabetes risk, beyond just classification. The results of the ROC, precision-recall, and calibration analyses showed that the proposed framework offers excellent class separability and reliable probability estimation for diabetes risk prediction.

After ROC, precision-recall, and calibration analyses, the prediction confidence analysis and decision curve analysis give more evaluation of the reliability and usefulness of the CatBoost model. Figure 9 shows the prediction confidence distribution, class-wise confidence scores and decision curve analysis for a multiclass diabetes risk classification. The prediction confidence distribution reveals that most of the samples correspond to high confidence levels (around 1.0), meaning that the model demonstrates high value of confidence in its estimations for a majority of individuals.

The class-wise confidence analysis shows that the High-Risk and Low-Risk classes are clustered around 1.0, suggesting that they have different physiological and metabolic properties which are well distinguished by the model. Alternatively, the Prediabetes category shows a relatively large range of confidence values for the samples, with some samples scoring lower. This observation is aligned with the confusion matrix and class-wise performance results, in which the class of Prediabetes has relatively low values. This drop in confidence may be due to the continuity of intermediate metabolic states, as the variables (fasting glucose, HbA1c, BMI, triglyceride level, waist circumference level) undergo gradual transitions between healthy and diabetic subjects, and thus the classification is more difficult. Also, the decision curve analysis shows that the CatBoost model offers a clear clinical benefit under a broad spectrum of threshold probabilities, suggesting that it is a clinically useful tool for diabetes risk stratification.

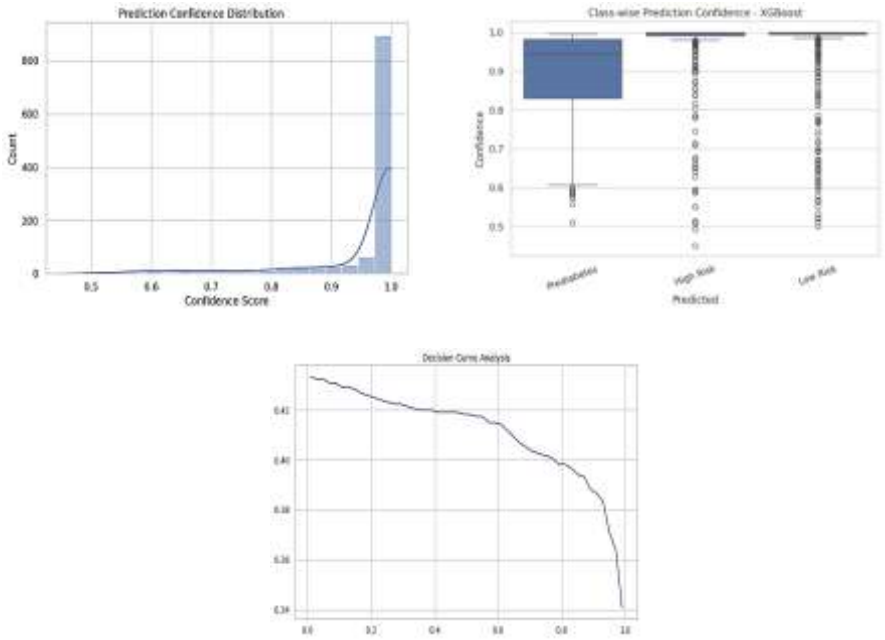


Figure 9. Prediction confidence distribution, class-wise confidence analysis, and decision curve analysis of the CatBoost model for multiclass diabetes risk classification

After the classification performance, confidence analysis and decision curve analysis, the interpretability of the CatBoost model was further investigated by the SHAP (Shapley Additive Explanations) value. The mean absolute SHAP values for most influential features are shown in Figure 10 for the three diabetes risk groups. SHAP analysis gives insights into the contribution of each physiological and lifestyle parameter to the model predictions, as well as identifying which parameters have the greatest influence on the diabetes progression. The body mass index (BMI) significantly contributed to the prediction process in all classes, more than other variables, suggesting that obesity-related factors are the key determinants of diabetes risk. The second most important characteristic was the waist circumference, which was followed by the HbA1c, physical activity, and fasting glucose levels. These results are similar to the feature selection and statistical analyses discussed previously in which these variables had significant correlations with the various diabetes risk groups. BMI and waist circumference are established markers of insulin resistance and metabolic dysfunction, and fasting glucose and HbA1c levels are direct measures of glycemic control and disease progression. The effects of blood pressure, daily calorie intake, triglycerides level, cholesterol level, and sugar intake were also included in the model predictions, but were relatively small.

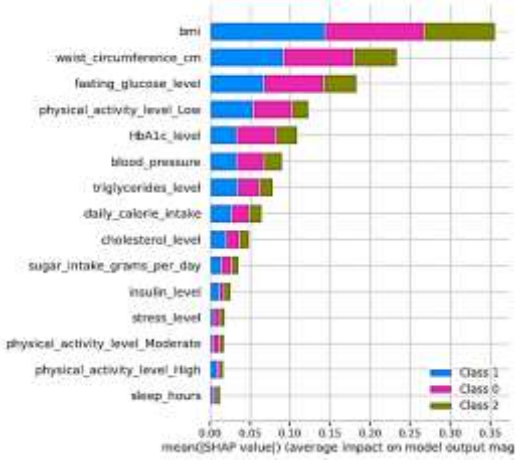


Figure 10. SHAP-based feature importance analysis of the CatBoost model for multiclass diabetes risk classification

Contrary to this, the insulin level, stress level, physical activity at moderate and high intensity and sleep duration showed relatively smaller SHAP values and thus seemed to be less influential on the classification results. However, the contributions show that diabetes risk is dependent on a mixture of metabolic, behavioural, and lifestyle factors, and not just on one. The SHAP analysis also demonstrates that the contribution of each feature varies across the three risk classes, as the underlying physiological features of the Low-Risk, Prediabetes and High-Risk individuals differ.

5. CONCLUSION

The present study proposed a comprehensive machine learning framework for risk of multiclass diabetes estimation based on physiological, metabolic and lifestyle parameters. The proposed framework dimensionality reduction, XAI, and various conventional and ensemble learning algorithms and various conventional and ensemble learning algorithms to classify the clients as Low Risk, Prediabetes, and High Risk. The experimental results showed that the ensemble and boosting methods always outperformed the traditional machine learning models, indicating their capability of modelling the complex nonlinear relationships between the health indicators. The most significant results were found in CatBoost with an accuracy of 94.0% and lowest log loss of 0.1501 and with the highest weighted ROC-AUC of 99.25% among all models. The significance analysis of the features based on Mutual Information and SHAP showed that waist circumference, BMI, fasting glucose, HbA1c level, triglycerides level and physical activity levels are recognized as highly influential features affecting the onset of diabetes. The proposed framework had good classification accuracy across all classes, although the classification accuracy for the Prediabetes class was relatively low due to its overlap with the Low-Risk and High-Risk classes. In addition, the developed models were robust and practically useful as confirmed by the following: ROC analysis, precision-recall analysis, calibration analysis, confidence analysis and decision curve analysis. In general, the findings prove that ensemble learning and XAI can be can be used to develop an effective and interpretable system for assessing the early-stage diabetes risk and providing personalized preventive health care. The proposed framework could be further validated in future work with larger real-world clinical datasets, and by investigating deep learning and hybrid frameworks to further enhance the identification of intermediate-risk individuals.

REFERENCES

[1] Abdalrada, A. S., Neamah, A. F., & Murad, H. (2024). Predicting diabetes disease occurrence using logistic regression: An early detection approach. *Iraqi Journal For Computer Science and Mathematics*, *5*(1), 6.

[2] Abdulhadi, N., & Al-Mousa, A. (2021, July). Diabetes detection using machine learning classification methods. In *2021 international conference on information technology (ICIT)* (pp. 350-354). IEEE.

[3] Afsaneh, E., Sharifdini, A., Ghazzaghi, H., & Ghobadi, M. Z. (2022). Recent applications of machine learning and deep learning models in the prediction, diagnosis, and management of diabetes: a comprehensive review. *Diabetology & Metabolic Syndrome*, *14*(1), 196.

[4] Aich, A., Murshed, M. M., Hewage, S., & Mayeaux, A. (2026). A copula based supervised filter for feature selection in machine learning driven diabetes risk prediction. *Scientific Reports*, *16*(1), 12132.

[5] Akshaya, N., Balaji, Y., Chennarao, J., Sathwik, P., & Dharnasi, P. (2026). Diabetic retinopathy diagnosis with deep learning. *International Journal of Research Publications in Engineering, Technology and Management (IJRPETM)*, *9*(2), 506-512.

[6] Alnowaiser, K. (2024). Improving healthcare prediction of diabetic patients using KNN imputed features and tri-ensemble model. *IEEE Access*, *12*, 16783-16793.

[7] Arya, M., Sastry G, H., Motwani, A., Kumar, S., & Zaguia, A. (2022). A novel extra tree ensemble optimized DL framework (ETEODL) for early detection of diabetes. *Frontiers in Public Health*, *9*, 797877.

[8] Azad, C., Bhushan, B., Sharma, R., Shankar, A., Singh, K. K., & Khamparia, A. (2022). Prediction model using SMOTE, genetic algorithm and decision tree (PMSGD) for classification of diabetes mellitus. *Multimedia Systems*, *28*(4), 1289-1307.

[9] Banday, M. Z., Sameer, A. S., & Nissar, S. (2020). Pathophysiology of diabetes: An overview. *Avicenna journal of medicine*, *10*(04), 174-188.

[10] Bilal, A., Imran, A., Baig, T. I., Liu, X., Long, H., Alzahrani, A., & Shafiq, M. (2024). Improved support vector machine based on CNN-SVD for vision-threatening diabetic retinopathy detection and classification. *Plos one*, *19*(1), e0295951.

[11] García-Ordás, M. T., Benavides, C., Benítez-Andrades, J. A., Alaiz-Moretón, H., & García-Rodríguez, I. (2021). Diabetes detection using deep learning techniques with oversampling and feature augmentation. *Computer Methods and Programs in Biomedicine*, *202*, 105968.

[12] Gourlay, A., Sutherland, C., & Radley, A. (2024). Point-of-care testing of HbA1c levels in community settings for people with established diabetes or people at risk of developing diabetes: a systematic review and meta-analysis. *Primary care diabetes*, *18*(1), 7-16.

[13] Gündoğdu, S. (2023). Efficient prediction of early-stage diabetes using XGBoost classifier with random forest feature selection technique. *Multimedia Tools and Applications*, *82*(22), 34163-34181.

[14] Hossain, M. J., Al-Mamun, M., & Islam, M. R. (2024). Diabetes mellitus, the fastest growing global public health concern: Early detection should be focused. *Health science reports*, *7*(3), e2004.

[15] Jaiswal, S., & Gupta, P. (2022, September). Ensemble approach: XGBoost, CATBoost, and LightGBM for diabetes mellitus risk prediction. In *2022 Second International Conference on Computer Science, Engineering and Applications (ICCSEA)* (pp. 1-6). IEEE.

[16] Khaleel, F. A., & Al-Bakry, A. M. (2023). Diagnosis of diabetes using machine learning algorithms. *Materials Today: Proceedings*, *80*, 3200-3203.

[17] Khan, A. A., Chaudhari, O., & Chandra, R. (2024). A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. *Expert Systems with Applications*, *244*, 122778.

[18] Khosravi, M., Zare, Z., Mojtabaeian, S. M., & Izadi, R. (2024). Artificial intelligence and decision-making in healthcare: a thematic analysis of a systematic review of reviews. *Health services research and managerial epidemiology*, *11*, 23333928241234863.

[19] Kudin, M., Kurniawan, A., & Insani, C. N. (2026). Analysis of Diabetes Classification Performance Improvement Using Ensemble Bagging and K-Fold. *Jurnal Sistem Cerdas*, *9*(1), 98-113.

[20] Kumari, S., Kumar, D., & Mittal, M. (2021). An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. *International Journal of Cognitive Computing in Engineering*, *2*, 40-46.

[21] Liu, R., Li, L., Shao, C., Cai, H., & Wang, Z. (2022). The impact of diabetes on vascular disease: progress from the perspective of epidemics and treatments. *Journal of diabetes research*, *2022*(1), 1531289.

[22] Mohapatra, S. K., Swain, J. K., & Mohanty, M. N. (2018, September). Detection of diabetes using multilayer perceptron. In *International Conference on Intelligent Computing and Applications: Proceedings of ICICA 2018* (pp. 109-116). Singapore: Springer Singapore.

[23] Noviyanti, C. N., & Alamsyah, A. (2024). Early detection of diabetes using random forest algorithm. *Journal of Information System Exploration and Research*, *2*(1).

[24] Sagenela, V. K., Gidijala, K., Karanam, S. R., & Sagar, N. S. S. (2026). Artificial Intelligence in Diabetic Retinopathy Detection: A Decade of Progress from Machine Learning to Transformer-Based Frameworks. *Archives of Computational Methods in Engineering*, 1-26.

[25] Sai, M. J., Chettri, P., Panigrahi, R., Garg, A., Bhoi, A. K., & Barsocchi, P. (2023). An ensemble of light gradient boosting machine and adaptive boosting for prediction of type-2 diabetes. *International Journal of Computational Intelligence Systems*, *16*(1), 14.

[26] Saxena, S., Mohapatra, D., Padhee, S., & Sahoo, G. K. (2023). Machine learning algorithms for diabetes detection: a comparative evaluation of performance of algorithms. *Evolutionary intelligence*, *16*(2), 587-603.

[27] Sharma, T., & Shah, M. (2021). A comprehensive review of machine learning techniques on diabetes detection. *Visual Computing for Industry, Biomedicine, and Art*, *4*(1), 30.

[28] Singh, D. P., Banerjee, T., Mahajan, S., Ramesh Chandra, K., Kumar, R., Phani, S., ... & Ashok, M. (2026). A comprehensive study on deep learning models for the detection of diabetic retinopathy using pathological images. *Archives of Computational Methods in Engineering*, *33*(1), 503-532.

[29] Singh, S., Wani, N. A., Kumar, R., & Bedi, J. (2025). DiaXplain: A transparent and interpretable artificial intelligence approach for Type-2 diabetes diagnosis through deep learning. *Computers and Electrical Engineering*, *126*, 110470.

[30] Vishard. (2025). *Diabetes risk prediction dataset* [Data set]. Kaggle. https://www.kaggle.com/datasets/vishardmehta/diabetes-risk-prediction-dataset

[31] 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.

[32] Zaki, M. G., & Imaduddin, H. (2026). Early Classification of Diabetes Risk in Productive Age Groups Using Machine Learning. *Jurnal Penelitian Sains Teknologi*, 75-90.