World Journal of Advanced Research
worldadvancedjournal.com
Online ISSN: 3107-8966
Received: 09-09-2026, Accepted: 09-10-2026, Published: 24-10-2026
Volume 2, Issue 4, 2026, Page No. 25-28

Predictive modeling of diabetes using ensemble machine learning and patient health indicators

Marek Pawlicki
Bydgoszcz University of Science and Technology, Bydgoszcz, Poland

Abstract
Appropriate management to reduce the risk of long-term complications. The increasing availability of patient health data has created opportunities for machine learning techniques to support diabetes risk prediction. This article presents a predictive modeling framework based on ensemble machine learning and routinely collected patient health indicators. The proposed framework focuses on combining complementary machine learning algorithms, particularly Random Forest, XGBoost, and Gradient Boosting, to identify patterns associated with diabetes. Relevant patient indicators may include age, body mass index (BMI), glucose level, blood pressure, insulin, pregnancy history, and other routinely collected clinical characteristics. Previous research has demonstrated that machine learning can identify useful patterns in diabetes-related health data, while ensemble methods can provide robust predictive performance across different datasets (Chang *et al*., 2022; Mohsen *et al*., 2023) [2, 6]. The proposed framework incorporates data preprocessing, feature selection, class-imbalance management, model training, ensemble construction, and performance evaluation using accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC). Explainable artificial intelligence techniques such as SHAP can additionally be used to identify the health indicators that contribute most strongly to model predictions. The framework can provide a foundation for developing data-driven decision-support systems for early diabetes risk assessment, although external validation and clinical evaluation are required before practical deployment.

Keywords: Diabetes mellitus, machine learning, ensemble learning, random forest, xgboost, gradient boosting, patient health indicators, predictive modeling, diabetes risk prediction

Introduction

Diabetes mellitus is a chronic metabolic disorder characterized by abnormal regulation of blood glucose. Early identification of individuals at elevated risk is important because timely intervention can support disease management and potentially reduce the burden of diabetes-related complications. Conventional approaches to diabetes assessment generally depend on clinical measurements, laboratory investigations, medical history, and professional interpretation. The growing availability of electronic health records and routinely collected patient information has created opportunities for computational approaches that can complement these traditional methods. Machine learning (ML) has increasingly been investigated for diabetes prediction because it can identify nonlinear relationships and interactions among multiple patient characteristics.

A scoping review of artificial intelligence approaches for diabetes risk prediction found that classical machine learning methods remain widely used, with Random Forest, support vector machines, and XGBoost among the commonly investigated approaches (Mohsen *et al*., 2023) [6]. Patient health indicators provide an important foundation for predictive modeling. Variables such as glucose, BMI, age, blood pressure, and insulin can contain information relevant to diabetes risk. Chang *et al*. (2022) [2], for example, evaluated several machine learning algorithms using health indicators for early diabetes prediction and reported that Random Forest achieved an accuracy of 82.26% in their particular dataset. These findings demonstrate the potential value of health-indicator-based machine learning while also emphasizing that model performance depends on the population and dataset used. Ensemble learning provides another approach to improving predictive modeling. Instead of relying on one classifier, ensemble methods combine multiple learners to obtain a final prediction. Random Forest uses multiple decision trees, while boosting methods such as Gradient Boosting and XGBoost sequentially develop models that focus on correcting previous prediction errors.

Recent research has continued to investigate these methods for diabetes prediction and related clinical outcomes (Bhuiyan *et al*., 2026; El Sherbiny *et al*., 2026) [1].

Therefore, this article proposes a predictive framework that combines ensemble machine learning with patient health indicators for diabetes risk assessment.

Problem Statement

Traditional diabetes assessment can require multiple clinical measurements and professional interpretation. Although these approaches remain essential, the increasing volume of patient health information creates an opportunity to develop computational systems capable of identifying patterns associated with diabetes. A further challenge is that the performance of individual machine learning algorithms can vary according to the characteristics of the dataset, population, feature distribution, and evaluation methodology. A model that performs well in one dataset may not necessarily achieve the same performance in another population. A scoping review of diabetes risk prediction studies highlighted substantial variation in algorithms, predictors, datasets, and evaluation approaches, making direct comparison between studies difficult (Mohsen *et al*., 2023) [6]. Consequently, there is a need for a systematic predictive framework that evaluates multiple ensemble algorithms using consistent preprocessing and evaluation procedures. Such a framework can help determine how effectively patient health indicators can be used to distinguish individuals with and without diabetes.

Literature Review

Machine Learning for Diabetes Prediction

Machine learning has become an important research area in healthcare because predictive algorithms can learn patterns from historical patient information. In diabetes research, ML models have been applied to early detection, diagnosis, risk prediction, and complication prediction. Kopitar *et al*. (2020) [4] demonstrated the application of machine learning-based prediction models for the early detection of type 2 diabetes. Similarly, Chang *et al*. (2022) [2] compared several algorithms using health indicators and demonstrated that machine learning can provide useful predictive information for diabetes classification. More recent studies have continued to evaluate tree-based ensemble models. A 2026 study using patient health records compared several classifiers, including Random Forest, XGBoost, and Gradient Boosting, and investigated explainability using SHAP. The study reported that glucose, insulin, BMI, and age were among the influential variables in its specific dataset (Bhuiyan *et al*., 2026) [1].

Random Forest

Random Forest is an ensemble algorithm that constructs multiple decision trees and combines their predictions. Its ability to model nonlinear relationships and interactions makes it suitable for structured healthcare datasets.

Random Forest has been frequently investigated in diabetes prediction. Mohsen *et al*. (2023) [6] reported that Random Forest was used in nine studies included in their review of AI-based diabetes risk prediction. Chang *et al*. (2022) [2] also found strong predictive performance for Random Forest in their health-indicator dataset.

XGBoost

Extreme Gradient Boosting (XGBoost) is a gradient-boosting algorithm designed to construct an ensemble of decision trees sequentially. Each new tree attempts to improve the errors produced by previous trees.

XGBoost has attracted substantial attention in medical prediction because of its ability to model complex relationships in structured data. Mohsen *et al*. (2023) [6] identified XGBoost among the commonly used machine learning algorithms in diabetes risk prediction research.

Recent research has also applied XGBoost to diabetes classification and patient health records. El Sherbiny *et al*. (2026), for example, developed an XGBoost-based framework for diabetes mellitus multiclassification and compared it with other machine learning algorithms. Such research supports continued investigation of boosting approaches for diabetes prediction.

Gradient Boosting

Gradient Boosting is another ensemble technique that sequentially combines weak learners to construct a stronger predictive model. It can capture nonlinear relationships between patient characteristics and diabetes outcomes. Gradient Boosting has been included in comparative diabetes prediction studies alongside Random Forest, support vector machines, and other algorithms. Its inclusion in the proposed framework allows researchers to examine whether different boosting strategies capture complementary information from patient health indicators.

Ensemble Learning

The central principle of ensemble learning is that multiple predictive models can be combined to produce a final prediction. The rationale is that different algorithms may capture different patterns in the same data. A recent systematic review of ensemble learning in predictive healthcare reported widespread use of hybrid and stacking approaches across several disease areas, including diabetes (Saha *et al*., 2026) [7]. However, reported performance should not automatically be generalized across datasets because differences in patient populations, variables, sample sizes, preprocessing, and validation methods can substantially influence results.

Proposed Methodology

Research Design

The proposed study adopts a quantitative predictive modeling design. Patient health records containing diabetes-related indicators would be processed and used to train supervised machine learning models. The dependent variable would represent diabetes status, while independent variables would consist of patient health indicators.

Patient Health Indicators

Potential predictors include:

The final variables should depend on the characteristics and quality of the selected dataset.

Data Preprocessing

Data preprocessing is an essential stage because healthcare datasets may contain missing values, inconsistent measurements, outliers, and imbalanced classes. The proposed preprocessing procedure includes:

1. Removing duplicate records.
2. Identifying missing or invalid values.
3. Applying appropriate imputation methods.
4. Detecting extreme or implausible observations.
5. Encoding categorical variables where necessary.
6. Scaling variables when required by particular algorithms.
7. Separating predictors from the target variable
8. Dividing the dataset into training and testing subsets.

Importantly, preprocessing operations should be fitted using training data only when appropriate to prevent information leakage between training and testing data.

Handling Class Imbalance

Diabetes datasets may contain unequal numbers of diabetic and non-diabetic observations. If substantial imbalance exists, techniques such as class weighting or Synthetic Minority Over-sampling Technique (SMOTE) can be considered. Recent diabetes prediction research has used SMOTE together with careful cross-validation to address class imbalance while avoiding leakage between training and test data (Bhuiyan *et al*., 2026) [1].

Feature Selection

Feature selection can be used to determine which patient indicators provide the greatest predictive contribution. Tree-based feature importance, permutation importance, or SHAP values can be used for this purpose. Feature selection may reduce unnecessary variables while improving interpretability. However, feature selection must also be

performed carefully within the training process to prevent information from the test set influencing model development.

ROC-AUC

The area under the receiver operating characteristic curve (ROC-AUC) measures the model's ability to distinguish between the target classes across different classification thresholds. Using several metrics provides a more comprehensive assessment than relying on accuracy alone. Diabetes prediction studies have similarly used combinations of accuracy, precision, recall, F1-score, and AUC to evaluate predictive models (Chang *et al*., 2022) [2].

Explainable Artificial Intelligence

Although predictive accuracy is important, understanding why a model generates a particular prediction is also valuable. Explainable artificial intelligence (XAI) techniques can therefore be incorporated into the proposed framework. SHAP (SHapley Additive exPlanations) can be used to estimate the contribution of individual features to model predictions. For example, a SHAP analysis could determine whether glucose, BMI, age, insulin, or blood pressure has a stronger contribution to a particular prediction. Recent patient-record-based research has applied SHAP analysis to machine learning diabetes prediction and identified variables such as insulin, glucose, BMI, and age as important predictors in the studied dataset (Bhuiyan *et al*., 2026) [1]. Explainability can improve transparency and help researchers investigate whether the model is relying on clinically meaningful information rather than irrelevant patterns.

Expected Results

Because this article does not use a supplied dataset or report an original experiment, numerical results should not be invented. The proposed study is expected to produce: The actual performance of each model should be reported only after experimentation on the selected dataset. Existing research demonstrates that the relative performance of Random Forest, XGBoost, and other algorithms can vary across datasets and populations. For example, a 2026 comparative study using the Pima Indians Diabetes dataset reported similar AUC values for Random Forest and XGBoost despite differences in their accuracy values (Emre, 2026) [3]. This illustrates why a specific study should report its own experimentally obtained results rather than assuming that one algorithm will always outperform another.

Discussion

The proposed framework integrates patient health indicators with ensemble machine learning to address diabetes prediction. The use of several ensemble algorithms provides an opportunity to compare different approaches to learning nonlinear relationships within clinical data. Random Forest can provide robustness through aggregation across multiple decision trees, while XGBoost and Gradient Boosting use sequential learning to improve predictive performance. Combining their predictions may potentially provide a more stable prediction than relying on a single algorithm, although this must be empirically tested. The use of health indicators is particularly relevant because many diabetes prediction datasets contain routinely available measurements. Research using health indicators has shown that machine learning models can identify patterns useful for diabetes classification (Chang *et al*., 2022) [2]. At the same time, model performance can be strongly influenced by dataset characteristics. Therefore, high performance reported in one study should not automatically be interpreted as evidence that the same model will perform similarly in another population. Another important consideration is model interpretability. Healthcare professionals may require information about which variables contribute to predictions rather than receiving only a binary classification. SHAP and related methods can help provide this information and make model behavior easier to investigate.

Limitations

Several limitations should be considered.

First, machine learning performance depends heavily on the quality and representativeness of the dataset. A model trained on a relatively small or geographically limited population may not generalize to other populations.

Second, missing values and measurement errors can influence model performance. Appropriate preprocessing and validation procedures are therefore essential.

Third, class imbalance may lead to misleading accuracy measurements. Multiple evaluation metrics should consequently be used.

Fourth, ensemble models can be more difficult to interpret than simple statistical models. Explainability techniques can partially address this issue but do not eliminate the need for careful validation.

Finally, predictive performance alone does not establish clinical usefulness. Before implementation in healthcare settings, models require appropriate external validation, calibration assessment, prospective evaluation, and consideration of clinical workflow.

Conclusion

This article proposes a predictive framework for diabetes assessment using ensemble machine learning and patient health indicators. The framework combines Random Forest, XGBoost, and Gradient Boosting with systematic data preprocessing, feature selection, class-imbalance management, and comprehensive model evaluation.

Existing research indicates that machine learning can effectively use patient health information for diabetes prediction, while ensemble approaches are frequently investigated because of their ability to model complex relationships (Chang *et al*., 2022; Mohsen *et al*., 2023) [2, 6]. The proposed approach further incorporates explainability through techniques such as SHAP to identify important health indicators and improve transparency.

However, no single algorithm should be assumed to be universally superior. The appropriate model should be determined through rigorous experimentation using a representative dataset and appropriate validation procedures. Ultimately, an adequately validated ensemble model could serve as a research foundation for data-driven diabetes risk assessment and clinical decision-support applications, rather than replacing professional diagnosis.

References

1. Bhuiyan T, Tasnim M, Khan M A N, Onik T A, Nabil M A, Alam, A, *et al.* Machine learning-based prediction

of diabetes using patient health records. Vascular and Endovascular Review,2026:9(1):446–454.
2. Chang V, Ganatra M A, Hall K, Golightly L, Xu Q A. An assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators. Healthcare Analytics,2022:2:100118.
3. Emre İ E. Predicting Type 2 diabetes using Random Forest and XGBoost algorithms: A comparative machine learning approach. Acıbadem Üniversitesi Sağlık Bilimleri Dergisi, 2026, 17.
4. Kopitar L, Kocbek P, Cilar L, Sheikh A, Stiglic G. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. Scientific Reports,2020:10:11981.
5. Bhatta R P. Ensemble based machine learning model for prediction of diabetes. NPRC Journal of Multidisciplinary Research,2026:3(2):116-129
6. Mohsen F, Al-Absi H R H, *et al*. A scoping review of artificial intelligence-based methods for diabetes risk prediction. npj Digital Medicine,2023:6:197.
7. Saha R, Saif S, Khan T, *et al*. A critical systematic review of ensemble learning methods for predictive healthcare. Discover Computing,2026:29:117.
8. Alsadi B, Musleh S, Al-Absi H R, Refaee M, Qureshi R, El Hajj, *et al*. An ensemble-based machine learning model for predicting type 2 diabetes and its effect on bone health. BMC Medical Informatics and Decision Making,2024:24(1):144.
9. Acheampong E, Adua E, Obirikorang C, Anto E O, Peprah-Yamoah E, Obirikorang Y, *et al*. Predictive modelling of metabolic syndrome in Ghanaian diabetic patients: an ensemble machine learning approach. Journal of Diabetes & Metabolic Disorders,2024:23(2): 2233-2249.
10. Ganie S M, Pramanik P K D, Bashir Malik M, Mallik S, Qin H. An ensemble learning approach for diabetes prediction using boosting techniques. Frontiers in Genetics,2023:14:1252159.
11. Ganie S M, Pramanik P K D, Bashir Malik M, Mallik S, Qin H. An ensemble learning approach for diabetes prediction using boosting techniques. Frontiers in Genetics,2023:14:1252159.
12. Fitriyani N L, Syafrudin M, Alfian G, Rhee J. Development of disease prediction model based on ensemble learning approach for diabetes and hypertension. Ieee Access,2019:7:144777-144789.
13. Abnoosian K, Farnoosh R, Behzadi M H. Prediction of diabetes disease using an ensemble of machine learning multi-classifier models. BMC bioinformatics,2023:24(1):337.
14. Datta D, Bhattacharya M, Rajest S S, Shynu T, Regin R, Priscila S S. Development of predictive model of diabetic using supervised machine learning classification algorithm of ensemble voting. International Journal of Bioinformatics Research and Applications,2023:19(3):151-169.
15. Liu J, Wang L, Zhang L, Zhang Z, Zhang S. Predictive analytics for blood glucose concentration: an empirical study using the tree-based ensemble approach. Library Hi Tech,2020:38(4):835-858.