World Journal of Current Research
worldcurrentjournal.com
Online ISSN: 3107-9636
Received: 07-09-2026, Accepted: 07-10-2026, Published: 22-10-2026
Volume 2, Issue 4, 2026, Page No. 4-8

Explainable ensemble machine learning for diabetes prediction and risk assessment

Gaith Rjoub
King Hussein School of Computing Sciences, Princess Sumaya University for Technology (PSUT), Amman, Jordan

Abstract
Diabetes mellitus is a major chronic metabolic disorder that requires early identification and appropriate risk assessment to reduce the likelihood of severe complications. The increasing availability of clinical, demographic, behavioral, and physiological data has created opportunities for machine learning-based predictive systems capable of identifying complex patterns associated with diabetes. However, the use of highly complex machine learning models in healthcare is often challenged by limited interpretability, making it difficult for healthcare professionals and other stakeholders to understand why a particular prediction has been generated. This study presents an explainable ensemble machine learning framework for diabetes prediction and risk assessment that combines multiple predictive algorithms with explainable artificial intelligence techniques. The proposed framework integrates ensemble classifiers such as Random Forest, Gradient Boosting, and XGBoost with feature selection and SHAP-based model interpretation. Patient characteristics, including demographic information, anthropometric measurements, clinical indicators, and relevant behavioral factors, can be used as predictive variables. The methodology involves data preprocessing, missing-value treatment, feature selection, model development, ensemble learning, performance evaluation, and explainability analysis. Model performance can be assessed using accuracy, precision, recall, F1-score, receiver operating characteristic area under the curve (ROC-AUC), and calibration measures. SHAP analysis can subsequently be used to determine the importance and direction of individual predictors. Recent research demonstrates that stacking ensemble models can achieve strong predictive performance while SHAP can provide interpretable information about important diabetes risk factors (Resma *et al*., 2026) [3]. Earlier research has similarly demonstrated the feasibility of combining ensemble classification with SHAP explanations for diabetes prediction (Khan *et al*., 2022) [1]. The proposed framework therefore emphasizes both predictive performance and transparency, providing a foundation for developing trustworthy machine learning systems for diabetes risk assessment. Nevertheless, external validation, calibration, and clinical assessment remain necessary before such models can be considered for real-world healthcare applications.

Keywords: Diabetes mellitus, ensemble machine learning, explainable artificial intelligence, SHAP, Random Forest, XGBoost, diabetes prediction, risk assessment, machine learning

Introduction

Diabetes mellitus is a chronic metabolic disorder associated with abnormal blood glucose regulation and a wide range of long-term health consequences. Early identification of individuals at elevated risk is important because timely assessment and intervention can support appropriate disease management and prevention strategies. The increasing availability of electronic health records, population surveys, and routinely collected patient health information has created new opportunities for computational methods to assist with diabetes prediction and risk assessment.

Machine learning has attracted considerable attention in healthcare because it can identify relationships among multiple variables without requiring all relationships to be explicitly specified in advance. Diabetes prediction models can use demographic, clinical, anthropometric, and behavioral information to distinguish individuals who are at higher or lower risk. Ensemble learning is particularly attractive because it combines multiple predictive models and can potentially reduce the weaknesses associated with relying on a single classifier.

Recent evidence demonstrates the growing role of ensemble learning in diabetes prediction. Resma *et al*. (2026) [3], in a population-based study of adults in northern Bangladesh, developed multiple machine learning models followed by a stacking ensemble. The stacking model achieved an AUC of 0.967 in that study, while SHAP analysis was used to identify important predictors, including age, family history of diabetes, BMI, weight, and selected dietary factors. These findings illustrate how predictive modeling and explainability can be incorporated within the same analytical framework.

Despite these advances, predictive accuracy alone is insufficient for many healthcare applications. A model may produce an accurate prediction without providing an understandable explanation for that prediction. This creates a potential barrier to trust and adoption. Explainable artificial intelligence (XAI) addresses this issue by providing methods for investigating how individual features contribute to model outputs. SHAP, based on Shapley values from cooperative game theory, is one such approach and can provide both global and individual-level explanations.

Research has demonstrated the application of SHAP and other explainability techniques to diabetes prediction. Khan *et al*. (2022) [1] developed an ensemble approach involving several machine learning algorithms and used SHAP to generate global and local explanations of diabetes predictions. Their work demonstrated how explainability can complement predictive modeling by providing information about the factors influencing individual predictions.

The combination of ensemble learning and explainable artificial intelligence therefore represents an important direction for diabetes prediction research. This article proposes an explainable ensemble machine learning

framework designed to predict diabetes and assess the factors contributing to individual risk predictions.

Problem Statement

The early identification of diabetes can be challenging because the disease is influenced by numerous interacting factors. Patient characteristics such as age, body mass index, glucose level, blood pressure, family history, and lifestyle-related variables may contribute to diabetes risk in different ways. Traditional statistical models can provide valuable information, but they may not fully capture complex nonlinear interactions among numerous predictors.

Machine learning can address some of these challenges by learning predictive relationships directly from data. However, complex machine learning algorithms can behave as black-box systems in which the reasoning behind an individual prediction is difficult to understand. This limitation is particularly important in healthcare, where predictions may influence subsequent clinical assessment or resource allocation.

Recent research has attempted to address this issue by integrating explainability into diabetes prediction. For example, a 2026 population-based study combined stacking ensemble learning with SHAP analysis to both predict diabetes and identify important risk factors. The researchers subsequently compared their findings with multivariable logistic regression, providing an additional validation perspective.

The central problem addressed by this study is therefore not simply how to obtain accurate diabetes predictions, but how to develop a predictive framework that combines strong ensemble learning with understandable explanations of model behavior.

Literature Review

Machine Learning for Diabetes Prediction

Machine learning has increasingly been applied to diabetes prediction because of its ability to analyze complex relationships among patient characteristics. Algorithms such as Random Forest, support vector machines, Gradient Boosting, XGBoost, and neural networks have been investigated for diabetes classification and risk prediction.

Khan *et al*. (2022) [1] evaluated multiple machine learning algorithms and an ensemble classifier for diabetes diagnosis. Their study incorporated missing-value imputation and class-balancing procedures before developing the predictive models. The researchers subsequently applied SHAP to provide global and local explanations of model predictions. Their weighted ensemble achieved a balanced accuracy of 90% and an F1-score of 89% within their experimental setting.

These results demonstrate the potential of combining several predictive models, while also illustrating the importance of model interpretation. Nevertheless, performance values obtained from one dataset should not be generalized automatically to other populations because differences in data quality, prevalence, feature distributions, and study design can substantially influence model performance.

Ensemble Machine Learning

Ensemble learning combines multiple predictive models to produce a final prediction. Common approaches include voting, bagging, boosting, blending, and stacking. The underlying principle is that different models may capture different patterns within the same dataset, allowing their combined predictions to potentially provide more robust results.

Random Forest is an example of bagging-based ensemble learning in which multiple decision trees are constructed and their predictions are aggregated. Gradient Boosting and XGBoost represent boosting-based approaches in which models are developed sequentially to reduce previous prediction errors.

Stacking extends the ensemble concept by using predictions from several base models as inputs to a meta-model. Recent diabetes research has demonstrated the application of this approach. Resma *et al*. (2026) [3] developed a stacking ensemble from multiple machine learning algorithms and reported stronger predictive performance for the ensemble than for the individual models in their northern Bangladesh population.

Explainable Artificial Intelligence

Explainable artificial intelligence refers to methods designed to make machine learning predictions more understandable. This is particularly important in healthcare because users may need to understand which patient characteristics contributed to a prediction.

SHAP is one of the most widely used model-explanation approaches. It assigns contribution values to features based on Shapley values, allowing researchers to examine both the overall importance of predictors and the contribution of individual variables to particular predictions.

Prendin *et al*. (2023) [2] demonstrated the value of SHAP for interpreting machine learning predictions in diabetes-related modeling and emphasized that explainability can help identify potential model biases and improve understanding of model behavior.

In diabetes prediction, SHAP can be used to determine whether variables such as glucose, BMI, age, blood pressure, insulin, family history, or behavioral characteristics are contributing positively or negatively to a model's predicted risk.

Explainable Ensemble Models

The combination of ensemble learning and explainability provides an opportunity to address two important requirements of healthcare prediction: predictive performance and transparency.

Khan *et al*. (2022) [1] combined several machine learning algorithms with a weighted ensemble classifier and applied SHAP to generate both global and local explanations. Their approach demonstrated that an ensemble model can be accompanied by interpretable information rather than functioning exclusively as a black-box predictor.

More recent evidence further supports this direction. Resma *et al*. (2026) [3] used SHAP to explain a stacking ensemble and identified demographic, anthropometric, family-history, and behavioral characteristics associated with diabetes predictions in their study population. The researchers also compared machine learning findings with multivariable logistic regression, strengthening the interpretive analysis within that specific study.

Proposed Methodology

The proposed research adopts a quantitative predictive modeling approach. The framework is designed to use patient health data to develop an ensemble classifier capable

of estimating diabetes risk while simultaneously providing explanations for its predictions.

Data Collection

The framework can be implemented using an appropriate publicly available diabetes dataset or an anonymized clinical dataset. The selected dataset should contain sufficient demographic, clinical, anthropometric, and behavioral information to support diabetes risk assessment.

Potential variables may include age, gender, glucose level, body mass index, blood pressure, insulin, family history, physical activity, dietary characteristics, and other relevant health indicators. The precise variables should be determined by the selected dataset and research objectives.

Data Preprocessing

Data preprocessing is an essential stage because healthcare datasets may contain missing observations, inconsistent values, outliers, duplicate records, and class imbalance.

Missing values should be examined and treated using an appropriate imputation strategy. Categorical variables may require encoding, while continuous variables may be standardized where appropriate. Duplicate observations should be identified, and implausible measurements should be investigated before model development.

Class imbalance should also be assessed. If one outcome class substantially outnumbers the other, techniques such as class weighting or oversampling may be considered. Khan *et al*. (2022) [1], for example, used SMOTETomek to address class imbalance in their diabetes ensemble study.

Importantly, preprocessing procedures must be performed in a manner that prevents information from the test dataset from influencing model training.

Feature Selection

Feature selection can improve model efficiency and interpretability by identifying variables that provide meaningful predictive information.

The proposed framework can employ methods such as recursive feature elimination, Random Forest feature importance, mutual information, correlation analysis, or statistical selection techniques. The selected features can then be used as inputs to the ensemble models.

Recent research has demonstrated the use of multiple feature-selection strategies in explainable diabetes prediction. Resma *et al*. (2026) [3], for example, combined recursive feature elimination, Random Forest importance, and Best First Search before constructing their stacking ensemble.

Risk Assessment

The proposed system should provide more than a simple diabetes or non-diabetes classification. Where appropriate, the ensemble can generate a probability representing the estimated risk associated with the model's prediction.

The resulting probabilities can be categorized into meaningful risk groups only after appropriate calibration and validation. For example, the research could investigate low-, moderate-, and high-risk categories based on validated probability thresholds. These categories should not be arbitrarily defined because poorly calibrated thresholds could produce misleading risk estimates.

Calibration assessment is therefore an important component of the proposed framework. Measures such as calibration curves, Brier score, and expected calibration error can complement discrimination measures such as ROC-AUC.

Model Explainability Using SHAP

SHAP analysis is a central component of the proposed framework. It can explain both the general behavior of the ensemble model and individual patient predictions.

Global SHAP analysis can rank variables according to their average contribution to model predictions. This may reveal whether glucose, age, BMI, blood pressure, family history, or other characteristics are particularly influential in the selected population.

Local SHAP explanations can provide a patient-specific interpretation. For example, a prediction may be influenced by several features simultaneously, with some increasing the predicted risk and others decreasing it. This information can provide a more detailed interpretation than simply reporting that an individual has been classified as high or low risk.

The interpretation should remain descriptive rather than causal. A feature having a large SHAP contribution does not establish that the feature causes diabetes. Instead, it indicates that the feature was influential in the model's prediction.

Recent population-based research illustrates this distinction. Resma *et al*. (2026) [3] used SHAP to identify important predictors of diabetes within their study population and then used logistic regression to examine associations independently.

Model Evaluation

The predictive performance of the individual models and the final ensemble should be evaluated using multiple measures. Accuracy provides an overall measure of correct classification, while precision indicates the proportion of positive predictions that are correct. Recall measures the ability to identify actual diabetes cases, and the F1-score combines precision and recall.

ROC-AUC can be used to assess the ability of the model to discriminate between diabetes and non-diabetes across different thresholds. When class imbalance is present, precision-recall analysis can provide additional information.

Calibration should also be evaluated because risk assessment requires reliable probability estimates. A model may demonstrate strong discrimination while producing poorly calibrated probabilities. Therefore, a comprehensive evaluation should consider discrimination, calibration, and clinical usefulness rather than relying exclusively on accuracy.

Resma *et al*. (2026) [3] evaluated their diabetes prediction models using accuracy, precision, recall, F1-score, and AUC and reported that their stacking ensemble achieved an AUC of 0.967 in their specific population. These results demonstrate the potential of ensemble learning but should not be interpreted as a universal benchmark because performance depends on the dataset and population.

Expected Results and Discussion

Because the present article proposes a research framework rather than reports an experiment using a supplied dataset, specific numerical results should not be invented. The proposed analysis would compare Random Forest, Gradient Boosting, XGBoost, and the final ensemble under identical validation conditions.

The principal expectation is that the ensemble may provide competitive or improved predictive performance by combining the complementary strengths of its constituent models. However, the ensemble should only be considered beneficial if improvements are demonstrated on independent test data and remain stable across cross-validation procedures.

The explainability component is expected to provide additional information regarding the factors influencing model predictions. SHAP analysis may identify variables that are consistently influential across the study population, while local explanations may demonstrate why specific observations receive particular risk predictions.

Recent studies support the feasibility of this combined approach. Khan *et al*. (2022) [1] demonstrated that an ensemble model could be paired with SHAP-based global and local explanations. Similarly, Resma *et al*. (2026) [3] integrated stacking ensemble learning and SHAP analysis to predict diabetes and identify important risk factors in a population-based study.

Nevertheless, high predictive accuracy should not automatically be interpreted as evidence of clinical effectiveness. A model trained on a particular population may perform differently when applied to another population because of differences in demographics, disease prevalence, healthcare systems, measurement procedures, and data distributions.

The interpretability of the model also requires careful consideration. SHAP values explain the contribution of variables to a model's output, but they do not demonstrate causation. Consequently, an influential variable identified by SHAP should not automatically be interpreted as a causal risk factor.

Clinical and Research Significance

An explainable ensemble model could provide value as a decision-support research tool by combining predictive classification with information about the factors underlying predictions. Instead of producing only a binary output, the system could provide an estimated risk probability and an explanation of the major features contributing to that prediction.

Such a framework could be particularly relevant in settings where large amounts of routinely collected patient data are available but resources for comprehensive screening are limited. However, its practical value would depend on external validation, calibration, usability, fairness, and integration with established clinical workflows.

Recent research in northern Bangladesh demonstrates the relevance of this direction in a low- and middle-income setting. Resma *et al*. (2026) [3] used population-level data and explainable ensemble learning to identify diabetes risk factors and develop a predictive model, demonstrating how machine learning could be applied to public-health-oriented risk assessment.

Limitations

The proposed framework has several limitations. Its performance will depend on the quality and representativeness of the selected dataset. Missing information, measurement errors, sampling bias, and class imbalance can affect model predictions.

Another limitation concerns generalizability. A model developed using one population may not perform equally well in another population. External validation is therefore essential before interpreting the model as broadly applicable.

Ensemble models may also be more complex than individual classifiers. Although SHAP can improve interpretability, explanations should not be regarded as complete descriptions of the biological mechanisms underlying diabetes.

Furthermore, predictive associations identified by machine learning do not establish causality. Variables identified as important by SHAP should therefore be interpreted as model predictors rather than automatically being treated as causal determinants.

Finally, a machine learning model should not replace clinical diagnosis. The proposed framework is intended as a predictive and research-support tool, and any future clinical implementation would require appropriate validation and professional oversight.

Conclusion

This article proposes an explainable ensemble machine learning framework for diabetes prediction and risk assessment. The proposed approach integrates Random Forest, Gradient Boosting, and XGBoost with feature selection, ensemble learning, probability-based risk assessment, and SHAP-based explainability.

The combination of ensemble learning and explainable artificial intelligence addresses two important challenges in healthcare prediction. Ensemble learning can provide a flexible mechanism for combining complementary predictive models, while explainability techniques can provide information about the variables contributing to model outputs.

Previous research has demonstrated the feasibility of combining ensemble learning with SHAP for diabetes prediction (Khan *et al*., 2022) [1], while recent population-based research has further demonstrated the use of stacking ensembles and SHAP to predict diabetes and identify important risk factors (Resma *et al*., 2026) [3].

The principal contribution of the proposed framework is its emphasis on both prediction and transparency. Rather than evaluating a model solely according to its accuracy, the framework considers discrimination, calibration, feature contribution, and interpretability. Future research should validate the framework using diverse populations, independent datasets, longitudinal health records, and prospective clinical studies. With appropriate validation, explainable ensemble learning could provide a useful foundation for data-driven diabetes risk assessment and decision-support research.

References

1. Khan A, *et al*. An ensemble approach for the prediction of diabetes mellitus using a soft voting classifier with an explainable AI. Diagnostics,2022:12(12):3094.
2. Prendin F, Pavan J, Cappon G, Del Favero S, Sparacino G, Facchinetti A. The importance of interpreting machine learning models for blood glucose prediction in diabetes, An analysis using SHAP. Scientific Reports,2023:13:16865.
3. Resma MNJ, Kayum MA, Bhowmik P, Khondaker MEA, Kibria MK. Explainable ensemble machine learning for predicting diabetes mellitus and identifying

key risk factors, A population-based study in northern Bangladesh. Scientific Reports,2026:16:24698.
4. Reátegui R, Ramírez-Cerna L, Bautista-Valarezo E. Interpretable diabetes risk prediction, A comparative study of tree-based algorithms using SHAP and LIME. Scientific Reports,2026.
5. Kibria HB, Nahiduzzaman M, Goni MOF, Ahsan M, Haider J. An ensemble approach for the prediction of diabetes mellitus using a soft voting classifier with an explainable AI. Sensors,2022:22(19):7268.
6. Yuan C, Liu Z, Li X, Zhou X, Wang D, Fan Y, *et al*. A dynamic weighted ensemble learning framework for cardiovascular risk prediction in type 2 diabetes, A comparative study with SHAP-based interpretability. Scientific Reports,2025:15(1):45029.
7. Bhatta RP. Ensemble based machine learning model for prediction of diabetes. NPRC Journal of Multidisciplinary Research,2026:3(2):116–129.
8. Shaheen I, Javaid N, Alrajeh N, Asim Y, Aslam S. Hi-le and hitcle, Ensemble learning approaches for early diabetes detection using deep learning and explainable artificial intelligence. IEEE Access,2024:12:66516–66538.
9. Tasin I, Nabil TU, Islam S, Khan R. Diabetes prediction using machine learning and explainable AI techniques. Healthcare Technology Letters,2023:10(1-2):1–10.
10. Elgendy IA, Hosny M, Albashrawi MA, Alsenan S. Dual-stage explainable ensemble learning model for diabetes diagnosis. Expert Systems with Applications,2025:274:126899.
11. Awan SF, Irfan K, Butt UM, Letchmunan S, Hassan FH. Explainable diabetes prediction using a stacked ensemble framework. PLOS ONE,2026:21(9):e0352313.
12. Ghadekar P, Shaikh U, Ner R, Patil S, Qazi R. Predicting heart disease risk in diabetic patients using a pipeline of ensemble learning and XAI-enhanced approaches. In 2024 1st International Conference on Cognitive, Green and Ubiquitous Computing (IC-CGU). IEEE, 2024, March:1–8.
13. Ahmad S, Hussain S, Arif M, Ansari MA. Early-stage diabetes prediction using a stacked ensemble model enhanced with SHAP explainability. Biomedical and Pharmacology Journal,2026:19(1):246–263.
14. Hasnat A, Himel MTM, Hasan MR, Parves KTB, Hasan MS. Diabetes prediction with explainable AI, A stacking ensemble approach for better accuracy and interpretability. In 2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN). IEEE, 2025, July:1–6.
15. Hossain MI, Palas MR, Quader MA, Mamun SH, Rahman MS. Explainable ensemble modeling for diabetes risk prediction using the real-world DiaBD cohort. In 2026 IEEE 2nd International Conference on Quantum Photonics, Artificial Intelligence & Networking (QPAIN). IEEE, 2026, April:1–6.