



Data-driven diabetes risk prediction using advanced ensemble learning methods

Chan Yeob Yeun

Department of Electrical Engineering, Computer Science Research, Center for Cyber-Physical Systems, Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates

Abstract

Diabetes mellitus is a major chronic metabolic disorder whose increasing prevalence has created a need for effective approaches to early risk identification and prevention. The rapid growth of electronic health records, population health surveys, and patient-generated health information has created opportunities for data-driven approaches to diabetes risk prediction. Machine learning can analyze large numbers of demographic, behavioral, and clinical variables to identify patterns associated with diabetes risk. Among these approaches, ensemble learning has demonstrated considerable potential because it combines multiple predictive models to improve robustness, generalization, and classification performance. This article examines the application of advanced ensemble learning methods for data-driven diabetes risk prediction, with particular attention to Random Forest, Gradient Boosting, XGBoost, LightGBM, CatBoost, voting, and stacking approaches. The proposed framework incorporates data preprocessing, feature selection, class-imbalance management, ensemble model development, hyperparameter optimization, and explainable artificial intelligence. Recent research indicates that advanced ensemble approaches can achieve strong predictive performance, although reported results vary considerably according to dataset characteristics, population, feature engineering, and validation strategy. A recent study using 253,680 observations from the Behavioral Risk Factor Surveillance System demonstrated the potential of combining XGBoost, Random Forest, CatBoost, LightGBM, and K-Nearest Neighbors for diabetes prediction after addressing class imbalance with SMOTE-ENN. The article argues that reliable diabetes risk prediction requires not only high predictive performance but also rigorous validation, interpretability, population diversity, and careful consideration of data quality.

Keywords: Diabetes mellitus, diabetes risk prediction, ensemble learning, machine learning, explainable artificial intelligence, class imbalance

Introduction

India's tribunal system has repeatedly generated the same constitutional anxiety: bodies created to exercise judicial or quasi-judicial power have often remained administratively dependent on the executive departments whose decisions Diabetes mellitus is a chronic metabolic condition that presents a substantial challenge to public health because of its increasing prevalence and its association with serious long-term complications. Early identification of individuals who are at elevated risk can support preventive interventions, lifestyle modification, clinical assessment, and appropriate management. Conventional risk assessment approaches often depend on a limited number of established clinical variables. However, modern healthcare systems generate increasingly large and heterogeneous datasets containing demographic characteristics, health behaviors, medical history, physiological measurements, and laboratory information. These data provide an opportunity to develop predictive models capable of identifying complex relationships associated with diabetes risk. Machine learning has emerged as an important computational approach for analyzing such health data. Unlike traditional statistical approaches that may impose assumptions about the relationship between predictors and outcomes, machine learning algorithms can capture nonlinear relationships and interactions among variables. Recent systematic research has identified a substantial body of literature investigating machine learning for diabetes prediction and has emphasized the importance of dataset characteristics, model selection, training strategies, and evaluation procedures in determining reported performance (Khokhar *et al.*, 2025) [2].

Ensemble learning represents an important development within machine learning because it combines predictions from multiple individual models. Rather than relying on a single classifier, ensemble methods can integrate different learners to improve predictive stability and generalization. A recent systematic review of ensemble learning in predictive healthcare found that stacking and hybrid ensemble approaches combined with feature selection were particularly prominent among high-performing studies, while also identifying continuing concerns regarding computational complexity, interpretability, class imbalance, and dataset bias (Saha *et al.*, 2026) [4]. The application of advanced ensemble learning to diabetes risk prediction is therefore important because diabetes-related datasets frequently contain nonlinear relationships, heterogeneous predictors, missing values, class imbalance, and potentially interacting risk factors. This article proposes a data-driven framework for using advanced ensemble models to address these challenges while incorporating explainability and rigorous evaluation.

Problem Statement

Despite the growing application of machine learning to diabetes prediction, several challenges remain. First, diabetes datasets can differ substantially in population characteristics, sample size, feature availability, and outcome definitions. Consequently, a model that produces strong performance on one dataset may not necessarily demonstrate equivalent performance when applied to another population. Second, diabetes prediction is often affected by class imbalance. If the number of non-diabetic

observations is considerably larger than the number of diabetic observations, a model may achieve high overall accuracy while failing to identify an important proportion of individuals at risk. Therefore, relying exclusively on accuracy can provide an incomplete assessment of predictive performance. Third, complex ensemble models can be difficult to interpret. Healthcare applications require greater transparency because users need to understand which variables contribute to predictions. Explainable artificial intelligence techniques such as SHAP can provide information about feature contributions and help researchers examine whether model behavior is consistent with clinically meaningful risk factors. These challenges demonstrate the need for a comprehensive data-driven framework that combines advanced ensemble learning, appropriate preprocessing, feature selection, class-imbalance management, robust validation, and explainability.

Aim of the Study

The aim of this study is to develop a data-driven framework for diabetes risk prediction using advanced ensemble learning methods. The framework is designed to investigate how different ensemble strategies can utilize demographic, behavioral, and clinical health indicators to identify individuals at increased risk of diabetes. The study further seeks to evaluate the predictive behavior of different ensemble algorithms using multiple performance measures and to investigate the contribution of individual risk factors through explainable artificial intelligence techniques.

Literature Review

Machine Learning in Diabetes Risk Prediction

Machine learning has been increasingly applied to diabetes prediction because of its ability to process large-scale health datasets and identify complex patterns. Predictive models can incorporate variables such as age, body mass index, blood pressure, glucose-related measurements, family history, physical activity, diet, smoking status, and other health indicators. Recent research demonstrates the potential of data-driven prediction using large population-level datasets. A 2025 study using the Behavioral Risk Factor Surveillance System dataset analyzed 253,680 survey responses containing 21 health-related features. The study applied SMOTE-ENN to address class imbalance and used XGBoost for feature selection before developing an ensemble involving XGBoost, Random Forest, CatBoost, LightGBM, and K-Nearest Neighbors. The resulting approach reported 96.40% accuracy on that dataset (Diabetes Prediction using Ensemble Learning Techniques, 2025) ^[1]. Such findings demonstrate the potential of combining multiple machine learning algorithms. However, reported accuracy should be interpreted within the context of the particular dataset and experimental methodology. High performance on a large benchmark dataset does not automatically establish that the model will perform similarly in another population.

Advanced Ensemble Learning

Ensemble learning combines multiple predictive models to obtain a final prediction. Bagging methods such as Random Forest construct multiple models independently and aggregate their outputs, whereas boosting methods such as Gradient Boosting, XGBoost, LightGBM, and CatBoost

build models sequentially to improve the weaknesses of previous learners. Random Forest is particularly useful for structured health data because it can model nonlinear relationships and interactions while providing measures of feature importance. XGBoost is a gradient-boosting method designed to optimize predictive performance while incorporating regularization and computational improvements. LightGBM is another gradient-boosting framework designed for efficient learning, particularly with large datasets. CatBoost is designed to handle categorical variables effectively and can be useful when healthcare datasets contain substantial categorical information. A recent study of Type 2 diabetes cardiovascular risk prediction developed a dynamic weighted ensemble containing Random Forest, Gradient Boosting Decision Tree, K-Nearest Neighbors, and XGBoost. The researchers argued that integrating models with complementary strengths could improve predictive accuracy and stability compared with relying on a single algorithm (Yuan *et al.*, 2025) ^[5].

Stacking and Hybrid Ensembles

Stacking represents an advanced ensemble strategy in which predictions from multiple base learners are provided to a meta-model that learns how to combine them. The approach can integrate models with different learning characteristics and may therefore capture complementary patterns in complex datasets. Recent evidence supports the increasing use of stacking and hybrid ensemble methods in healthcare prediction. A systematic review published in 2026 found that approximately 40% of the reviewed predictive-healthcare studies used stacking or hybrid ensemble models with feature selection, while noting that such models can face challenges involving computational cost and interpretability (Saha *et al.*, 2026) ^[4]. In diabetes prediction specifically, a 2026 population-based study developed an explainable stacking ensemble and reported an accuracy of 91.67% and an AUC of 0.967 within its study population. The authors also used SHAP to identify important predictors, including age, family history of diabetes, body mass index, weight, dietary behaviors, income, and gender (Resma *et al.*, 2026) ^[3]. These findings suggest that stacking can be useful when different algorithms capture complementary aspects of the data. Nevertheless, model performance must be validated carefully because the reported results are specific to the study population and methodology.

Proposed Data-Driven Framework

Data Collection

The proposed framework begins with the collection of relevant diabetes-related health data. Depending on availability, the dataset may be obtained from electronic health records, public health surveys, hospital databases, or validated public machine learning repositories. The predictors may include demographic, behavioral, lifestyle, and clinical characteristics. Variables such as age, body mass index, blood pressure, glucose measurements, family history, physical activity, diet, smoking behavior, and other health indicators can provide useful information for estimating diabetes risk. The target variable should be clearly defined before model development. For binary prediction, the target can represent diabetic versus non-diabetic status. Alternatively, the framework can be adapted

to multiple risk categories where the available data support meaningful risk stratification.

Data Preprocessing

Data preprocessing is essential for ensuring that the models receive reliable and consistent input. The dataset should first be examined for missing values, duplicated records, inconsistent measurements, extreme observations, and incorrectly encoded variables. Missing values can be treated using statistically appropriate imputation procedures. Categorical variables can be encoded into numerical representations when required by the selected algorithms. Although many tree-based ensemble algorithms are relatively insensitive to feature scaling, scaling may still be useful when the ensemble includes algorithms such as K-Nearest Neighbors or Support Vector Machines. Class imbalance requires particular attention. Resampling approaches such as SMOTE, undersampling, or hybrid methods such as SMOTE-ENN can be considered when appropriate. The 2025 BRFS-based ensemble study used SMOTE-ENN specifically to address class imbalance before developing its ensemble model (Diabetes Prediction using Ensemble Learning Techniques, 2025) ^[1]. Importantly, resampling should be performed within the training process rather than before separating the test dataset. This helps prevent information leakage and provides a more realistic estimate of model performance.

Feature Selection

Feature selection is another important component of the proposed framework. Large health datasets may contain redundant, weakly informative, or highly correlated variables. Removing irrelevant features can reduce model complexity and potentially improve generalization. Several approaches can be considered, including correlation analysis, recursive feature elimination, mutual information, statistical testing, permutation importance, and model-based feature importance. XGBoost and Random Forest can also provide information about feature contribution that can assist in identifying influential variables. Recent ensemble-learning research suggests that feature selection is frequently associated with strong performance in stacking and hybrid frameworks (Saha *et al.*, 2026) ^[4].

Advanced Ensemble Model Development

Random Forest

Random Forest provides a bagging-based baseline for the proposed framework. Multiple decision trees are constructed using different bootstrap samples and subsets of predictor variables. Their predictions are subsequently aggregated to produce the final classification. The primary advantage of Random Forest is its robustness and ability to model nonlinear relationships without requiring strong assumptions about the functional form of the data. It can also provide useful estimates of feature importance, making it suitable for exploratory analysis of diabetes risk factors.

Gradient Boosting

Gradient Boosting develops decision trees sequentially, with each new tree attempting to reduce the prediction errors of the existing ensemble. This sequential optimization allows the model to learn complex patterns that may not be captured by simpler models.

Gradient Boosting can be particularly useful when diabetes risk depends on interactions among several demographic, behavioral, and clinical variables. However, careful hyperparameter optimization is necessary to control model complexity and reduce overfitting.

XGBoost

XGBoost is one of the most widely used advanced boosting methods for structured datasets. It incorporates regularization and computational optimization into the gradient-boosting framework. Its ability to capture nonlinear relationships and interactions makes it particularly suitable for diabetes prediction. Recent diabetes research has repeatedly incorporated XGBoost into ensemble frameworks. A 2025 study combined XGBoost with Random Forest, CatBoost, LightGBM, and K-Nearest Neighbors for diabetes prediction using a large BRFS dataset (Diabetes Prediction using Ensemble Learning Techniques, 2025) ^[1]. Another recent study used XGBoost as part of a dynamic weighted ensemble alongside Random Forest, Gradient Boosting Decision Tree, and K-Nearest Neighbors, demonstrating the growing use of heterogeneous ensemble architectures for health prediction (Yuan *et al.*, 2025) ^[5].

LightGBM and CatBoost

LightGBM and CatBoost provide additional boosting algorithms for comparison. LightGBM is designed for efficient gradient boosting and can be particularly useful when dealing with large datasets. CatBoost is designed to handle categorical variables effectively and can therefore be useful for health datasets containing demographic and behavioral categories. The inclusion of several boosting algorithms makes it possible to construct heterogeneous ensembles in which models learn different aspects of the underlying data. Rather than assuming that one algorithm is universally optimal, the ensemble can combine their predictions to potentially improve robustness.

Stacking Ensemble

The proposed framework can integrate Random Forest, XGBoost, LightGBM, CatBoost, and Gradient Boosting through a stacking architecture. The base models independently generate predictions, and a meta-learner then learns how to combine these predictions. A carefully designed stacking framework should generate meta-level training data using cross-validation to avoid information leakage. The final meta-model can then learn which base-model predictions should receive greater importance for different observations. Recent population-based diabetes research provides evidence for the usefulness of stacking. Resma *et al.* (2026) ^[3] reported that their stacking ensemble achieved an AUC of 0.967, exceeding the performance of the individual models examined in that study.

Model Optimization

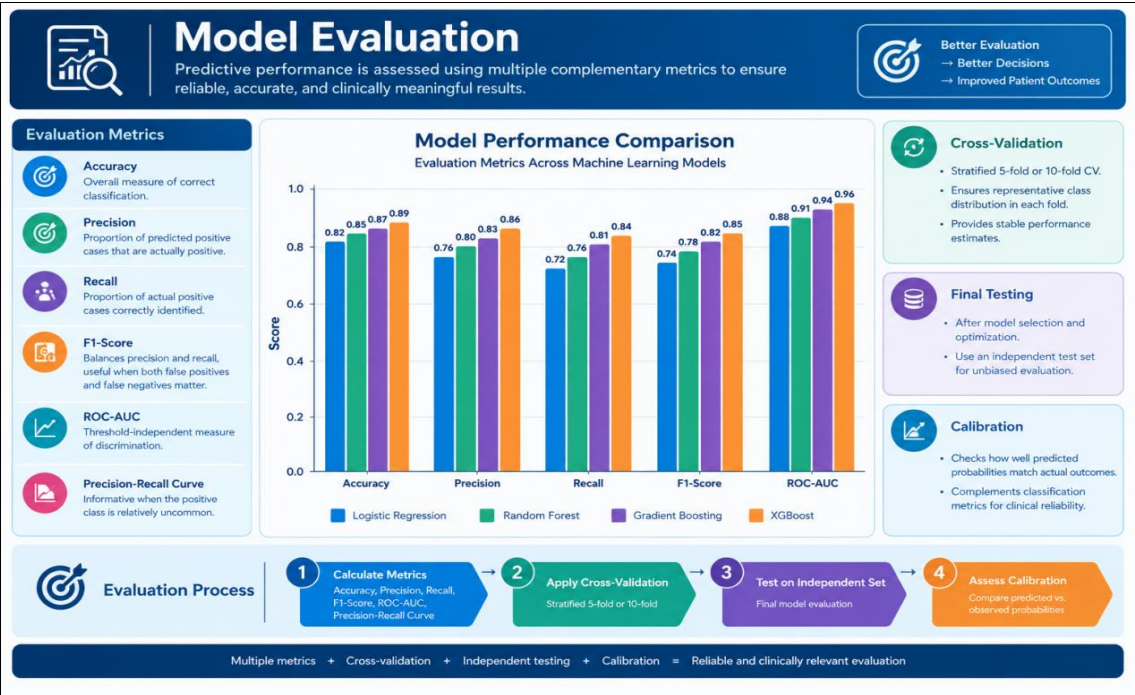
Hyperparameter optimization is an important stage in developing advanced ensemble models. Parameters such as the number of trees, maximum tree depth, learning rate, subsampling ratio, and regularization parameters can significantly influence predictive performance. Grid search, random search, Bayesian optimization, and evolutionary optimization can be considered for identifying appropriate parameter configurations. The optimization process should

be performed using the training data and cross-validation rather than the independent test set. A recent diabetes prediction study developed a GA-XGBoost stacking model and incorporated feature selection, resampling, XGBoost optimization, and SHAP-based interpretation into its methodology (Diabetes prediction model based on GA-XGBoost and stacking ensemble algorithm, 2024). This illustrates how optimization and ensemble integration can be combined within a broader data-driven framework.

Model Evaluation

The proposed framework should evaluate predictive performance using multiple complementary metrics. Accuracy provides an overall measure of correct classification, while precision measures the proportion of predicted positive cases that are actually positive. Recall measures the proportion of actual positive cases correctly identified by

the model. The F1-score combines precision and recall and can be useful when both false-positive and false-negative errors are important. ROC-AUC provides a threshold-independent measure of discrimination, while precision-recall curves can be particularly informative when the positive class is relatively uncommon. Cross-validation should be employed to assess the stability of the models. A stratified five-fold or ten-fold cross-validation procedure can help ensure that each fold contains a representative distribution of the target classes. After model selection and optimization, a completely independent test dataset should be used for final evaluation. The evaluation should also consider calibration. A diabetes risk model should ideally not only distinguish between higher- and lower-risk individuals but should also provide probabilities that correspond reasonably well to observed outcomes. Calibration analysis can therefore complement traditional classification metrics.



Explainable Artificial Intelligence

Explainability is a critical component of data-driven diabetes risk prediction. Advanced ensemble models can contain hundreds or thousands of individual decision rules, making it difficult to understand why a particular prediction was generated. SHAP is particularly useful for explaining ensemble models because it can estimate the contribution of individual features to a model's prediction. Global SHAP analysis can identify variables that have substantial influence across the dataset, while local SHAP explanations can show why the model generated a particular prediction for an individual observation. Recent research has demonstrated the usefulness of SHAP in diabetes prediction. Resma *et al.* (2026) [3] used SHAP to identify age, family history, body mass index, weight, dietary behaviors, income, and gender as important predictors within their population-based model. Explainability should nevertheless be interpreted carefully. A feature that contributes strongly to a prediction is not necessarily a causal determinant of diabetes. SHAP and related methods primarily explain model behavior rather than establish causal relationships.

Expected Results and Discussion

The proposed data-driven framework is expected to demonstrate that advanced ensemble methods can effectively capture complex patterns in diabetes-related health data. Random Forest may provide a stable and robust baseline, while boosting algorithms such as XGBoost, LightGBM, CatBoost, and Gradient Boosting may capture more complex nonlinear interactions. A heterogeneous stacking ensemble may potentially improve predictive performance by combining complementary information from different algorithms. Recent evidence supports this possibility. In the BRFSS-based study, an ensemble incorporating XGBoost, Random Forest, CatBoost, LightGBM, and K-Nearest Neighbors achieved 96.40% accuracy on the investigated dataset (Diabetes Prediction using Ensemble Learning Techniques, 2025) [1]. Similarly, a 2026 population-based study reported an AUC of 0.967 for a stacking ensemble, compared with lower AUC values for individual models in that particular study (Resma *et al.*, 2026) [3]. These findings provide evidence for the potential of ensemble learning but should not be interpreted as

universal performance benchmarks. Differences in population, dataset size, feature definitions, preprocessing, class balancing, and validation procedures can substantially influence reported results. The most meaningful outcome of the proposed framework would therefore be a comprehensive comparison of models under identical experimental conditions. Such an approach would make it possible to determine whether improvements result from the underlying algorithm, feature selection, class-balancing strategy, hyperparameter optimization, or ensemble combination.

Clinical and Research Significance

Data-driven diabetes prediction has potential value for early risk assessment and population-level screening research. By processing multiple health indicators simultaneously, ensemble models can identify combinations of characteristics associated with elevated diabetes risk. The integration of explainability can further improve the research value of these systems by providing information about the variables influencing predictions. This may assist researchers in identifying important patterns and examining whether the model's behavior corresponds to established knowledge about diabetes risk. However, machine learning prediction should be viewed as a decision-support approach rather than a replacement for professional medical diagnosis. Before deployment in healthcare settings, predictive models require external validation, appropriate calibration, assessment of population fairness, privacy protections, and prospective evaluation.

Challenges and Limitations

Several limitations may affect the development of data-driven ensemble models for diabetes prediction. Dataset bias is an important concern because models trained using a particular population may not generalize to individuals with different demographic, socioeconomic, or clinical characteristics. A recent review of ensemble learning in healthcare identified dataset bias and class imbalance as continuing challenges for predictive healthcare applications (Saha *et al.*, 2026) [4]. Another limitation is the possibility of data leakage. Incorrectly performing feature selection, oversampling, or preprocessing before cross-validation can result in overly optimistic performance estimates. Careful experimental design is therefore necessary. Computational complexity can also increase when several advanced algorithms are combined into a stacking or hybrid ensemble. Although improved predictive performance may justify additional complexity in research settings, practical implementation requires consideration of computational resources, inference time, maintainability, and interpretability. Finally, high accuracy does not guarantee clinical usefulness. A model should be evaluated using independent populations and clinically meaningful measures before being considered for real-world application.

Future Research Directions

Future research can extend this framework by incorporating longitudinal patient data rather than relying exclusively on cross-sectional observations. Longitudinal information could allow models to learn how risk factors and health indicators change over time. Another promising direction involves multimodal data integration. Combining clinical records with behavioral, dietary, laboratory, wearable-

device, and other health-related information may provide a more comprehensive representation of diabetes risk. Research on precision diabetes care increasingly emphasizes the potential of integrating heterogeneous health data, although privacy, interoperability, and generalization remain important challenges. Adaptive ensemble models may also be investigated to allow predictive systems to respond to changes in patient characteristics and healthcare environments. Recent reviews have highlighted adaptive and explainable ensemble learning as important areas for future predictive-healthcare research (Saha *et al.*, 2026) [4]. External validation should also become a central component of future studies. Instead of evaluating a model using only one dataset, researchers should test models across multiple populations and healthcare settings. Such validation would provide stronger evidence regarding model generalizability and robustness.

Conclusion

Data-driven diabetes risk prediction using advanced ensemble learning methods represents a promising direction for machine learning-based healthcare research. The increasing availability of demographic, behavioral, clinical, and population-level health data provides an opportunity to identify complex patterns associated with diabetes risk. Advanced ensemble methods such as Random Forest, Gradient Boosting, XGBoost, LightGBM, CatBoost, voting, and stacking can provide complementary predictive capabilities. Recent studies demonstrate that combining these algorithms can produce strong predictive performance, although the reported results remain dependent on the dataset and experimental methodology. A reliable diabetes prediction framework should therefore extend beyond algorithm selection. Data quality, preprocessing, class-imbalance management, feature selection, hyperparameter optimization, cross-validation, calibration, and explainability should all be incorporated into the modeling process. SHAP and related explainable AI techniques can provide valuable insight into model behavior, while external validation can determine whether predictions generalize beyond the original development population. Ultimately, advanced ensemble learning provides a strong methodological foundation for data-driven diabetes risk prediction, but its successful application depends on rigorous validation, transparent reporting, representative data, and responsible integration with clinical decision-making.

References

1. Diabetes Prediction using Ensemble Learning Techniques. *Procedia Computer Science*,2025;258:3155–3164.
2. Khokhar PB, Gravino C, Palomba F. Advances in artificial intelligence for diabetes prediction: Insights from a systematic literature review. *Artificial Intelligence in Medicine*,2025;164:103132.
3. Resma, *et al.* 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. Saha R, Saif S, Khan T, Alabdultif A. A critical systematic review of ensemble learning methods for predictive healthcare. *Discover Computing*,2026;29:117.

5. 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.
6. Dinh A, Miertschin S, Young A, Mohanty SD. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Medical Informatics and Decision Making*, 2019;19(1):211.
7. Ganie SM, Pramanik PKD, Bashir Malik M, Mallik S, Qin H. An ensemble learning approach for diabetes prediction using boosting techniques. *Frontiers in Genetics*, 2023;14:1252159.
8. Al Reshan MS, Amin S, Zeb MA, Sulaiman A, Alshahrani H, Shaikh A, *et al.* An innovative ensemble deep learning clinical decision support system for diabetes prediction. *IEEE Access*, 2024;12:106193–106210.
9. Bhatta RP. Ensemble based machine learning model for prediction of diabetes. *NPRC Journal of Multidisciplinary Research*, 2026;3(2):116–129.
10. Gogoi LVN, Bose R, Majumder S, Majumder S. A statistically validated ensemble learning framework for diabetes prediction using clinical and metabolic features. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 2026;15(1):198.
11. Vyas H, Aggarwal M, Goyal K, Gupta K, Verma K. Enhancing Diabetes Risk Assessment with Ensemble Learning Methods. In: *International Conference on Advanced Informatics for Computing Research*. Cham: Springer Nature Switzerland, 2023, 66–77.
12. Shobha T, Pradeep S, Patil S, Subramanya J, Mitra GS. A data-driven machine learning model for effective diabetes diagnosis. *Scientific Reports*, 2026.
13. Qin Y, Wu J, Xiao W, Wang K, Huang A, Liu B, *et al.* Machine learning models for data-driven prediction of diabetes by lifestyle type. *International Journal of Environmental Research and Public Health*, 2022;19(22):15027.
14. Hasib E, Rahman S, Imtiaz MS, Hamim SA, Mridha MF. A Data-Driven Approach to Type 2 Diabetes Prediction Using Machine Learning Techniques. In: *2026 IEEE 2nd International Conference on Quantum Photonics, Artificial Intelligence & Networking (QPAIN)*. IEEE, 2026, 1–6.