



Performance analysis of bagging and boosting algorithms for diabetes prediction

Robin Sommer

International Computer Science Institute (ICSI), Lawrence Berkeley National Laboratory (LBNL), Berkeley, California, USA

Abstract

Diabetes mellitus is a major chronic disease that requires early identification to support timely intervention and reduce the risk of long-term complications. The increasing availability of electronic health records, demographic information, lifestyle indicators, and clinical measurements has created opportunities for machine learning-based diabetes prediction. Among machine learning approaches, ensemble learning has attracted considerable attention because it combines multiple predictive models to improve classification performance and generalization. This article presents a performance-oriented analysis of two important ensemble learning strategies, namely bagging and boosting, for diabetes prediction. Bagging methods such as Random Forest construct multiple decision trees independently using different bootstrap samples, while boosting methods such as Gradient Boosting and Extreme Gradient Boosting (XGBoost) sequentially develop weak learners to correct previous prediction errors. Recent studies indicate that both approaches can effectively model nonlinear relationships in diabetes-related data, although their performance depends on data quality, class distribution, feature selection, preprocessing, and validation methodology. The proposed analysis focuses on comparing bagging and boosting according to accuracy, precision, recall, F1-score, receiver operating characteristic area under the curve, computational requirements, robustness, and interpretability. The study provides a methodological framework for identifying the strengths and limitations of these ensemble approaches and emphasizes the importance of evaluating models using multiple performance measures rather than relying exclusively on accuracy.

Keywords: Diabetes prediction, ensemble learning, bagging, boosting, machine learning

Introduction

Diabetes mellitus is a chronic metabolic disorder characterized by abnormal blood glucose regulation and represents an important challenge for healthcare systems. Early identification of individuals who are at increased risk can support preventive interventions and further clinical assessment. Traditional approaches to diabetes assessment generally rely on clinical measurements and diagnostic procedures, but machine learning has increasingly been investigated as a complementary approach for identifying patterns associated with diabetes risk.

Machine learning algorithms can process demographic, behavioral, and clinical variables and learn relationships between patient characteristics and diabetes outcomes. A recent systematic review of 53 studies reported extensive use of machine learning methods for diabetes prediction and highlighted the importance of appropriate datasets, validation strategies, performance measures, and interpretability techniques (Khokhar *et al.*, 2025)^[2].

Among these approaches, ensemble learning is particularly relevant because it combines multiple predictive learners rather than depending on a single model. Two major ensemble strategies are bagging and boosting. Bagging constructs several models independently and combines their predictions, primarily reducing variance and improving stability. Random Forest is one of the most widely used bagging algorithms and has been applied extensively to healthcare prediction problems. Boosting, in contrast, builds models sequentially, with subsequent learners concentrating on errors made by previous learners. Gradient Boosting and XGBoost are important examples of this approach and have demonstrated strong performance in structured healthcare datasets.

Recent diabetes studies provide evidence that both Random Forest and boosting algorithms can capture complex relationships among demographic, behavioral, and clinical variables. For example, a recent population-based study comparing several machine learning algorithms reported an accuracy of 85.65% and an AUC of 0.939 for Random Forest, while XGBoost achieved an accuracy of 88.69% and an AUC of 0.955 in the study dataset (Resma *et al.*, 2026)^[3]. These findings demonstrate why a systematic comparison between bagging and boosting is relevant for diabetes prediction.

Problem Statement

Although numerous machine learning algorithms have been applied to diabetes prediction, selecting an appropriate model remains challenging because algorithmic performance varies according to the characteristics of the dataset and the experimental methodology. Diabetes datasets may contain missing observations, imbalanced classes, correlated variables, categorical attributes, and nonlinear relationships between risk factors and outcomes. Consequently, an algorithm that performs strongly on one dataset may not produce the same performance on another dataset.

A further challenge is that studies sometimes emphasize accuracy without sufficiently considering recall, precision, F1-score, AUC, calibration, and generalization. In medical prediction, a model that achieves high overall accuracy may still perform poorly in identifying individuals belonging to a minority class. Therefore, a meaningful performance analysis should examine multiple evaluation criteria and consider the influence of preprocessing, class balancing, feature selection, and cross-validation.

Aim of the Study

The aim of this study is to conduct a comparative performance analysis of bagging and boosting algorithms for diabetes prediction. The study focuses on determining how Random Forest as a representative bagging technique compares with boosting algorithms such as Gradient Boosting and XGBoost when applied to structured diabetes-related data.

The analysis considers predictive accuracy, precision, recall, F1-score, ROC-AUC, computational efficiency, robustness, and model interpretability. Rather than assuming that one ensemble strategy is universally superior, the study evaluates how each approach behaves under the same experimental conditions.

Literature Review

Machine Learning for Diabetes Prediction

Machine learning has become an important research area in diabetes prediction because it can identify complex relationships among patient characteristics that may be difficult to represent using conventional statistical approaches. Variables frequently investigated in diabetes prediction include age, body mass index, blood pressure, glucose-related measurements, family history, physical activity, dietary characteristics, and other demographic or behavioral indicators.

Khokhar *et al.* (2025) ^[2] conducted a systematic literature review of machine learning applications for diabetes prediction and identified substantial variation in datasets, algorithms, training strategies, and evaluation procedures. Their review emphasized that differences in datasets and validation methods make direct comparisons between reported performance values difficult.

More recent research has continued to demonstrate the usefulness of ensemble methods. A 2026 ^[8] study evaluating Random Forest, XGBoost, and other machine learning models found that XGBoost produced strong performance for diabetes classification, while Random Forest also demonstrated robust predictive capability (Resma *et al.*, 2026) ^[3].

Bagging Algorithms

Bagging, or bootstrap aggregation, is an ensemble learning strategy designed primarily to reduce model variance. In a typical bagging procedure, multiple bootstrap samples are generated from the original training dataset. A separate model is trained on each sample, and their predictions are subsequently aggregated.

Random Forest extends the bagging concept by constructing multiple decision trees while introducing additional randomness through feature selection. The independent trees contribute to the final prediction through majority voting for classification tasks. This process can improve stability and reduce the tendency of individual decision trees to overfit.

The suitability of Random Forest for diabetes prediction is associated with its ability to model nonlinear relationships and interactions between predictor variables. Recent diabetes research has used Random Forest as an important benchmark against boosting algorithms. In a population-based diabetes study, Random Forest achieved an AUC of 0.939, indicating strong discrimination within that particular dataset (Resma *et al.*, 2026) ^[3].

Boosting Algorithms

Boosting follows a different ensemble strategy. Instead of developing learners independently, boosting constructs

models sequentially. Each new learner attempts to improve the weaknesses of the existing ensemble. Gradient Boosting develops a sequence of decision trees by minimizing a specified loss function, while XGBoost extends gradient boosting with additional optimization and regularization mechanisms.

XGBoost has become particularly prominent in structured healthcare prediction because it can model nonlinear relationships and interactions while incorporating regularization to control model complexity. Recent research on diabetes classification has compared XGBoost with Random Forest, Logistic Regression, Support Vector Machines, Decision Trees, and K-Nearest Neighbors. In one 2026 study, XGBoost obtained $98.79 \pm 1.13\%$ mean five-fold cross-validation accuracy, while Random Forest achieved $98.30 \pm 0.79\%$ in the same experimental setting (El Sherbiny *et al.*, 2026) ^[1].

These results should not be interpreted as evidence that XGBoost will always outperform Random Forest because the reported values are specific to the dataset, preprocessing procedure, classification task, and validation strategy used in that study.

Comparative Evidence

Recent research has directly examined bagging and boosting strategies for diabetes risk prediction. A 2025 ^[5] study evaluated bagging, boosting, and other ensemble approaches across three diabetes-related datasets, including the CDC Diabetes Health Indicators dataset, the Early Stage Diabetes Risk Prediction System dataset, and the PIMA Indian Diabetes dataset. The researchers also examined the influence of SMOTE and feature selection on model performance. Their findings demonstrated that performance varied across datasets and algorithms, emphasizing the importance of dataset characteristics and preprocessing when comparing ensemble approaches.

Another recent study using the BRFSS Health Indicator dataset investigated an ensemble comprising XGBoost, Random Forest, CatBoost, LightGBM, and KNN. The study reported 96.40% accuracy for its proposed ensemble approach, illustrating the potential of combining multiple learners for diabetes prediction (Diabetes Prediction using Ensemble Learning Techniques, 2025) ^[5].

Proposed Methodology

Data Collection

The proposed study can use a publicly available diabetes dataset containing demographic, behavioral, and clinical health indicators. Candidate variables may include age, body mass index, blood pressure, glucose-related measurements, family history, physical activity, smoking status, and other relevant patient characteristics. The dependent variable should represent diabetes status or an appropriately defined diabetes-risk category.

The selected dataset should be divided into training and testing subsets before model development. A stratified division is preferable for classification because it helps preserve the distribution of the target classes between the training and testing data.

Data Preprocessing

Data preprocessing is an essential stage because poor-quality input data can negatively influence model performance. Missing values should be investigated and

treated using an appropriate imputation strategy. Categorical variables may require encoding, while numerical variables should be examined for abnormal or implausible values.

Class imbalance should also be investigated. If one diabetes class is substantially smaller than another, the model may become biased toward the majority class. Techniques such as class weighting or Synthetic Minority Oversampling Technique can be considered when appropriate. Recent diabetes research has demonstrated that SMOTE can influence predictive performance, particularly when class imbalance is substantial (Enhancing Diabetes Risk Prediction, 2025)^[5].

Importantly, preprocessing and resampling operations should be performed within the training data or cross-validation folds rather than using information from the independent test set. This prevents information leakage and produces a more reliable estimate of model generalization.

Feature Selection

Feature selection can be applied to identify variables that contribute most strongly to diabetes prediction. Correlation analysis, recursive feature elimination, feature importance, permutation importance, or model-based selection can be considered.

Random Forest provides measures of feature importance that can assist in identifying influential predictors, while boosting algorithms can similarly provide information regarding the contribution of individual variables. However, feature importance should not automatically be interpreted as causal evidence. A variable may be predictive without directly causing diabetes.

Bagging Model

Random Forest is proposed as the principal bagging algorithm. Multiple decision trees are trained using bootstrap samples and randomly selected subsets of features. The final diabetes prediction is determined by aggregating the predictions generated by the individual trees.

The primary advantage of this approach is improved stability compared with a single decision tree. Because the trees are trained independently, errors made by one tree do not necessarily dominate the entire model. Random Forest can therefore provide a robust baseline for comparison with boosting approaches.

Boosting Models

Gradient Boosting and XGBoost are proposed as representative boosting algorithms. Gradient Boosting develops decision trees sequentially, with each successive tree attempting to reduce the errors of the current ensemble. XGBoost extends this framework through regularization and computational optimizations that can improve efficiency and control model complexity.

The boosting models should be trained using the same training dataset and evaluated using the same validation strategy as the Random Forest model. Hyperparameters such as tree depth, learning rate, number of estimators, and subsampling parameters should be optimized using cross-validation rather than tuning against the independent test set.

Performance Evaluation

A comprehensive comparison should use several performance metrics. Accuracy measures the proportion of

correctly classified observations, but it can be misleading when the classes are imbalanced. Precision measures the proportion of predicted positive cases that are actually positive, whereas recall measures the proportion of actual positive cases correctly identified by the model.

The F1-score provides a combined measure of precision and recall and can be particularly useful when both types of classification errors are important. ROC-AUC measures the ability of the model to discriminate between classes across different classification thresholds. For imbalanced datasets, precision-recall analysis and class-specific measures should also be considered.

Cross-validation should be used to examine model stability. Five-fold or ten-fold stratified cross-validation can provide multiple training and validation partitions and reduce dependence on a single data split. A final independent test set should remain untouched until the end of the model-development process.

Recent diabetes research illustrates the importance of cross-validation. El Sherbiny *et al.* (2026)^[1], for example, reported both holdout and five-fold cross-validation results and noted differences between individual split performance and cross-validated performance. This demonstrates why model evaluation should not rely solely on a single accuracy value.

Comparative Analysis of Bagging and Boosting

Bagging and boosting differ fundamentally in how their component models are constructed. Bagging develops learners independently and aggregates their predictions, making it particularly effective for reducing variance and improving stability. Random Forest is therefore generally robust to variations in individual training samples and can perform effectively on complex tabular datasets.

Boosting, on the other hand, develops learners sequentially and focuses subsequent models on errors generated by earlier learners. This can allow boosting algorithms to capture complex patterns and achieve strong predictive discrimination. However, their sequential structure can make hyperparameter selection more important and may increase computational complexity.

Recent empirical evidence illustrates these differences. In a 2026 population-based study, XGBoost achieved an AUC of 0.955 compared with 0.939 for Random Forest, although both models performed strongly within the study population (Resma *et al.*, 2026). In another 2026^[3] study, XGBoost achieved a mean cross-validation accuracy of 98.79%, compared with 98.30% for Random Forest (El Sherbiny *et al.*, 2026)^[1]. These results support the value of comparing both strategies, while also demonstrating that performance differences should be interpreted within the context of the dataset and experimental design.

Explainability and Feature Importance

Interpretability is particularly important when machine learning models are used in healthcare-related research. Although Random Forest and boosting algorithms can provide feature importance measures, more advanced explainability methods can be used to understand individual predictions.

SHAP, or SHapley Additive exPlanations, is one approach that can explain how individual features contribute to a model's prediction. Recent diabetes research has applied SHAP to ensemble models to identify influential risk factors

and provide both global and individual-level explanations. In a population-based study, SHAP analysis identified variables including age, family history, body mass index, weight, diet, income, and gender among important predictors in the investigated dataset (Resma *et al.*, 2026)^[3]. Explainability does not mean that a predictive relationship represents a causal relationship. Instead, it helps researchers and healthcare professionals understand which variables contributed to a model's prediction. This distinction is important when interpreting machine learning results in health-related applications.

Expected Results and Discussion

The proposed comparative analysis is expected to demonstrate that both bagging and boosting algorithms can provide useful predictive performance for diabetes classification. Random Forest is expected to provide a stable and relatively robust baseline because of its bootstrap aggregation and randomized feature selection. Boosting algorithms such as Gradient Boosting and XGBoost may demonstrate stronger predictive discrimination when the dataset contains complex nonlinear relationships and interactions.

However, the expected relative performance should not be treated as predetermined. Differences between models may result from dataset size, feature characteristics, class imbalance, preprocessing techniques, hyperparameter optimization, and validation strategy. The recent literature shows that XGBoost and Random Forest can both perform strongly, but their relative performance varies across datasets and experimental settings.

The analysis should therefore emphasize the complete performance profile of each algorithm. A model with slightly higher accuracy but substantially lower recall may not be preferable for a screening-oriented application. Similarly, a model with strong predictive performance but poor calibration or excessive computational requirements may present limitations for practical deployment.

Significance of the Study

A comparative analysis of bagging and boosting algorithms can contribute to the development of more reliable machine learning approaches for diabetes prediction. By examining multiple evaluation metrics under a common experimental framework, the study can provide a clearer understanding of the circumstances under which different ensemble strategies perform effectively.

The study may also support future research into explainable and clinically relevant prediction systems. Recent systematic reviews emphasize that future diabetes prediction research should not focus exclusively on predictive accuracy but should also address interpretability, validation, dataset diversity, and integration with healthcare practice (Khokhar *et al.*, 2025)^[2].

Limitations

Several limitations should be considered when conducting this analysis. First, the performance of machine learning algorithms is strongly dependent on the characteristics and representativeness of the dataset. A model trained on one population may not generalize effectively to another population with different demographic, socioeconomic, behavioral, or clinical characteristics.

Second, preprocessing and class-balancing procedures can substantially influence reported performance. Consequently, comparisons should use identical data-processing procedures and evaluation protocols wherever possible. Third, high predictive performance on a public benchmark dataset does not necessarily establish clinical usefulness. External validation using independent populations is required before a prediction model can be considered reliable for broader healthcare applications.

Finally, feature importance should be interpreted carefully because predictive association does not establish causality. Explainability techniques can clarify model behavior but cannot replace clinical validation.

Conclusion

This article presents a professional framework for analyzing the performance of bagging and boosting algorithms for diabetes prediction. Random Forest represents a major bagging approach that improves predictive stability by combining independently trained decision trees, whereas Gradient Boosting and XGBoost represent boosting strategies that sequentially improve predictive performance by focusing on previous errors. Current research demonstrates that both approaches can achieve strong performance in diabetes prediction, although their relative effectiveness depends on the dataset, preprocessing strategy, class distribution, feature characteristics, and validation methodology.

A rigorous comparison should therefore evaluate accuracy, precision, recall, F1-score, ROC-AUC, computational efficiency, stability, and interpretability rather than relying on a single metric. Recent evidence shows strong performance from both Random Forest and XGBoost in different diabetes datasets, supporting the value of evaluating bagging and boosting under consistent experimental conditions. Ultimately, a carefully validated comparative framework can contribute to the development of reliable machine learning models for early diabetes risk assessment and provide a foundation for future research involving explainable and clinically validated artificial intelligence systems.

References

1. El Sherbiny MM, Abdelfattah MG, Takieldean AE, El-Din Mostafa H, Nafea HB. An XGBoost-based predictive framework for diabetes mellitus multi-classification. *Scientific Reports*,2026;16:25519.
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. Enhancing diabetes risk prediction, A comparative evaluation of bagging, boosting, and ensemble classifiers with SMOTE oversampling. *Informatics in Medicine Unlocked*,2025;57:101661.
5. Diabetes prediction using ensemble learning techniques. *Procedia Computer Science*,2025;258:3155–3164.
6. Taser PY. Application of bagging and boosting approaches using decision tree-based algorithms in

- diabetes risk prediction. In Proceedings. MDPI,2021:74(1):6.
7. Oliullah K, Rasel MH, Islam MM, Islam MR, Wadud MAH, Whaiduzzaman M. A stacked ensemble machine learning approach for the prediction of diabetes. *Journal of Diabetes & Metabolic Disorders*,2024;23(1):603–617.
 8. Bhatta RP. Ensemble based machine learning model for prediction of diabetes. *NPRC Journal of Multidisciplinary Research*,2026;3(2):116–129.
 9. Yadav DC, Pal S. An experimental study of diversity of diabetes disease features by bagging and boosting ensemble method with rule-based machine learning classifier algorithms. *SN Computer Science*,2021;2(1):50.
 10. Perveen S, Shahbaz M, Guergachi A, Keshavjee K. Performance analysis of data mining classification techniques to predict diabetes. *Procedia Computer Science*,2016;82:115–121.
 11. Sai MJ, Chettri P, Panigrahi R, Garg A, Bhoi AK, Barsocchi P. An ensemble of light gradient boosting machine and adaptive boosting for prediction of type-2 diabetes. *International Journal of Computational Intelligence Systems*,2023;16(1):14.
 12. Alam MS, Ferdous M, Neera NS. Enhancing diabetes prediction, An improved boosting algorithm for diabetes prediction. *International Journal of Advanced Computer Science & Applications*,2024;15(5):1273.
 13. Anbananthen KSM, Busst MBMA, Kannan R, Kannan S. A comparative performance analysis of hybrid and classical machine learning method in predicting diabetes. *Emerging Science Journal*,2023;7(1):102–115.
 14. Joshi R, Alehegn M. Analysis and prediction of diabetes diseases using machine learning algorithm, Ensemble approach. *International Research Journal of Engineering and Technology*,2017;4(10):426–435.
 15. Chandramouli A, Hyma VR, Tanmayi PS, Santoshi TG, Priyanka BJEC. Diabetes prediction using hybrid bagging classifier. *Entertainment Computing*,2023;47:100593.