IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double- Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, July 2026
Impact Factor: 8.2

A Review of Xgboost and Gradient Boosting Techniques for Heart Disease and Stroke Prediction

Manoj Kumar Yadav[1] **and Dr. Nilesh Vasant Ingale**[2]
[1]Research Scholar, [2]Associate Professor
Department of Computer Science
Vikrant University, Gwalior, M.P, India

Abstract: *Cardiovascular diseases, including heart disease and stroke, remain major causes of mortality and disability worldwide. Early prediction of cardiovascular risk is therefore essential for timely prevention, diagnosis, and clinical intervention. In recent years, machine learning techniques have increasingly been applied to healthcare data to improve the prediction of heart disease and stroke. Among these approaches, Gradient Boosting Machines and Extreme Gradient Boosting (XGBoost) have attracted considerable attention because of their ability to model complex, nonlinear relationships among clinical, demographic, lifestyle, laboratory, and imaging variables. This review examines the principles, applications, advantages, limitations, and comparative performance of XGBoost and gradient boosting techniques in heart disease and stroke prediction.*

Keywords: XGBoost, Gradient Boosting, Heart Disease Prediction, Stroke Prediction, Machine Learning.

I. INTRODUCTION

Heart disease and stroke constitute major public health challenges because they are associated with substantial mortality, morbidity, disability, and healthcare expenditure. The development of cardiovascular disease is influenced by multiple interacting factors, including age, sex, hypertension, diabetes, obesity, smoking, dyslipidemia, family history, physical inactivity, dietary habits, and previous medical conditions. Conventional risk prediction approaches often rely on statistical models that assume relatively simple relationships among predictors. However, real-world clinical data frequently contain nonlinear associations, interactions, missing values, and complex patterns that may not be adequately captured by traditional methods.

Machine learning has emerged as an important approach for analysing such complex healthcare datasets. Supervised machine learning algorithms can learn relationships between patient characteristics and clinical outcomes from historical data and subsequently estimate risk for new individuals. Decision trees, random forests, support vector machines, neural networks, logistic regression, and boosting algorithms have all been investigated for cardiovascular prediction.

Among ensemble methods, gradient boosting has demonstrated particular promise. Gradient boosting combines many weak learners, usually decision trees, into a strong predictive model. Each successive learner attempts to reduce errors made by the preceding model. XGBoost, or Extreme Gradient Boosting, is an advanced implementation of gradient-boosted decision trees that introduces regularization and computational optimization. These characteristics make it particularly suitable for structured tabular healthcare datasets.

A major meta-analysis of machine learning for cardiovascular disease prediction reported that boosting algorithms showed strong overall predictive ability, with a pooled AUC of 0.88 for coronary artery disease prediction and 0.91 for stroke prediction. However, the investigators also emphasized substantial methodological heterogeneity across studies, indicating that algorithm performance should not be interpreted independently of data characteristics and study design (Krittanawong et al., 2020).

Recent reviews have similarly highlighted the growing importance of machine learning in cardiovascular prediction while emphasizing challenges related to data quality, class imbalance, explainability, ethical concerns, and clinical implementation.

II. MACHINE LEARNING IN CARDIOVASCULAR PREDICTION

Machine learning models identify patterns from data rather than relying exclusively on manually specified relationships. In cardiovascular prediction, the input variables may include demographic information, medical history, laboratory measurements, physiological variables, lifestyle characteristics, electrocardiographic findings, imaging information, and electronic health record data.

The general machine learning workflow includes:

1. Collection of clinical and demographic data.
2. Data cleaning and preprocessing.
3. Management of missing values.
4. Encoding of categorical variables.
5. Feature selection or feature engineering.
6. Division into training and testing datasets.
7. Model development.
8. Hyperparameter optimization.
9. Internal and external validation.
10. Assessment of discrimination, calibration, and clinical usefulness.
11. Interpretation of model predictions.

The increasing availability of electronic health records and large clinical databases has created opportunities for predictive models that can integrate multiple risk factors simultaneously. However, high predictive accuracy alone does not guarantee clinical usefulness. Models should also demonstrate adequate calibration, reproducibility, transparency, fairness, and external validity.

III. CONCEPT OF GRADIENT BOOSTING

Gradient boosting is an ensemble learning technique in which several weak predictive models are combined sequentially to create a strong model. In most applications, the weak learners are shallow decision trees.

The process begins with an initial prediction. The model then calculates the residual errors between predicted and actual outcomes. A new decision tree is trained to predict these errors. This process continues iteratively, with each new learner improving the overall prediction.

The general principle can be expressed as:

Final Model = Initial Model + Sequential Error-Correcting Weak Learners

The method is called "gradient boosting" because the algorithm minimizes a specified loss function through an optimization process based on gradient information.

Important hyperparameters include:

1. Number of estimators
2. Learning rate
3. Maximum tree depth
4. Minimum samples per leaf
5. Subsampling rate
6. Loss function
7. Tree complexity

Gradient boosting is powerful because it can capture nonlinear relationships and interactions among variables. However, poorly tuned models may overfit the training dataset.

Recent cardiovascular machine learning literature describes gradient boosting as a sequential approach in which each new model is trained to correct residual errors from earlier learners, allowing the algorithm to capture complex predictive patterns.

Table 1. Basic Characteristics of Conventional Gradient Boosting

Characteristic	Description
Learning strategy	Sequential ensemble learning
Base learner	Usually decision trees
Main objective	Reduce errors of previous learners
Strength	Captures nonlinear relationships
Data type	Particularly effective for structured/tabular data
Major limitation	Can overfit if poorly tuned
Important parameters	Learning rate, tree depth, estimators, subsampling
Clinical application	Risk prediction, classification and outcome prediction

IV. EXTREME GRADIENT BOOSTING (XGBOOST)

XGBoost is an optimized implementation of gradient boosting designed for high predictive performance and computational efficiency. It extends conventional gradient boosting through several important improvements.

First, XGBoost incorporates regularization into the objective function. Regularization penalizes excessive model complexity and helps reduce overfitting. Second, XGBoost supports efficient parallel computation, making it suitable for large datasets. Third, it provides effective handling of missing values. Fourth, it includes tree-pruning and optimization procedures that improve computational efficiency.

The objective function in XGBoost includes two major components:

Objective Function = Training Loss + Regularization Penalty

The loss function measures prediction error, whereas the regularization component penalizes unnecessary model complexity. This balance is particularly important in healthcare data, where datasets may contain numerous correlated variables and a limited number of positive clinical events.

XGBoost has been widely used for classification, regression, ranking, and risk prediction. In cardiovascular applications, it can process variables such as age, blood pressure, cholesterol, glucose, body mass index, smoking status, previous cardiovascular history, and laboratory measurements.

A recent systematic review of heart disease prediction literature describes XGBoost as an extension of gradient-boosted decision trees that incorporates regularization, parallel computation, and mechanisms for handling missing data, making it attractive for structured clinical datasets.

Table 2. Comparison Between Gradient Boosting and XGBoost

Feature	Conventional Gradient Boosting	XGBoost
Basic approach	Sequential boosting	Optimized gradient boosting
Regularization	Limited/basic	Explicit L1 and L2 regularization
Computational efficiency	Moderate	High
Parallel processing	Limited	Supported
Missing value handling	Often requires preprocessing	Built-in handling mechanisms
Scalability	Moderate	High
Hyperparameter control	Standard	Extensive
Overfitting control	Learning rate and tree constraints	Regularization, subsampling and tree constraints
Healthcare suitability	High	Very high for structured clinical data
Interpretability	Moderate	Moderate to high with SHAP and feature importance

IJARSCT
ISSN: 2581-9429
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double- Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Volume 6, Issue 4, July 2026
Impact Factor: 8.2

V. APPLICATION OF GRADIENT BOOSTING IN HEART DISEASE PREDICTION

Heart disease prediction generally involves identifying individuals who have existing disease or who are at increased risk of future cardiovascular events. Common predictors include age, sex, blood pressure, cholesterol, diabetes status, chest pain characteristics, electrocardiographic variables, exercise-related findings, and lifestyle factors.

Gradient boosting is particularly useful because the influence of these variables is rarely linear. For example, the relationship between age and cardiovascular risk may interact with diabetes, hypertension, smoking, and cholesterol levels. Tree-based boosting models can automatically identify such complex interactions.

Studies comparing machine learning algorithms have reported strong performance for boosting approaches, although the best algorithm varies across datasets. A systematic survey of heart disease prediction studies found that model performance depends substantially on the dataset, feature set, preprocessing methods, and validation procedures rather than on a single universally superior algorithm.

In a systematic review of ischemic heart disease prediction, XGBoost was identified as a scalable boosting method applicable to coronary artery disease and myocardial infarction prediction. The review described evidence showing strong predictive performance when larger clinical datasets were used.

More recent cardiovascular research has also reported high performance for optimized XGBoost models. For example, a 2025 study combining multiple feature selection and improved particle swarm optimization reported improved performance compared with a baseline XGBoost model, illustrating the potential value of feature selection and hyperparameter optimization.

VI. APPLICATION OF XGBOOST IN CORONARY ARTERY DISEASE PREDICTION

Coronary artery disease prediction is an important application because early identification of high-risk individuals can support preventive interventions and clinical assessment.

XGBoost can integrate a large number of clinical variables and model nonlinear interactions among them. Its regularization mechanism is particularly valuable when datasets contain multiple predictors and the risk of overfitting is high.

A recent study on coronary artery disease model development reported strong performance for XGBoost, including favorable discrimination in external and replication validation cohorts. This finding is important because external validation provides stronger evidence than internal validation alone.

However, very high AUC values should be interpreted carefully. Model performance can be influenced by sample selection, predictor availability, outcome definition, class balance, and possible data leakage. Therefore, models reporting excellent internal performance should ideally be tested in independent populations before clinical deployment.

VII. XGBOOST AND GRADIENT BOOSTING FOR STROKE PREDICTION

Stroke prediction presents several challenges because stroke is relatively less frequent than non-stroke outcomes in many datasets. This creates class imbalance, where the majority class can dominate model training. Algorithms may achieve apparently high accuracy simply by predicting the majority class.

For this reason, evaluation should include sensitivity, specificity, precision, recall, F1-score, ROC-AUC, and preferably precision-recall metrics. Data balancing approaches such as SMOTE may also be considered.

A 2024 systematic review of machine learning algorithms for stroke prediction found that Random Forest was identified as the most efficient model in some studies, while XGBoost, stacking, support vector machines, and other algorithms were reported as top-performing methods in other studies. This variation demonstrates that no single model consistently dominates every stroke dataset.

A broader 2025 systematic review of stroke prediction research found that ensemble approaches achieved particularly high reported accuracy for stroke occurrence prediction, although considerable heterogeneity among datasets and study designs limited direct comparison across studies.

Recent work using SMOTE and XGBoost reported strong classification performance and used SHAP analysis to identify age and hypertension-related variables among the dominant predictors. The study illustrates the potential benefits of combining imbalance correction, boosting, cross-validation, and explainable AI techniques.

Table 3. Typical Predictors Used in Heart Disease and Stroke Prediction

Predictor Category	Heart Disease Prediction	Stroke Prediction
Demographic factors	Age, sex	Age, sex
Blood pressure	Hypertension, systolic/diastolic BP	Hypertension, blood pressure
Metabolic factors	Diabetes, glucose	Diabetes, glucose
Lipid profile	Cholesterol, triglycerides	Cholesterol and lipid abnormalities
Lifestyle	Smoking, diet, physical activity	Smoking, inactivity
Anthropometric factors	BMI, obesity	BMI, obesity
Medical history	CAD, family history, previous events	Atrial fibrillation, hypertension, previous stroke/TIA
Laboratory data	Glucose, lipid profile	Glucose and biochemical markers
Clinical findings	ECG and chest pain variables	Neurological and cardiovascular variables

VIII. COMPARATIVE PERFORMANCE OF BOOSTING TECHNIQUES

The comparative performance of machine learning algorithms depends on the prediction target and data characteristics. A model that performs best for one population may not perform best for another.

Krittanawong et al. (2020) conducted a meta-analysis involving more than three million individuals across 103 cohorts. Their findings suggested promising predictive performance for machine learning in cardiovascular disease, including pooled AUC values of approximately 0.88 for boosting algorithms in coronary artery disease prediction and 0.91 for stroke prediction. Nevertheless, heterogeneity across studies was substantial.

Recent reviews continue to identify ensemble learning as an important strategy for cardiovascular prediction, but they emphasize that claims of algorithm superiority should be interpreted in the context of dataset composition, preprocessing, validation, and clinical setting.

Table 4. Comparative Strengths of Common Machine Learning Algorithms

Algorithm	Major Strength	Major Limitation	Cardiovascular Prediction Use
Logistic Regression	High interpretability	Limited nonlinear modeling	Baseline risk prediction
Decision Tree	Easy interpretation	High overfitting risk	Clinical decision support
Random Forest	Strong performance and robustness	Large models may be less interpretable	Heart disease and stroke
SVM	Effective in high-dimensional spaces	Parameter sensitive	Cardiovascular classification
Gradient Boosting	Strong nonlinear modeling	Can overfit	Risk prediction
XGBoost	High efficiency and regularization	Requires hyperparameter tuning	Structured clinical datasets
Neural Networks	Captures highly complex patterns	Data intensive and less interpretable	Imaging and large datasets

IX. EVALUATION METRICS

The evaluation of cardiovascular prediction models should not depend on accuracy alone.

IJARSCT

International Journal of Advanced Research in Science, Communication and Technology

International Open-Access, Double- Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

ISSN: 2581-9429 Volume 6, Issue 4, July 2026 Impact Factor: 8.2

1. Accuracy

Accuracy represents the proportion of correctly classified observations:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

However, accuracy may be misleading in highly imbalanced stroke datasets.

2. Sensitivity

Sensitivity measures the ability to correctly identify positive cases:

$$Sensitivity = TP / (TP + FN)$$

In cardiovascular screening, high sensitivity is important because missing a high-risk patient may have serious consequences.

3. Specificity

Specificity measures the ability to correctly identify negative cases:

$$Specificity = TN / (TN + FP)$$

4. Precision

Precision indicates the proportion of predicted positive cases that are truly positive:

$$Precision = TP / (TP + FP)$$

5. F1-Score

The F1-score balances precision and recall and is useful when class imbalance exists.

6. Area Under the ROC Curve

AUC evaluates the model's ability to distinguish between positive and negative outcomes across different classification thresholds.

7. Calibration

Calibration examines whether predicted probabilities correspond to actual event frequencies. A model with excellent discrimination but poor calibration may be unsuitable for clinical risk estimation.

8. Decision Curve Analysis

Decision curve analysis evaluates the potential clinical net benefit of using a prediction model across different decision thresholds.

Table 5. Recommended Metrics for Cardiovascular Prediction Models

Metric	Main Purpose	Importance
Accuracy	Overall correctness	Limited for imbalanced data
Sensitivity	Identify true disease cases	Very high
Specificity	Identify true non-disease cases	High
Precision	Reliability of positive predictions	High
F1-score	Balance precision and recall	High
ROC-AUC	Overall discrimination	Very high
PR-AUC	Performance with rare outcomes	Very high for imbalanced datasets
Calibration	Accuracy of predicted probabilities	Essential
Decision Curve Analysis	Clinical usefulness	Highly recommended

X. EXPLAIN ABILITY AND CLINICAL INTERPRETATION

One of the major concerns regarding machine learning is the “black box” problem. Clinicians require information about why a model has assigned a particular risk score.

XGBoost provides several approaches to interpretation, including:

1. Feature importance scores
2. Partial dependence plots
3. SHAP values
4. Individual prediction explanations

DOI: 10.48175/568

SHAP analysis is particularly valuable because it can provide both global and patient-level explanations. In recent stroke prediction research using XGBoost, SHAP analysis was used to identify major predictors and provide interpretable feature rankings.

Interpretability is essential for clinical acceptance. A highly accurate model that cannot be understood or audited may face challenges in routine healthcare implementation.

XI. MAJOR ADVANTAGES OF XGBOOST IN CARDIOVASCULAR PREDICTION

XGBoost offers several advantages:

1. **High predictive performance:** It can model nonlinear relationships and complex interactions.
2. **Regularization:** L1 and L2 penalties help reduce overfitting.
3. **Handling of structured data:** It is especially suitable for clinical and electronic health record datasets.
4. **Missing value handling:** It can efficiently manage incomplete data.
5. **Scalability:** XGBoost can process large datasets efficiently.
6. **Feature importance:** Important predictors can be identified.
7. **Flexibility:** It can be applied to classification and regression problems.
8. **Compatibility with explainable AI:** SHAP and related approaches can improve interpretability.

Table 6. Advantages and Limitations of XGBoost

Advantages	Limitations
High predictive power	Requires careful hyperparameter tuning
Captures nonlinear relationships	Can overfit small datasets
Handles complex feature interactions	Interpretation may be difficult without explainability tools
Includes regularization	Performance may vary across populations
Efficient for large datasets	Computational cost may increase with extensive tuning
Handles missing values	Dataset bias can be learned and amplified
Provides feature importance	External validation is often lacking

XII. CHALLENGES IN HEART DISEASE AND STROKE PREDICTION

Despite promising results, several challenges remain.

1. Dataset Imbalance

Stroke events are often rare compared with non-stroke cases. Accuracy alone may therefore provide an unrealistic impression of model quality.

2. Small and Single-Center Datasets

Many published models are developed using limited datasets. Such models may not generalize to other hospitals or populations.

3. Data Leakage

Improper preprocessing, feature selection, or oversampling before data partitioning can lead to information leakage.

4. Lack of External Validation

Many studies rely exclusively on internal validation. Independent external validation is necessary before clinical deployment.

5. Calibration Problems

High AUC does not guarantee accurate risk probabilities.

6. Algorithmic Bias

Models trained on unrepresentative populations may perform differently across demographic groups.

7. Clinical Integration

Integration with electronic health records, workflow design, clinician acceptance, privacy, and regulation remain major barriers.

Recent reviews of heart disease and stroke prediction consistently identify issues related to dataset heterogeneity, population representation, privacy, bias, and limited real-world clinical validation.

XIII. CONCLUSION

XGBoost and gradient boosting techniques have emerged as important machine learning approaches for heart disease and stroke prediction. Their ability to model nonlinear relationships, capture complex interactions, and process structured healthcare data makes them particularly suitable for cardiovascular risk assessment. XGBoost further improves conventional gradient boosting through regularization, computational efficiency, scalable implementation, and effective handling of missing data. Current evidence indicates that boosting algorithms can achieve strong predictive discrimination in cardiovascular applications, and recent studies continue to demonstrate the potential of XGBoost for coronary artery disease, cardiovascular events, and stroke prediction. Nevertheless, reported performance varies substantially across datasets and methodological designs.

The future clinical value of XGBoost will depend not only on improved accuracy but also on robust external validation, calibration, interpretability, fairness, transparency, and successful integration into healthcare workflows. Prediction models should therefore be evaluated as clinical decision-support tools rather than merely as computational classifiers. With rigorous methodological development and validation, XGBoost and related gradient boosting methods have considerable potential to support earlier identification of individuals at high risk of heart disease and stroke.

REFERENCES

1. Asadi, F., et al. (2024). The most efficient machine learning algorithms in stroke prediction: A systematic review. *Health Science Reports, 7*, e70062.
2. Begum, S. S., Swamy, A., Dhanka, S., Abu-Shareha, A. A., Khan, M., & Husain, I. (2026). Generative adversarial networks and hyperparameter-optimized XGBoost for enhanced heart disease prediction. *Scientific Reports, 16*, 11326.
3. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery.
4. Chen, Y., et al. (2024). Construction and validation of a predictive model for coronary artery disease using extreme gradient boosting. *Journal of Clinical Medicine*, 13.
5. Choi, J. H., et al. (2023). XGBoost-based simple three-item model accurately predicts outcomes of acute ischemic stroke. *Journal of Clinical Medicine, 12*.
6. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189–1232.
7. Haseeb, S., Ansari, U., & Ali, H. (2024). Using machine learning extreme gradient boosting model to predict major adverse cardiovascular events: A systematic review. *JACC: Cardiovascular Interventions, 17*(4_Supplement), S55.
8. Krittanawong, C., Virk, H. U. H., Bangalore, S., Wang, Z., Johnson, K. W., Pinotti, R., Zhang, H. J., Kaplin, S., Narasimhan, B., Kitai, T., Baber, U., Halperin, J. L., & Tang, W. H. W. (2020). Machine learning prediction in cardiovascular diseases: A meta-analysis. *Scientific Reports, 10*, 16057.
9. Kumar, R. K. R., Garg, S. G., Kaur, R., Johar, M. G. M., Singh, S., Menon, S. V., & Kumar, P. (2025). A comprehensive review of machine learning for heart disease prediction: Challenges, trends, ethical considerations, and future directions. *Frontiers in Artificial Intelligence, 8*, 1583459.
10. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems, 30*.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double- Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, July 2026
Impact Factor: 8.2

11. Lundberg, S. M., Erion, G. G., Chen, H., DeGrave, A. J., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*, 56–67.
12. Omotehinwa, T. O., et al. (2024). Machine learning techniques for cardiovascular disease prediction using healthcare data. *Frontiers in Artificial Intelligence*.
13. Patel, J., et al. (2023). Machine-learning algorithms for ischemic heart disease prediction: A systematic review. *Journal of Clinical Medicine, 12*.
14. Shinde, P., Sanghavi, M., & Tran, T. A. (2025). A survey on machine learning techniques for heart disease prediction. *SN Computer Science, 6*, 334.
15. Srinivas, K., & Katarya, R. (2022). Machine learning approaches for cardiovascular disease prediction and healthcare analytics. *International Journal of Health Information Systems and Informatics*.
16. Teja, R., & Rayalu, S. (2025). HistGradientBoosting and advanced machine learning approaches for structured healthcare prediction. *International Journal of Artificial Intelligence and Data Science*.
17. Tiwari, A., et al. (2022). Comparative evaluation of ensemble machine learning techniques for clinical prediction. *Healthcare Analytics, 2*, 100071.
18. Wu, Y., et al. (2025). Machine learning techniques for stroke prediction: A systematic review of algorithms, datasets, and regional gaps. *International Journal of Medical Informatics, 203*, 106041.
19. Zhang, Y., et al. (2025). Stroke prediction using synthetic minority over-sampling technique and extreme gradient boosting. *Computer Methods in Biomechanics and Biomedical Engineering*.
20. Zhao, X., et al. (2025). Prediction of cardiovascular disease based on multiple feature selection and improved PSO-XGBoost model. *Scientific Reports, 15*.