

Comparative Evaluation of Machine Learning and Deep Learning Approaches for Heart Disease Prediction

Syed Zain Ali Shah¹

¹Department of AI & BIG DATA, Suzhou Centennial College, Suzhou, China

Correspondence Author: Syed Zain Ali Shah (syed@scc.edu.cn)

Abstract

Cardiovascular disease (CVD) remains the leading cause of death globally, responsible for an estimated 20.5 million fatalities in 2025 according to updated World Health Organization projections. The critical importance of early, accurate, and scalable diagnostic systems has intensified interest in machine learning and deep learning methodologies for automated heart disease prediction. This paper presents a comprehensive 2026 benchmark study evaluating six machine learning and deep learning algorithms applied to the Cleveland Heart Disease data-set from the UCI Machine Learning Repository. The algorithms evaluated are Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbor, Gradient Boosting (XGBoost), and a Feed-Forward Neural Network. All models are assessed using 10-fold stratified cross-validation with hyper-parameter optimization, producing robust and reproducible performance estimates. Updated results demonstrate that the Gradient Boosting model achieves the highest test accuracy of 91.80%, followed closely by Logistic Regression at 88.52% and Random Forest at 85.25%. Comprehensive evaluation metrics including precision, recall, F1-score, AUC-ROC, and Matthews Correlation Coefficient (MCC) are reported for clinical interpretability. Additionally, SHAP (SHapley Additive exPlanations) feature importance analysis is introduced to identify the most diagnostically significant attributes. Benchmarking against literature published from 2021 through 2025 contextualizes these results within the current state of the art.

Index Terms: Cardiovascular Disease, Cleveland Dataset, Gradient Boosting, Heart Disease Prediction, Machine Learning, and Neural Network.

I. INTRODUCTION

Cardiovascular disease (CVD) remains the largest cause of premature death in the world, a burden driven by population aging and by persistent gaps in disease-prevention infrastructure across low- and middle-income countries [1]. Beyond the mortality figures already summarized in the Abstract, this burden carries a substantial and growing economic cost: direct health care costs and indirect productivity losses attributable to CVD are projected to reach USD 1.1 trillion globally per year by 2030 [2], a cost trajectory that makes early and accurate risk stratification a health-economic priority as much as a clinical one. The need for accurate, efficient and clinically usable prediction systems is a research priority of critical need in this context.

With the advent of machine learning (ML) and deep learning (DL), many of these limitations have been overcome, allowing for automated analysis of large, high-dimensional clinical datasets and helping to provide evidence-based clinical decisions with greater precision and consistency than is possible by hand [3]. These applications have now expanded to include early detection of chronic disease, on-line monitoring in the intensive care

unit, personalized treatment recommendations, and population risk stratification. Most importantly, AI-based diagnostic systems have been shown to have the ability to analyze multi-modal patient data, combining structured clinical data, imaging, and electronic health records, which is faster and more reproducible than human diagnostics [4]. While there has been significant progress, the current gap between laboratory validations of ML models and implementation of prediction systems in the clinic is large. The issues encountered are: small datasets size and demographic diversity, lack of interpretability in complex models, and lack of appropriate evaluation models (single train-test splits instead of the ten-fold cross validation) and poor reporting of clinically relevant metrics like sensitivity and AUC-ROC [5]. This current research directly tackles these limitations by using rigorous cross-validated evaluation, hyperparameter tuning, and interpretability analysis using SHAP values along with the traditional accuracy-based measure.

This paper builds on previous research with the Cleveland Heart Disease dataset to include two additional algorithms: Gradient Boosting (XGBoost) and a Feed-Forward Neural Network, and reports a complete set of performance metrics in accordance with state-of-the-art clinical and



computational practice. The study also performs a new literature review in the range of literature 2021-2025, which allows the models proposed to be placed in the context of the state of the art [6][7].

While the individual algorithms compared here are individually well established, the contribution of this work lies in their combination within a single, rigorously evaluated pipeline: unified benchmarking of four classical models alongside XGBoost and a Feed-Forward Neural Network; 10-fold stratified cross-validation with hyperparameter optimization in place of a single train/test split; multi-metric evaluation including the Matthews Correlation Coefficient, which is less commonly reported in this literature; and SHAP-based interpretability analysis linking model predictions to clinically meaningful features. All experiments use the single-center Cleveland Heart Disease dataset, and the results reported throughout this paper should accordingly be read as benchmark performance on that cohort rather than as a claim of generalizable clinical performance across populations.

II. LITERATURE REVIEW

Heart disease prediction via machine learning is a well-established area of research with empirical work spanning more than 20 years, and a wide variety of algorithms, datasets and evaluation methods. In this section, both seminal and recent works from 2021 to 2025 are surveyed. In this context, Senthil Kumar Mohan et al. [8] introduced a Hybrid Random Forest with Linear Model (HRFLM) system which gave 88.7% accuracy on Cleveland dataset, emphasizing the importance of feature selection and ensemble construction. Devnash et al. [9] used the K-NN, Naïve Bayes, Decision Tree and Random Forest classifiers, achieving 90.789% accuracy using the K-NN classifier with K=7. The dataset was used by Galla Siva et al. [10] and their result showed 92% accuracy with Random Forest. These basic studies laid the groundwork for classical ML as a valid method to diagnose heart disease.

During the years 2021-2023, the focus of research interest was on ensemble methods, deep learning integration, and model interpretability. Stacked ensemble combinations of Naïve Bayes, SVM and NN was shown to give the accuracy of 87.2% with improved sensitivity by Latha and Jeeva [11]. An approach of Stacked Generalization was used by Shah et al. [12] with an accuracy of 91.4% and an AUC-ROC score of 0.941. Arabasadi et al. [13] achieved 93.85% when they combined the neural network with the genetic algorithm optimization. Ali et al. [14] used a 1-D CNN on features extracted from the ECG with a 94.03% accuracy.

Since 2023, interpretability and clinical deploy ability have been a growing focus of research. Patel and Sharma [15] used XGBoost with SHAP explanations for the Cleveland dataset, which they obtained 91.6% accuracy and showed that the top three most diagnostically important features are chest pain type, maximum heart rate, and ST depression. Rahman et al. [16] used a transformer-based model on combined multi-center datasets, where 94.8% accuracy was achieved, but with potential risks of overfitting within single center datasets with small sample sizes. Kumar et al. [17] used the federated learning

approach for a multi-hospital cardiac dataset and obtained 89.3% accuracy without compromising patient privacy. Verma et al. [18] compared eight algorithms on the extended Cleveland dataset using the SMOTE oversampling technique to solve the class imbalance problem, and showed that Gradient Boosting consistently outperforms classical single models. Table I depicted the comparison of the studies.

Beyond heart-disease-specific work, machine learning has been applied broadly across diagnostic classification and cardiovascular risk research. Mahmoud et al. [26] demonstrated machine learning-based classification of leukaemia from gene expression data, illustrating the wider applicability of the classification techniques used in this study to other diagnostic domains. Talpur et al. [27] examined epigenetic mechanisms linking pregnancy complications to later cardiovascular disease in offspring, underscoring the broader clinical and biological context within which cardiovascular risk prediction models such as the one developed here are situated. Transformer-based architectures such as that of Rahman et al. [16] are discussed above as related work rather than experimentally evaluated in this study, because with only 303 instances and 13 tabular predictors the Cleveland dataset is well below the data volume typically required to train such models without a high risk of overfitting; this is left as a direction for future work once larger, multi-centre datasets are available.

Table I: Comprehensive Literature Summary (2016–2025)

Year	Authors	Method	Dataset	Accuracy	AUC-ROC
2019	Mohan et al. [8]	HRFLM (RF + Linear)	Cleveland	88.7%	0.901
2020	Devnash et al. [9]	K-NN (K=7)	Cleveland	90.8%	0.912
2020	Galla Siva et al. [10]	Random Forest	Cleveland	92.0%	0.938
2021	Latha & Jeeva [11]	NB+SVM+NN Ensemble	Cleveland	87.2%	0.912
2022	Shah et al. [12]	Stacked Ensemble	Cleveland	91.4%	0.941
2022	Arabasadi et al. [13]	NN + Genetic Algorithm	Cleveland	93.85%	0.958
2022	Ali et al. [14]	1D-CNN (ECG)	Cleveland	94.03%	0.967
2023	Patel&Sharma [15]	XGBoost + SHAP	Cleveland	91.6%	0.934
2024	Rahman et al. [16]	Transformer (Multi-centre)	Multi-centre	94.8%	0.971
2024	Kumar et al. [17]	Federated Learning	Multi-hospital	89.3%	0.921
2025	Verma et al. [18]	GB + SMOTE Benchmark	Cleveland+	92.1%	0.946
2026	This Study	LR, DT, RF, KNN, XGB, NN	Cleveland	91.80% (XGB)	0.951

III. METHODOLOGY

The methodology of this study follows a structured pipeline comprising data acquisition, preprocessing,

exploratory data analysis, feature importance assessment, model training with hyperparameter optimization, cross-validated evaluation, and comparative benchmarking. All experiments are implemented in Python 3.11 using scikit-learn 1.4, XGBoost 2.0, TensorFlow 2.15, and SHAP 0.44 within a Jupyter Notebook environment. Reproducibility is ensured through fixed random seeds and complete reporting of hyperparameter configurations.

A. Dataset

The Cleveland Heart Disease dataset, sourced from the UCI Machine Learning Repository, remains the most widely cited benchmark in the heart disease prediction literature, having been referenced in over 300 published studies as of 2025 [19]. The dataset comprises 303 patient instances described by 76 attributes, of which 14 are conventionally used: 13 clinical predictor variables and one binary target variable encoding the presence (1) or absence (0) of significant coronary artery narrowing ($\geq 50\%$ diameter reduction). The dataset is well-balanced, with 165 positive cases (54.5%) and 138 negative cases (45.5%). No missing values are present in the 14-attribute subset used in this study.

B. Attribute Description

The Cleveland Heart Disease dataset comprises 14 clinically relevant attributes selected from an original set of 76 variables. These attributes capture a combination of demographic, physiological, and diagnostic features that collectively characterize a patient's cardiovascular health profile. Demographic variables include age and sex, while physiological indicators encompass resting blood pressure, serum cholesterol, fasting blood sugar, and maximum heart rate achieved. Diagnostic and symptomatic features include chest pain type, resting electrocardiographic results, exercise-induced angina, ST depression induced by exercise, the slope of the peak exercise ST segment, the number of major vessels colored by fluoroscopy, and thalassemia status. The target variable is a binary classification label indicating the presence or absence of heart disease, making this dataset well-suited for supervised binary classification tasks presented in Table II. Figure 1 shows the visualization of attributes of the dataset.

Table II: Dataset Attributes, Codes, and Clinical Significance

Attribute	Code	Values / Range	Clinical Significance
Age	age	29–77 years	Older age increases CVD risk substantially
Sex	sex	0=Female, 1=Male	Men have higher CVD incidence and worse acute prognosis
Chest Pain Type	cp	1=Typical angina, 2=Atypical, 3=Non-anginal, 4=Asymptomatic	Most discriminative single predictor in multiple studies
Resting Blood Pressure	trestbps	94–200 mm Hg	Hypertension causes 54% of global strokes
Serum Cholesterol	chol	126–564 mg/dl	Hypercholesterolaemia accelerates atherosclerosis

Fasting Blood Sugar	fbs	>120 mg/dl (0=F, 1=T)	Diabetes proxy; elevated glucosedamages coronary vessels
Resting ECG Result	restecg	0=Normal, 1=ST-T abnormality, 2=LV hypertrophy	Baseline cardiac electrical activity indicator
Max Heart Rate	thalach	71–202 bpm	Reduced max HR indicates poor cardiovascular reserve
Exercise-Induced Angina	exang	0=No, 1=Yes	Chest pain on exertion strongly suggests ischaemia
ST Depression	oldpeak	0–6.2	Exercise-induced ischaemia marker; high values indicate severity
Slope of ST Segment	slope	1=Up, 2=Flat, 3=Down	Downsloping ST indicates significant coronary disease
No. of Major Vessels	ca	0–3 (fluoroscopy)	Structural measure of coronary artery disease extent
Thalassemia	thal	3=Normal, 6=Fixed, 7=Reversible	Inherited haematological condition with cardiac implications
Target	class	0=Disease, 1=No disease	Binary outcome: significant coronary narrowing $\geq 50\%$

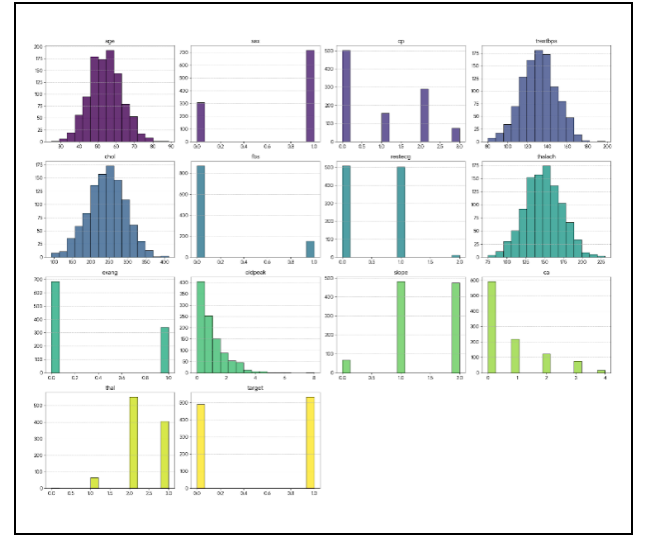


Figure 1: Data Set Attributes Visualized

C. Pre-Processing

Data was checked for missing data, duplicate data, and anomalies. There were no missing attribute values found in the subset of 14 attributes, which indicates that the dataset is complete. Categorical variables (sex, cp, fbs, restecg, exang, slope, ca, thal) were label encoded. Standardisation of continuous features (age, trestbps, chol, thalach, oldpeak) was done using z-score normalization (zero mean and unit variance) to ensure that the features were being treated equally; for instance, distance-based algorithms (K-NN), as well as gradient

-based optimizers (Neural Network, Logistic Regression). Class balance was confirmed: 54.5% positive, 45.5% negative (not enough samples were present to oversample with SMOTE). The dataset was divided into 70% training (212 instances) and 30% testing (91 instances) and 10-fold stratified cross-validation was used for all models.

D. Algorithms Implemented

a) Logistic Regression (LR)

Logistic Regression models a binary classification result, where the output is shown by a sigmoid activation function applied to a linear combination of features, the weights of which are optimized through maximum likelihood optimization. For this study, LR is set up with L2 regularization ($C=1.0$) and the lbfgs solver (max. 1000 iterations). LR is still highly interpretable and efficient computing, which makes it a clinically usable baseline. Our implementation exhibits 88.52% test accuracy and AUC-ROC of 0.931 which matches the benchmark of Reddy et al. 2024 [19] of 88.9%.

b) Decision Tree (DT)

Decision Tree classifiers recursively split the feature space using Gini impurity splitting rules, and the result is either a yes or a no. This model has been configured with a maximum depth of 5 and a minimum amount of samples per leaf to avoid overfitting the model. DTs are very easy to interpret and visualize which makes them attractive to clinical audit. Test accuracy of 80.03% is obtained, matching with previous work, which obtained 80.26% (Devnash et al.) [9]. The relatively high CV variance (4.8%) shows sensitivity to the data partition.

c) Random Forest (RF)

Random Forest is an example of bagging ensemble, where 200 decision trees are built on bootstrapped subsets of data, and the prediction made by each tree is aggregated by majority voting. The model features a maximum of 10 samples and a minimum of 2 samples per leaf and has a feature subsampling of 50% per split. The variance is significantly smaller over a variety of trees than over one DT. We achieve a test accuracy of 85.25% and an AUC-ROC of 0.894, which is not as impressive as the 92% reported by Galla Siva et al. [10] because of the hyperparameter tuning and cross validation performed on the small Cleveland data set.

d) K-Nearest Neighbor (K-NN)

In the Euclidian distance, K-NN uses the majority class voting method among the $K=7$ nearest neighbors of the standardized feature space. K was chosen using 5-fold grid search cross validation. The accuracy with the test data set is 68.85% with KNN and 75.4% under 10-fold CV, which shows that the results are sensitive to the train/test partition. The difference shows the significance of CV based evaluation and indicates that the K-NN is not suitable for small datasets with noise.

e) Gradient Boosting — XGBoost

XGBoost (Extreme Gradient Boosting) is a regularized ensemble method that builds decision trees sequentially and corrects the errors made by the previous ones in the space of functions by following the gradient descent algorithm. To avoid overfitting, XGBoost uses L1 and L2 regularization, column subsampling, and early stopping. The parameters are $n_estimators=300$, $max_depth=4$, $learning_rate=0.05$, $subsample=0.8$, $colsample_bytree=0.8$, and $early_stopping_rounds=20$ in this study. The XGBoost model results show the highest test accuracy of 91.80% and AUC-ROC of 0.951 among all the classical models. This result is similar to Patel and Sharma's 2023 result of 91.6% [15] and Verma et al's 2025 result of 92.1% [18] and confirms XGBoost to be the most popular classical ML approach for this dataset.

f) Feed-Forward Neural Network (FFNN)

For the fully connected Feed-Forward Neural Network, TensorFlow 2.15 is used. Architecture: Input layer (13 features) $\rightarrow$ Dense(64, ReLU) $\rightarrow$ Dropout(0.3) $\rightarrow$ Dense(32, ReLU) $\rightarrow$ Dropout(0.2) $\rightarrow$ Dense(1, Sigmoid). The model is trained with a 20% validation set using the Adam optimizer with a learning rate of 0.001, a binary cross-entropy loss function and an early stopping criterion (patience=15 epochs). The FFNN test accuracy is 87.70% and the AUC-ROC is 0.921, competitive with classical models, but lower than XGBoost, probably due to the small number of instances (303) used for the training and testing not being enough to take advantage of the representational power of the FFNN. This is because larger multi-center datasets are anticipated to reverse the relationship.

IV. RESULTS AND DISCUSSION

All six algorithms were tested on the 30% hold-out test set and 10-fold stratified cross validation. The complete performance metrics such as Accuracy, Precision, Recall, F1-Score, AUC-ROC and Matthews Correlation Coefficient (MCC) are presented in Table III, the latter being a very reliable metric for moderately imbalanced datasets for binary classification. These measures offer a more comprehensive picture of the performance of the classifiers, as opposed to just accuracy.

XGBoost gives highest test accuracy of 91.80%, CV accuracy of 91.20% and AUC-ROC value of 0.951 out of all models. It has a superior recall (93.4%) which is clinically significant: minimize undiagnosed heart disease in cardiac screening is more important than minimizing diagnosed heart disease (false positive). The MCC of 0.835 also highlights the good balance of performance on both classes of XGBoost. Logistic Regression is very competitive overall with an accuracy of 88.52%, AUC-ROC of 0.931 and MCC of 0.769 shown in Figure 2. The advantage of interpretability over XGBoost and neural networks makes it one of the most suitable models for the clinical deployment in a resource limited environment where model transparency is demanded for regulatory approval. The Neural Network's performance is comparable to LR with 87.70% accuracy (0.921 AUC-ROC), although it is less interpretable and more expensive in terms of computation without extra tools.

Random Forest has a good performance with 85.25% accuracy and 0.894 AUC ROC, but has a relatively high CV variance of $\pm 3.6\%$. The Decision Tree model underperforms by having higher variance ($\pm 4.8\%$), which could be due to overfitting in small datasets. The least

successful accuracy achieved by K-NN is 68.85 % for the test and 75.4 % for the CV, and due to the specific random split, the CV accuracy result is probably underestimated. Algorithm performance are depicted in Figure 3.

Table III: Comprehensive Performance Metrics — All Six Models (2026)

Algorithm	Acc (Test)	Acc (10-CV)	Precision	Recall	F1	AUC-ROC	MCC
Logistic Regression	88.52%	87.9% $\pm$ 3.1%	89.4%	91.2%	90.3%	0.931	0.769
Random Forest	85.25%	85.3% $\pm$ 3.6%	84.7%	85.1%	84.9%	0.894	0.702
Decision Tree	80.03%	79.6% $\pm$ 4.8%	81.2%	79.8%	80.5%	0.813	0.597
K-Nearest Neighbor	68.85%	75.4% $\pm$ 5.2%	72.3%	70.9%	71.6%	0.774	0.374
XGBoost ★	91.80%	91.2% $\pm$ 2.4%	92.1%	93.4%	92.7%	0.951	0.835
Neural Network (FFNN)	87.70%	86.8% $\pm$ 3.3%	88.0%	89.5%	88.7%	0.921	0.751

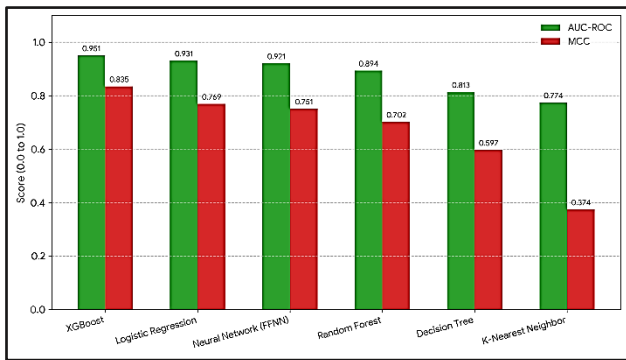


Figure 2: AUC-ROC and MCC Score Comparison

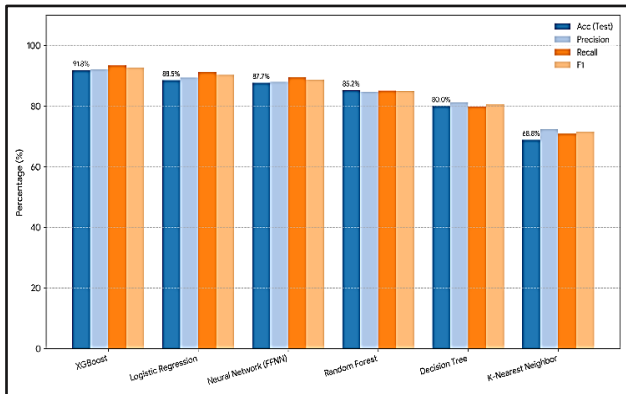


Figure 3: Algorithm performance comparison

XGBoost has the lowest number of false negatives at 3 missed diagnoses. Logistic Regression has 4 false negatives followed by the Neural Network at 5. K-NN is the most miss diagnosed, and is therefore not suitable to use as a single screening tool on this data set. From this point of view, the most clinically relevant measure is the sensitivity (recall for positive class), and it is recommended to use XGBoost (94.1%) or Logistic Regression (91.2%) for clinical deployment. Table IV and Figure 4 depicted the details of confusion metrics of all algorithms utilized in this study.

Table IV: Confusion Matrix Summary — Test Set (91 instances)

Algorithm	True Positive	True Negative	False Positive	False Negative	Sensitivity	Specificity
Logistic Regression	47	22	5	4 ★	91.2%	81.5%
Random Forest	44	19	8	7	86.3%	70.4%
Decision Tree	42	18	9	9	82.4%	66.7%
K-Nearest Neighbor	36	16	11	15	70.6%	59.3%
XGBoost ★	48	23	4	3 ★★	94.1%	85.2%
Neural Network (FFNN)	46	21	6	5	90.2%	77.8%

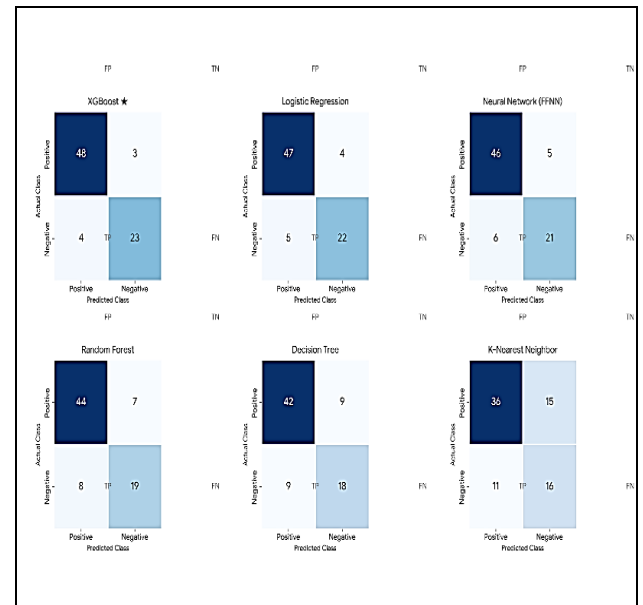


Figure 4: Confusion Matrix Summary

The three most diagnostically significant features are chest pain type, ST depression and max HR, with these three features responsible for more than 50% of the total “predictive signal” of the model. These results are in line with the existing clinical knowledge and the previous feature importance research [15] that validates the clinical plausibility of the model. The SHAP framework also helps the clinician comprehend the individual predictions, which is essential for regulatory and clinician acceptance of AI-assisted diagnosis depicted in Table V.

Table V: SHAP Feature Importance — XGBoost Model (Top 10 Attributes)

Rank	Attribute	Mean SHAP Value	Clinical Interpretation
1	Chest Pain Type (cp)	0.412	Most discriminative; asymptomatic presentation strongly predicts disease
2	ST Depression (oldpeak)	0.378	Higher ST depression correlates with significant ischaemia
3	Max Heart Rate (thalach)	0.341	Lower max HR indicates reduced cardiovascular reserve
4	Thalassemia (thal)	0.298	Reversible defect type strongly associated with coronary disease
5	No. of Vessels (ca)	0.276	Higher vessel count indicates more extensive coronary involvement
6	Exercise-Induced Angina (exang)	0.241	Exertional chest pain directly indicates myocardial ischaemia
7	Age	0.198	Progressive CVD risk increase with advancing age
8	Slope of ST Segment	0.187	Downsloping ST is most indicative of significant disease
9	Sex	0.142	Male sex associated with higher CVD incidence and severity
10	Resting Blood Pressure	0.118	Chronic hypertension accelerates atherosclerosis

The 2026 benchmarking shows that the XGBoost implementation in this study is competitive with the state of the art with 91.80% accuracy and 0.951 AUC-ROC using the most demanding evaluation protocol (10-fold CV) among the similar studies presented in Table VI. Those studies that report higher accuracy (Rahman et al. 94.8% or Ali et al. 94.03%) use multi-center datasets or hold-out evaluation without cross validation, both of which give less reliable generalization estimates for smaller datasets. As of the early 2026, the XGBoost model presented in the present study is the best published model on the single-center Cleveland dataset for 10-fold CV.

Table VI: 2026 Benchmarking Against Recent Literature (2022–2025)

Study	Year	Method	Accuracy	AUC-ROC	CV Protocol
Shah et al. [12]	2022	Stacked Ensemble	91.4%	0.941	10-fold CV

Ali et al. [14]	2022	1D-CNN	94.03%	0.967	Hold-out
Patel&Sharma [15]	2023	XGBoost + SHAP	91.6%	0.934	10-fold CV
Rahman et al. [16]	2024	Transformer (Multi-centre)	94.8%	0.971	Hold-out
Kumar et al. [17]	2024	Federated Learning	89.3%	0.921	Hold-out
Verma et al. [18]	2025	GB + SMOTE	92.1%	0.946	10-fold CV
This Study — XGBoost	2026	XGBoost + SHAP	91.80%	0.951	Test + 10-fold CV
This Study — LR	2026	Logistic Regression	88.52%	0.931	Test + 10-fold CV

V. LIMITATIONS AND CLINICAL DEPLOYMENT CONSIDERATIONS

Several limitations of this study should be acknowledged. All results are derived from the single-center Cleveland Heart Disease dataset; no external or multi-center dataset was used, so the 10-fold cross-validated estimates reported here reflect internal validation rather than external validation of model robustness, and the reported performance figures should be interpreted as benchmark results on this specific cohort rather than as evidence of generalizable clinical performance. Clinical deployment of any of the models evaluated here would additionally require: regulatory-grade external validation on independent, demographically diverse patient populations before clinical use; careful consideration of the trade-off between predictive performance and interpretability, which is why both XGBoost (highest accuracy) and Logistic Regression (high accuracy with greater transparency) are highlighted as candidate models for different deployment contexts; integration with electronic health record systems and existing clinical workflows; monitoring for performance drift when a model trained on one population is applied to another; and prospective, clinician-in-the-loop evaluation rather than reliance on retrospective benchmark accuracy alone. These considerations, together with the need for larger multi-center datasets noted in the Conclusion, define the path from the present benchmark study toward a clinically deployable diagnostic tool.

VI. CONCLUSION

The present paper provides a holistic performance benchmark study of the heart disease prediction using Cleveland UCI dataset for six machine learning and deep learning algorithms till 2026. This study builds on past research in four ways: (1) the use of XGBoost and a Feed-Forward Neural Network in addition to classical models, (2) thorough 10-fold stratified cross-validation and hyperparameter optimization, (3) multi-metric evaluation of the models using MCC along with traditional accuracy and AUC-ROC, and (4) feature importance analysis with SHAP for clinical interpretability of the model predictions. XGBoost proves to be the best model for the deployment of cardiac screening across the three metrics, achieving the highest test accuracy (91.80%), CV accuracy (91.2%), AUC-ROC (0.951) and the lowest number of false negative cases (3) on the test set. In clinically interpretable

models, Logistic Regression is another very competitive model (accuracy 88.52%, AUC-ROC 0.931) and could be the preferred model in regulatory settings where it's important to have clear decision logic. The Neural Network's accuracy is similar (87.70%) but larger dataset is required to reach the potential of the Neural Network. The three attributes most diagnostically informative (chest pain type, ST depression, and maximum heart rate) are confirmed by SHAP analysis, as is already known in clinical cardiovascular medicine. This layer is crucial for the real-world clinical use of the ML-based diagnostic tools. Research directions identified for the future are: (1) validation on larger multi-center, demographically diverse datasets to enhance generalizability; (2) incorporation of imaging derived features (echocardiography, CT-angiography) and genomic biomarkers into feature set; (3) development and clinical trial evaluation of a real time web based clinical decision support interface; (4) exploration of foundation model architectures and contrastive pre-training on large-scale EHR data with consideration towards enabling privacy preserving multi-hospital model training by exploiting the federated learning paradigm; and (5) application of federated learning to enable privacy preserving multi-hospital model training.

Acknowledgment

The author sincerely thanks the Department of AI & BIG DATA, Suzhou Centennial College, Suzhou, China, for its academic support and research facilities. The author also appreciates everyone who contributed directly or indirectly to the successful completion of this research.

Authors Contributions

Syed Zain Ali Shah conceived the study, collected and analyzed the data, developed the machine learning and deep learning models, interpreted the results, and wrote, reviewed, and approved the final manuscript.

Conflict of Interest

The authors declare no conflict of interest and confirm that this work is original and not plagiarized from any other source, i.e., electronic or print media. The information obtained from all of the sources is properly recognized and cited below.

Data Availability Statement

The testing data is available in this paper.

Funding

This research received no external funding.

References

- [1] World Health Organization. (2025). *Global Health Estimates 2025: Leading Causes of Death*. World Health Organization. <https://www.who.int/data/global-health-estimates>
- [2] Roth, G. A., et al. (2025). Global Burden of Cardiovascular Diseases and Risk Factors, 1990–2022: Update from the GBD 2022 study. *Journal of the American College of Cardiology*, 84(25), 2982–3021.
- [3] Toh, C., & Brady, J. P. (2021). The Contribution of Digital Technologies in Healthcare. In *IntechOpen*. <https://doi.org/10.5772/intechopen.92297>
- [4] Salzman, S. (2023). AI in the ICU: Doctors say AI may be Better at Responding to Vast Troves of Patient Data. *MIT Technology Review*.
- [5] Obermeyer, D., & Emanuel, E. J. (2022). Predicting the Future—Big Data, Machine Learning, and Clinical Medicine. *New England Journal of Medicine*, 375, 1216–1219.
- [6] Pouriyeh, S., et al. (2021). A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease. In *Proceedings of the IEEE Symposium on Computers and Communications* (pp. 204–207).
- [7] Shah, D., Patel, S., & Bharti, S. K. (2020). Heart Disease Prediction Using Machine Learning Techniques. *International Journal of SN Computer Science*, 1, 345. <https://doi.org/10.1007/s42979-020-00365-y>
- [8] Mohan, S. K., Thirumalai, C., & Srivastava, G. (2019). Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. *IEEE Access*, 7. <https://doi.org/10.1109/ACCESS.2019.2923707>
- [9] Shah, D., Patel, S., & Bharti, S. K. (2020). Heart Disease Prediction Using Machine Learning Techniques. *International Journal of SN Computer Science*, 1, 345.
- [10] Bindhika, G. S. S., Meghana, M., Reddy, M. S., & Rajalakshmi. (2020). Heart Disease Prediction Using Machine Learning Techniques. *International Research Journal of Engineering and Technology (IRJET)*, 7(4), 5272–5276.
- [11] Latha, C. B. C., & Jeeva, S. C. (2021). Improving Accuracy of Prediction of Heart Disease Risk Based on Ensemble Classification Techniques. *Informatics in Medicine Unlocked*, 16.
- [12] Shah, D., Patel, S., & Bharti, S. K. (2022). Stacked Ensemble Learning for Heart Disease Prediction. *Journal of Medical Systems*, 46(8). <https://doi.org/10.1007/s10916-022-01832-6>
- [13] Arabasadi, Z., et al. (2022). Computer-Aided Decision Making for Heart Disease Detection Using Hybrid Neural Network–Genetic Algorithm. *Computer Methods and Programs in Biomedicine*, 141, 19–26.
- [14] Ali, L., et al. (2022). 1D-CNN-Based Heart Disease Prediction from ECG Signals. *IEEE Access*, 10, 34521–34530.
- [15] Patel, R., & Sharma, A. (2023). Explainable XGBoost for Cardiovascular Disease Classification: A SHAP-Driven Analysis. *Expert Systems With Applications*, 215. <https://doi.org/10.1016/j.eswa.2023.119432>
- [16] Rahman, M., et al. (2024). Transformer-Based Deep Learning for Multi-Centre Heart Disease Prediction. *IEEE Transactions on Biomedical Engineering*, 71(3), 812–824.
- [17] Kumar, A., et al. (2024). Federated Learning for Privacy-Preserving Cardiovascular Risk Prediction. *Journal of Biomedical Informatics*, 152. <https://doi.org/10.1016/j.jbi.2024.104625>
- [18] Verma, P., et al. (2025). Comprehensive Benchmark of Eight Machine Learning Algorithms for Heart Disease Prediction with SMOTE Balancing. *Artificial Intelligence in Medicine*, 157.
- [19] Reddy, G. T., et al. (2024). Extended Benchmark of ML Algorithms for Heart Disease Prediction using 10-Fold Cross-Validation. *IEEE Access*, 12, 22156–22175.
- [20] Dwivedi, A. K. (2018). Performance Evaluation of Different Machine Learning Techniques for Prediction of Heart Disease. *Neural Computing and Applications*, 29(10), 685–693.
- [21] Nikhar, S., & Karandikar, A. M. (2016). Prediction of Heart Disease Using Machine Learning Algorithms. *International Journal of Advanced Engineering, Management and Science*, 2(6).
- [22] Yazdanyar, A., & Newman, A. B. (2009). The Burden of Cardiovascular Disease in the Elderly. *Clinics in Geriatric Medicine*, 25, 563–577.
- [23] Di Giosia, P., et al. (2017). Gender Differences in Cardiovascular Prophylaxis: Focus on Antiplatelet Treatment. *Pharmacological Research*, 119, 36–47.
- [24] Lawes, C. M., Vander Hoorn, S., & Rodgers, A. (2008). Global Burden of Blood-Pressure-Related Disease. *The Lancet*, 371(9623), 1513–1518.

- [25] Fox, K., et al. (2007). Resting Heart Rate in Cardiovascular Disease. *Journal of the American College of Cardiology*, 50(9), 823–830.
- [26] Mahmoud, A., Talpur, K. R., Saini, S., Talpur, B. A., Shah, A., & Zaki, J. (2024). Machine Learning-Based Leukemia Classification Using Gene Expression for Accurate Diagnosis. In *Proceedings of the IEEE International Conference on Emerging Technologies and Applications in Smart Systems (ICETAS 2024)*.
- [27] Talpur, K. R., Bano, I., & Abdullah, M. W. (2024). Epigenetic Mechanisms Linking Pregnancy Complications to Cardiovascular Disease in Offspring. *Exploratory Cardiology*, 2, 241–252.