2025 6th International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE) | 979-8-3315-3946-7/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICBASE66587.2025.11181459

Heart Disease Prediction using Optimized Machine Learning Models

Sisi Liu
Dept. of Computer Science
Guangdong University of Science and Technology
Dongguan, China
liusisi@gdust.edu.cn

***Abstract*—Heart disease remains the leading cause of mortality worldwide, emphasizing the importance of early prediction, which is often challenging. This study proposes an optimized machine learning framework to evaluate the performance of various algorithms in predicting heart disease. Five algorithms—logistic regression (LR), decision tree (DT), random forest (RF), XGBoost (XGB), and multilayer perceptron (MLP)—were applied to the Cleveland dataset. Hyperparameters were optimized through Bayesian optimization and randomized grid search. Results indicated that RF and XGB achieved the highest accuracy (90.0%) and the largest AUC values (0.960 and 0.950, respectively). Logistic regression and decision trees demonstrated comparable accuracy (89.0%), while multilayer perceptron showed comparatively limited performance (88.0%). This work highlights the potential of machine learning techniques in the early detection of heart disease.**

Keywords—heart disease prediction, machine learning optimization, random forest (RF), XGBoost (XGB)

I. INTRODUCTION

According to reports by the World Health Organization (WHO), cardiovascular diseases caused 17.9 million deaths worldwide in 2019, accounting for 32% of total mortality. Projections indicate that this figure will increase to 23 million by 2030, reinforcing the significance of cardiovascular diseases as a critical public health issue. Early detection and intervention may substantially lower mortality risks and mitigate the societal and economic burdens associated with cardiovascular diseases. Machine learning (ML), a subset of artificial intelligence, has become a vital tool in biomedical research, notably in applications such as medical image analysis [1] and cancer cell detection [2]. ML methods can unveil hidden correlations within physiological signals, such as electrocardiograms (ECGs) and phonocardiograms, which are indiscernible to the human eye, enabling extraction of diagnostically valuable insights. Consequently, the application of ML in cardiovascular disease prediction holds considerable promise for clinical impact.

This study evaluated the effectiveness of machine learning (ML) models in predicting heart disease, utilizing the Cleveland Heart Disease dataset from the UCI Machine Learning Repository. The primary objectives of this study were threefold:

- To develop and validate models capable of predicting patients at risk of cardiovascular events.
- To conduct a comparative analysis of ML algorithms to identify the most effective method.
- To thoroughly evaluate model performance across multiple metrics.

To accomplish this, various ML algorithms were implemented, and hyperparameters optimized via both randomized grid search and Bayesian optimization. The scope of the study encompassed:

- Conducting a correlation analysis between features and disease outcomes.
- Employing five ML classifiers: logistic regression (LR), decision tree (DT), random forest (RF), XGBoost (XGB), and multilayer perceptron (MLP).
- Hyperparameter tuning conducted through randomized grid search and Bayesian optimization.
- Performance assessment based on accuracy, recall, precision, F1-score, and AUC-ROC metrics.

II. RELATED WORK

As a powerful data analytics tool, ML has been extensively applied in healthcare. Recent studies worldwide have employed diverse ML algorithms for cardiac disease classification, thereby enhancing clinical decision-making. Bhatt, Patel, Ghetia and Mazzeo [3] employed four models—RF, DT, MLP, and XGB—on a Kaggle dataset. Their results demonstrated that MLP, when combined with cross-validation, outperformed other algorithms, achieving 87.28% accuracy. Chandrasekhar and Peddakrishna [4] implemented six ML algorithms—RF, KNN, LR, Naïve Bayes (NB), Gradient Boosting (GB), and AdaBoost—using GridSearchCV with 5-fold cross-validation. A soft-voting ensemble classifier further enhanced performance on the IEEE Dataport dataset. Bhowmik et al. [5] trained three models—LR, RF, and SVM—on the Cleveland dataset from the UCI Machine Learning Repository, evaluating metrics such as precision, accuracy, recall, F1-score, and AUC-ROC. Additional studies [6], [7], [8], and [9] have investigated ML approaches for cardiac prediction.

Deep learning (DL), a specialized subset of ML, demonstrates exceptional capabilities in processing large-scale data. Zhou, Dai and Hou [10] conducted a systematic review of 64 articles (2018–2023) on DL-based heart disease prediction. The review categorized algorithms into three types:

This work was supported by the Natural Science Project of Guangdong University of Science and Technology (Grant No. GKY-2022BSQD-44 and GKY-2024KYYBK-14).

- Standard deep learning architectures.
- Extended deep learning (ETDL) frameworks.
- Integrated deep learning approaches combining DL with other techniques.

Ur, Yousef and Adee [11] proposed a transformer model based on self-attention for cardiovascular disease (CVD) risk prediction. When tested on the Cleveland dataset, this model outperformed existing benchmarks, achieving superior accuracy.

Extensive research confirms that ML and DL techniques have produced promising results in predicting heart disease. Researchers have utilized algorithms such as RF, SVM, and LR for classification, employing feature selection to reduce dimensionality and improve model performance..

III. Research Methodology

The methodological workflow is depicted in Fig. 1. The pipeline includes data preprocessing, exploratory data analysis, correlation analysis, machine learning modeling, and model evaluation.

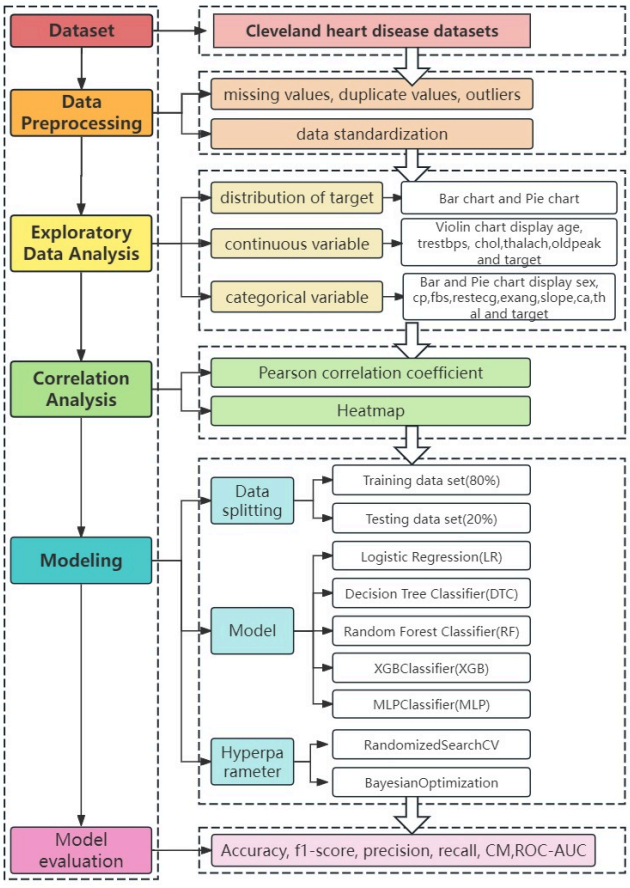

Fig. 1. Workflow diagram of heart disease prediction.

A. Dataset

The dataset was initially assembled through a multi-institutional collaboration involving the Cleveland Clinic Foundation, the Hungarian Institute of Cardiology (Budapest), VA Medical Center, and Long Beach Medical Center. This dataset is a vital resource in cardiology research, aimed at helping researchers and clinicians better understand cardiac risk factors and develop more accurate predictive models. The Cleveland dataset edition remains the most frequently utilized version in academic research. Over time, it has been extensively employed in machine learning and statistical analyses, establishing itself as a standard benchmark for evaluating algorithms in cardiovascular prediction tasks.

B. Data preporcessing

Irrelevant features were eliminated. Categorical variables were label-encoded, while continuous variables (age, trestbps, chol, thalach, oldpeak) exhibited wide value ranges, requiring normalization using z-score as in (1).

$$X\_normalized = X - mean(X) / std(X) \quad (1)$$

C. Exploratory Data Analysis

Exploratory data analysis (EDA) utilized statistical graphics and summary statistics to gain initial insights into data features. This approach helped researchers identify underlying patterns and relationships, establishing the foundation for further analysis and modeling.

EDA focused on two aspects: the distribution of the features and the relationships between features and the target variable. Appropriate statistical visualizations were selected based on the mixed data types (continuous and categorical features).

D. Correlation analysis

The correlation coefficient measures the strength and direction of the linear relationship between two variables, with values from -1 to 1. The magnitude of the coefficient indicates the strength of correlation. Pearson correlation coefficients were calculated to assess the relationships between features.

E. Maching learning techniques

Five ML algorithms—Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), XGBoost (XGB), and Multilayer Perceptron (MLP)—were compared. Hyperparameters were optimized to improve model performance through a combined approach of randomized grid search and Bayesian optimization.

1) ML algorithms: The study compared ML algorithms, including LR, which functions as a transparent linear classifier capable of directly predicting disease probabilities and estimating risk coefficients for features. The decision tree's graphical structure provides intuitive visualization of decision pathways, aiding clinicians in understanding the logic behind high-risk patient identification. Random Forests effectively reduce overfitting via ensemble learning using multiple decision trees, while also automatically assessing feature importance. XGBoost, as the leading gradient boosting algorithm, offers significant advantages in managing class imbalance and capturing complex feature interactions. Multilayer Perceptron (MLP), a fundamental neural network architecture, models complex nonlinear relationships among clinical features. This range of algorithms—from simple to complex, and linear to nonlinear—encompasses mainstream

machine learning paradigms, balancing the needs for accuracy and interpretability in clinical applications.

2) Randomized grid search: Randomized grid search evaluates hyperparameter combinations by randomly sampling from the parameter space. This method is more efficient for large hyperparameter spaces, as it can identify near-optimal solutions with less computational effort. In this study, hyperparameters for DT, RF, XGB, and MLP were optimized using randomized grid search.

3) Bayesian optimization: XGBoost offers numerous hyperparameters, including learning rate, number of trees, tree depth, regularization parameters, and subsample ratio. The hyperparameter tuning process is inherently complex. To address this, a hybrid approach combining both randomized grid search and Bayesian optimization was used for tuning XGBoost hyperparameters.

4) Performance evaluation: Model performance was assessed using metrics such as the confusion matrix, accuracy, precision, recall, F1-score, ROC curve, and AUC. Table I presents the confusion matrix employed for heart disease prediction.

TABLE I. CONFUSION MATRIX

	Predicted = Positive (Class 1)	**Predicted = Negative (Class 0)**
Actual = Positive (Class 1)	TP (True Positive)	FN (False Negative)
Actual = Negative (Class 0)	FP (False Positive)	TN (True Negative)

Accuracy is defined as the proportion of correctly classified samples—both true positives and true negatives—among all samples, calculated as:

$$accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (2)$$

IV. Results and Discussion

This study utilized several ML models to predict heart disease. This section reports the results and compares the performances of the different algorithms.

A. Exploratory data analysis

1) Distribution and analysis of continuous features: The dataset contains five continuous features, including age, resting blood pressure (trestbps), serum cholesterol (chol), maximum heart rate (thalach), and ST segment depression (oldpeak). Distributions of features were visualized via histograms, and the relationships between features and the target label were analyzed using violin plots, as shown in Fig. 2.

Significant clinical distinctions were observed between cardiovascular disease patients and healthy controls, particularly in age, maximum heart rate (thalach), and ST depression (oldpeak) suggest that these factors are key determinants of cardiovascular disease. Conversely, trestbps and cholesterol show less distinct differences, with minimal separation in visual analysis, indicating a weaker association.

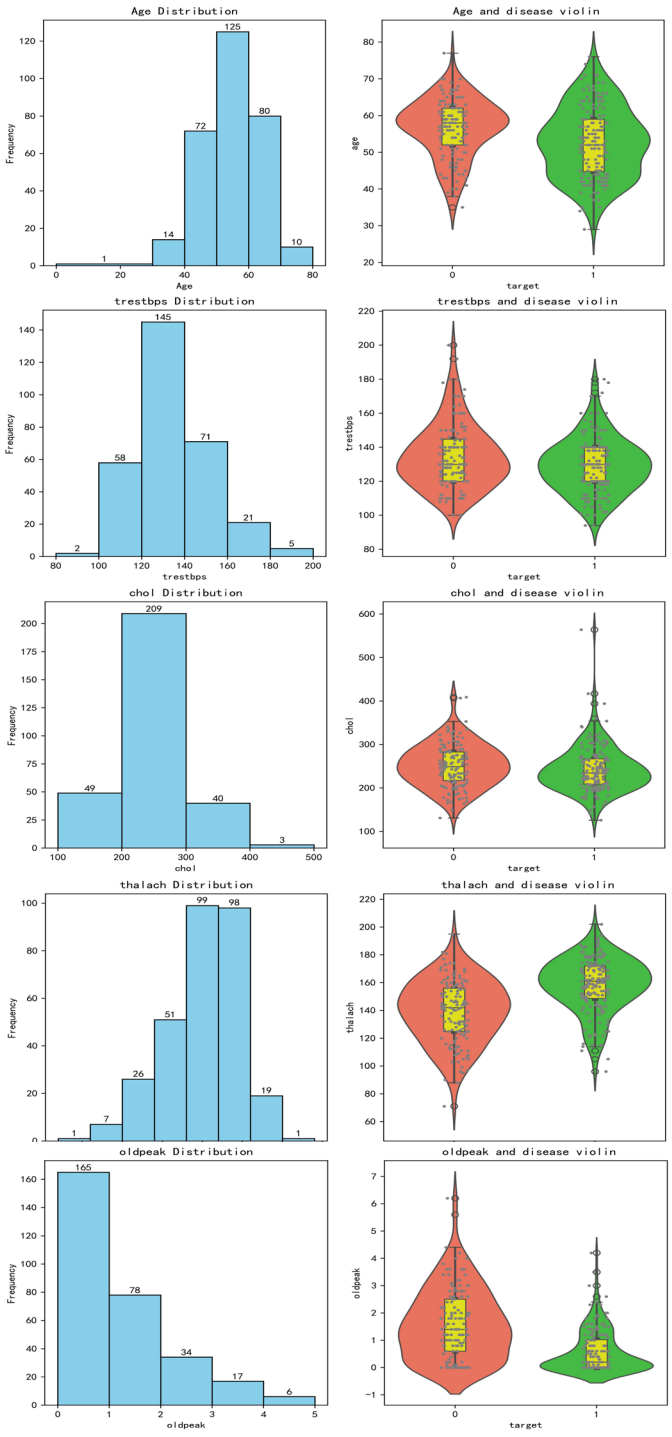

Fig. 2. Continuous variables.

2) Distribution and analysis of categorical features: The categorical features include sex, chest pain type (cp), fasting blood sugar (fbs), resting electrocardiogram results (restecg), exercise-induced angina (exang), slope of the ST segment during peak exercise (slope), number of affected vessels (ca), and thalassemia (thal). Distributions of categorical features were illustrated via pie charts, and their relationships with the outcome variable were analyzed using bar charts, as shown in

Fig. 3. The prevalence rates across categories are summarized in Table II.

Both figures and tables indicate that the prevalence rates for FBS >120 mg/dL and ⩽120 mg/dL are 54.86% and 51.11%, respectively, showing similar distributions. Conversely, the other seven categorical variables demonstrate considerable prevalence differences across categories, suggesting they are influential for cardiovascular risk.

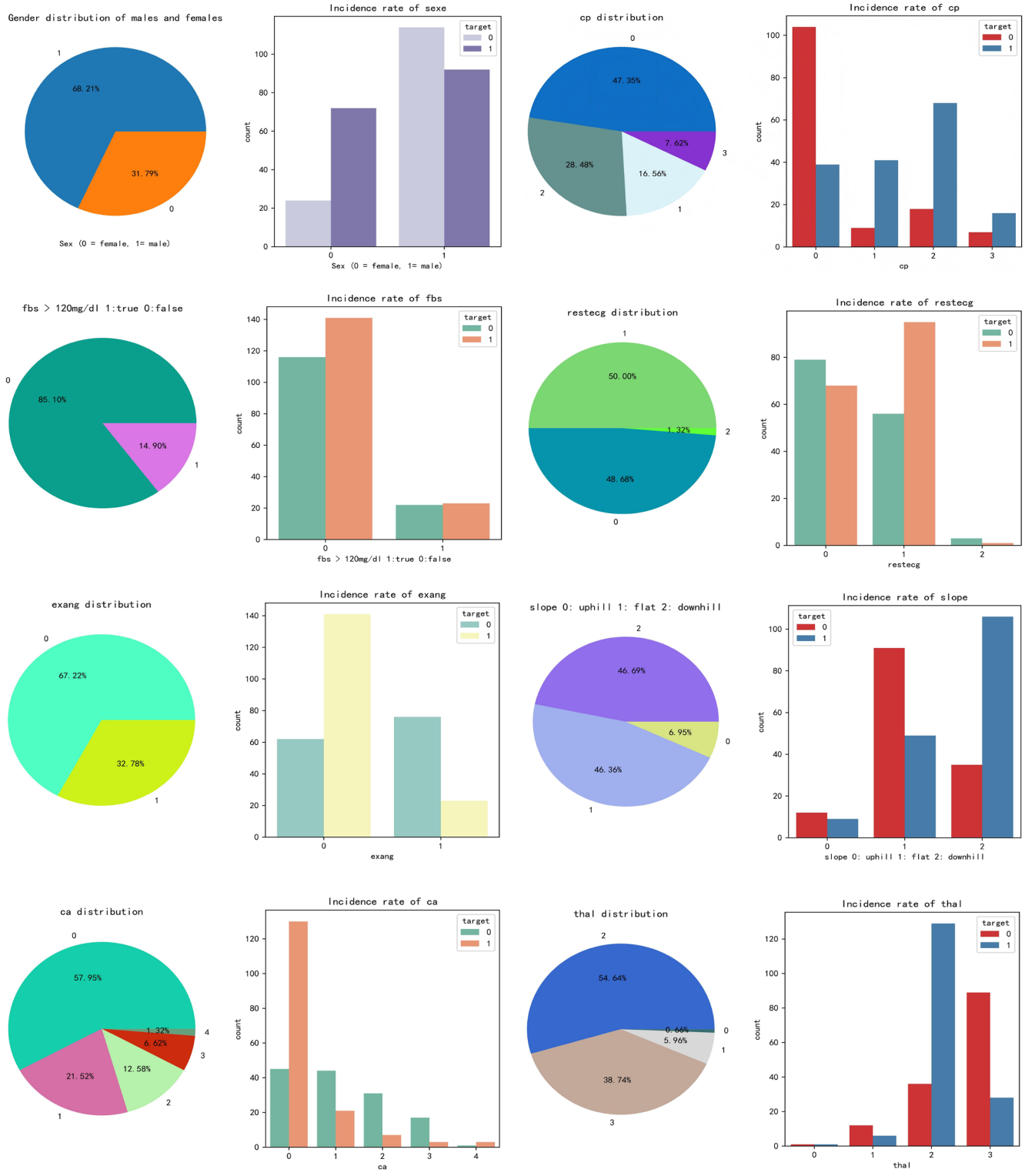

Fig. 3. Categorical variable

TABLE II. PREVALENCE RATES OF CATEGORICAL VARIABLES

Attribute	Description and Rate
sex	1 = male 44.66% 0 = female 75%
cp	0: typical angina 27.27% 1: atypical angina 82% 2: non-anginal pain 79.07% 3: asymptomatic 69.57%
fbs	1 = true 54.86% 0 = false 51.11%
restecg	0: normal 46.26% 1: having ST-T wave abnormality 62.91% 2: showing probable or definite left ventricular hypertrophy by Estes' criteria 25%
exang	1 = yes 69.46% 0 = no 23.23%
slope	0: upsloping 42.86% 1: flat 35% 2: downsloping 75.18%
ca	number of major vessels colored by fluoroscopy 0: 74.29%; 1:32.31%; 2: 18.42%; 3:15%; 4:75%
thal	0 = normal; 50% 1 = fixed defect; 33.33% 2 = reversable defect; 78.18% 3 = other 23.93%

B. *Correlation analysis*

Pearson correlation coefficients across all features were computed and visualized via a correlation heatmap (Fig. 4). Target-specific correlations are illustrated in a ranked bar plot (Fig. 5), revealing significant variability in predictive strength among the features. Notably, cholesterol (chol) and fasting blood sugar (fbs) exhibit the weakest correlations (coefficients below 0.1), whereas exang, cp, oldpeak, and thal demonstrate the strongest correlations (coefficients exceeding 0.4).

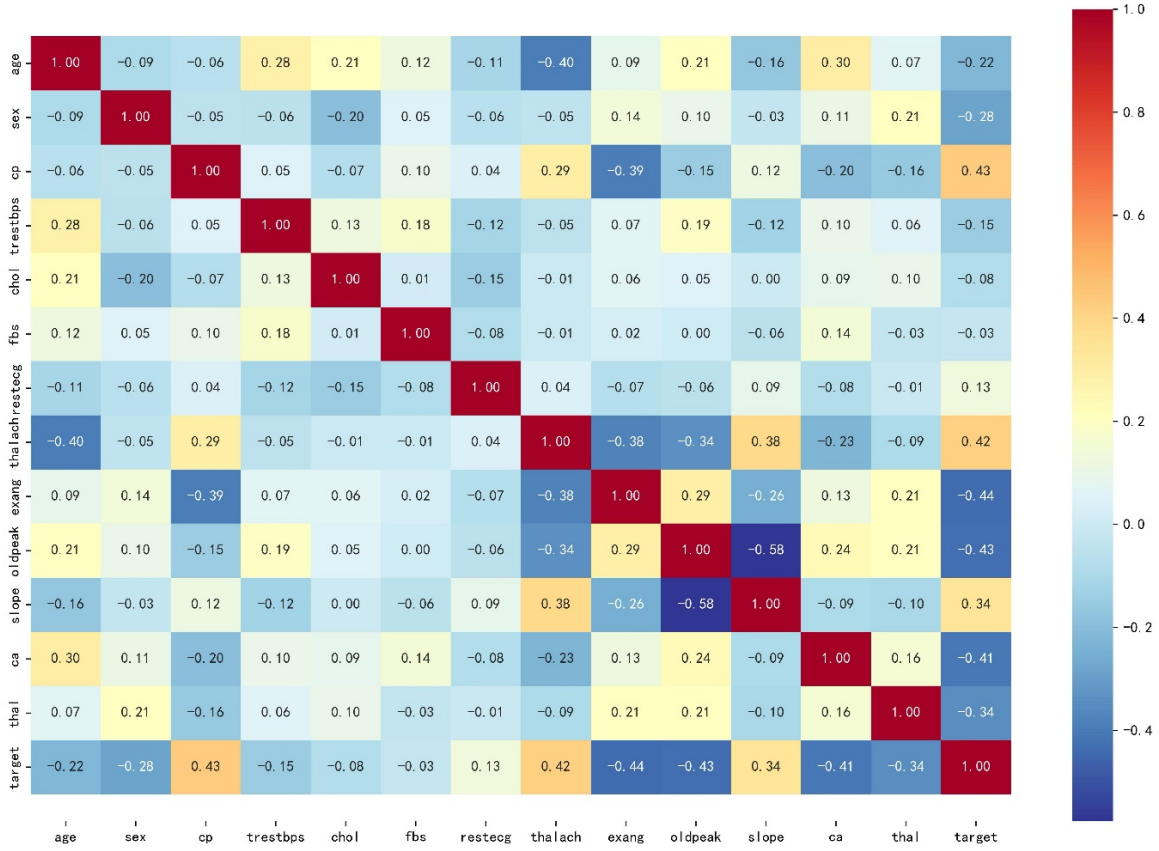

Fig. 4. Correlation value visualization using the heat map.

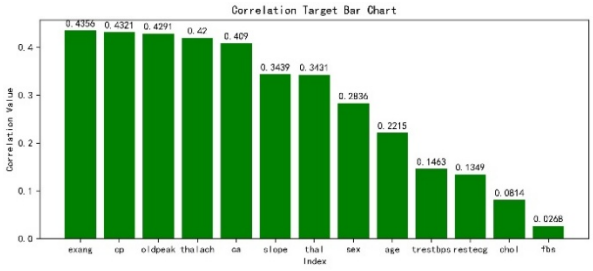

Fig. 5. Ranked bar chart.

C. *Fiture selection*

The initial dataset for heart disease prediction comprised 76 features related to patient health status. Through rigorous clinical relevance assessment, 13 predictors with high clinical significance and predictive value were retained, excluding the target variable. These features encompass key cardiovascular risk factors spanning three domains: demographic variables (age, sex), physiological indicators (trestbps, chol, thalach, oldpeak), and clinical examination results (cp, restecg, exang,

slope, ca, thal, fbs). The remaining 62 features were excluded due to statistically insignificant correlations with cardiac outcomes.

Despite very low correlation coefficients of chol (r = 0.08) and fbs (r = 0.02) with the target variable, removing these features or applying PCA (retaining 95% of the variance) did not enhance model performance. As a result, all 13 features were employed in this study to develop the machine learning models.

D. *Optimization of hyperparameters*

Hyperparameter tuning critically impacts model performance and generalizability. This study employed both randomized grid search and Bayesian optimization algorithms for hyperparameter tuning. Specifically, LR, DT, RF, and MLP models—each with relatively fewer tunable parameters—used grid search, whereas the more complex XGBoost (XGB) employed a hybrid approach combining grid search and Bayesian optimization. The optimal hyperparameter configurations for each model are summarized in Table III.

TABLE III. HYPERPARAMETER TUNING VALUES

Models	Hyperparameter Tuning Values
LR	default parameters

DT	max_depth = 5;
	max_features = 8;
	min_samples_split = 19

RF	Bootstrap = False;
	max_depth = 3;
	max_features = 2;
	min_samples_leaf = 2;
	min_samples_split = 19;
	n_estimators = 500

XGB	colsample_bynode = 0.97;
	colsample_bytree = 0.52;
	gamma = 5.08;
	learning_rate = 0.12;
	max_depth = 28;
	min_child_weight = 1.82;
	n_estimators = 235;
	reg_lambda = 6.65;
	subsample = 0.81

MLP	alpha = 0.001;
	early_stopping = True;
	hidden_layer_sizes = (100, 10, 10);
	learning_rate = 'adaptive';
	learning_rate_init = 0.2;
	solver = 'lbfgs'

E. *Comparative Analysis of ML Models*

The goal of this study was to identify the most effective machine learning algorithm for heart disease classification. This section summarizes the results and identifies the top-performing models based on multiple performance metrics.

As shown in Table IV and Fig. 6, performance differences among the models were modest, with all achieving satisfactory results. Notably, RF and XGB achieved the highest accuracy (90%) and superior AUC scores of 0.96 and 0.95, respectively, maintaining consistent performance across other metrics such as precision, recall, and F1-score, which approached 90%. LR and DT followed closely with 89% accuracy, while MLP achieved slightly lower accuracy at 88%, indicating potential for hyperparameter refinement.

TABLE IV. PERFORMANCE OF FIVE ML MODELS

Models	Accuracy	Precision	Recall	F1-Score	ROC-AUC
LR	0.89	0.90	0.88	0.89	0.94
DT	0.89	0.91	0.88	0.89	0.90
RF	0.90	0.90	0.90	0.90	0.96
XGB	0.90	0.91	0.89	0.90	0.95
MLP	0.88	0.89	0.87	0.88	0.94

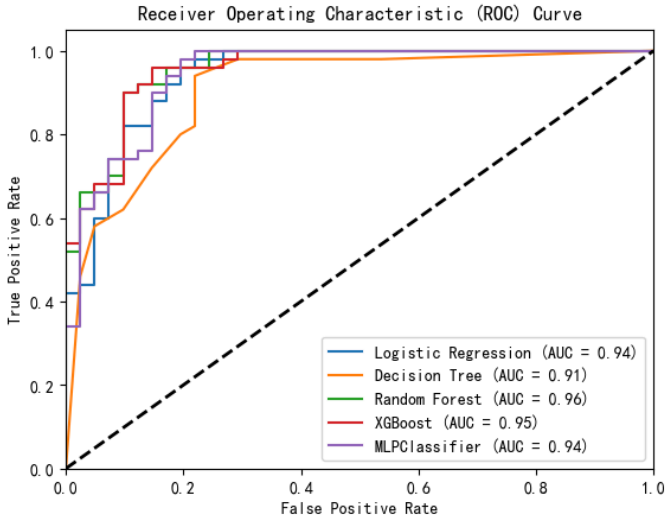

Fig. 6. Roc curve of five ML models.

V. CONCLUSIONS

This study developed and evaluated five machine learning models—logistic regression (LR), decision tree (DT), random forest (RF), XGBoost (XGB), and multilayer perceptron (MLP)—for heart disease prediction using the Cleveland dataset. Hyperparameters were optimized via randomized grid search and Bayesian optimization. Model performance was assessed based on accuracy, precision, recall, F1-score, and AUC-ROC metrics.

Results demonstrated that RF and XGB achieved the highest performance, both attaining 90% accuracy with AUC values of 0.96 and 0.95, respectively, while maintaining approximately 90% in precision, recall, and F1-score. LR and DT followed closely with 89% accuracy, whereas MLP's accuracy was slightly lower at 88%, indicating room for architectural refinement.

These findings confirm the efficacy of RF and XGB in cardiovascular risk prediction, providing reliable decision-support tools for clinicians. Limitations include reliance on a single dataset and moderate sample size, which may restrict model generalizability. Future work will focus on multi-center validation and integrating multimodal clinical data to enhance robustness.

ACKNOWLEDGMENT

This work was supported by the "Research on Student Profiling and Learning Analytics Based on Data Mining" (GKY-2023BSQD-48) and "Keyword Extraction Analysis Based on LDA Topic Modeling" (GKY-2024KYYBK-14) grants from Guangdong University of Science and Technology.

REFERENCES

[1] Taleb, C. Lippert, T. Klein, and M. Nabi, "Multimodal self-supervised learning for medical image analysis," In International Conference on Information Processing in Medical Imaging, vol. 12729, pp. 661–673, 2021.

[2] V.R. Allugunti, "Breast cancer detection based on thermographic images using machine learning and deep learning algorithms," Int. J. Eng. Comput. Sci., vol. 4, no. 1, pp. 49-56, 2022.

[3] C.M. Bhatt, P. Patel, T. Ghetia, and P.L. Mazzeo, "Effective heart disease prediction using machine learning techniques," Algorithms, vol. 16, pp. 88, 2023.

[4] N. Chandrasekhar, S. Peddakrishna, "Enhancing heart disease prediction accuracy through machine learning techniques and optimization," Processes, vol. 11, no. 4, 2023.

[5] P. K. Bhowmik, M. N. L. Miah, M. K. Uddin, M. M. H. Sizan, L. Pant, M. R. Islam, and N. Gurung, "Advancing heart disease prediction through machine learning: techniques and insights for improved cardiovascular health," British Journal of Nursing Studies, vol. 4, no. 2, pp. 35-50, 2024.

[6] G.A. Ansari, S.S. Bhat, M.D. Ansari, S. Ahmad, J. Nazeer, A.E.M. Eljialy, and Y. Jabeen Nazee, "Performance evaluation of machine learning techniques (MLT) for heart disease prediction comput.," Math. Methods Med.., pp. 1-10, 2023.

[7] K. Arumugam, Mohd Naved, Priyanka P. Shinde, Orlando Leiva-Chauca, Antonio Huaman-Osorio, and Tatiana Gonzales-Yanac, "Multiple disease prediction using machine learning algorithms," Materials Today: Proceedings, vol. 80, no. 3, pp. 3682-3685, 2023.

[8] M. K. Saini, "Machine learning techniques for precise heart disease prediction," International Journal of Science and Research (IJSR), vol. 12, no. 12, pp. 2121-2129, 2023.

[9] Salam H. Bani Hani and Muayyad M. Ahmad, "Machine-learning algorithms for ischemic heart disease prediction: a systematic review," Current Cardiology Reviews, vol. 19, no. 1, pp. 87-99, 2022.

[10] Zhou, P. Dai, A. Hou, "A comprehensive review of deep learning-based models for heart disease prediction," Artif Intell Rev.., vol. 57, no. 9, pp. 263-263, 2024.

[11] A.R. Ur, A. Yousef, Z. Adee, "Enhancing heart disease prediction using a self-attention-based transformer model," Scientific Reports, vol. 14, no. 1, pp. 514-514, 2024.