Predicting Cardiovascular Disease Using Machine Learning: A Comparative Study of Four Classification Models

Faith Nnakwe, Ryan Bouapheng, Ritesh Dumpala
Department of Computer Science
Georgia State University
Atlanta, Georgia
ujunwannakwe@gmail.com
riteshdumpala@gmail.com
Laosrb@gmail.com

***Abstract*— Cardiovascular disease (CVD) remains the leading cause of death in the United States, with nearly one million deaths reported in 2023. Early prediction is essential for facilitating timely clinical intervention, yet many machine learning approaches rely on complex architectures that offer limited interpretability. This study evaluates several machine learning models for predicting CVD using a structured clinical dataset. Logistic Regression, Random Forest, Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP) classifiers were trained and assessed using accuracy, precision, recall, F1-score, confusion matrices, ROC curves, and feature importance analysis. Logistic Regression and Random Forest demonstrated the strongest overall performance, while MLP and SVM underperformed relative to expectations on this dataset. Logistic Regression achieved the highest F1-score and provided superior interpretability, supporting its suitability for clinical decision support. The study concludes with a discussion of model behaviors, limitations, and directions for future research.**

***Keywords*—cardiovascular disease, machine learning, logistic regression, random forest, neural networks, support vector machine, ROC curve, feature importance**

I. INTRODUCTION

Cardiovascular disease (CVD) continues to be one of the leading causes of mortality in the United States, accounting for over 900,000 deaths in 2023 alone, according to the Centers for Disease Control and Prevention (CDC) [1]. Early identification of individuals at elevated risk is critical, as timely intervention can significantly reduce complications, improve long-term health outcomes, and lessen the overwhelming economic burden associated with heart disease. While traditional methods are clinically valuable, they may overlook subtle, complex interactions among patient characteristics such as age, cholesterol levels, chest pain type, and exercise-induced abnormalities. As a result, machine learning approaches have gained increased recognition as tools capable of highlighting predictive patterns from structured clinical datasets[2].

Existing research has explored a range of predictive machine learning methods, including logistic regression, neural networks, support vector machines, and ensemble learning; yet, there remains an ongoing debate on how to balance the predictive performance of a model with interpretability in high-stakes healthcare settings [3]. Complex models may capture nonlinear patterns but often lack transparency, making it hard for adoption by clinicians who must understand how predictions are generated. Conversely, simpler models offer interpretability but may underfit important physiological relationships.

This study addresses these challenges by evaluating four widely used machine learning classifiers for cardiovascular disease prediction: Logistic Regression, Random Forest, Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP). Using a structured linear clinical dataset and standardized preprocessing, each model is assessed through classification metrics, namely, accuracy, precision, recall, F1-score, confusion matrices, ROC curves, and feature importance analysis. Particular attention is given to both a model's predictive reliability and the ease of interpretability in the clinical field.

Our results show that Logistic Regression and Random Forest achieved the strongest overall performance, with Logistic Regression emerging as the best-performing model based on its F1-score and ease of interpretability. These findings demonstrate the effectiveness of machine-learning–based screening tools while emphasizing the importance of transparency and interpretability in clinical applications. The study concludes by discussing model behavior, limitations, and recommendations for future work, including incorporating additional clinical variables and exploring hybrid or explainable AI frameworks.

Link to code: **CVD-Github-Link**

II. MATERIALS AND METHODS

A. Data explanation and characterization

The dataset used in this study comprises 1,000 patient records and 14 attributes, including demographic information, clinical measurements, and a target label indicating the presence of cardiovascular disease. The dataset shape was (1000, 14), and no missing values or duplicate rows were found across any column, ensuring data completeness and reliability. The feature set includes: age, gender, chestpain, resting BP, serumcholestrol, fastingbloodsugar, restingrelectro, maxheartrate, exerciseangina, oldpeak, slope, noofmajorvessels, and target. The target variable (0 = No Disease, 1 = Disease) exhibited the following distribution:

- 580 patients (58%) were labeled with the disease.
- 420 patients (42%) had no disease.

This distribution indicates a mildly imbalanced dataset, but does not warrant resampling. Several exploratory visualizations, including histograms for numeric features and countplots for categorical variables by class, were generated to examine patterns in the data. Additional exploratory visualizations, such as violin plots and KDE density curves, were also generated to assess distribution shape and class-level separation for numeric features (figures not shown).

The correlation heatmap shown in Figure 1 highlights the strongest relationships among clinical variables. Features such as slope, chestpain, and noofmajorvessels were the features most strongly associated with the disease outcome.

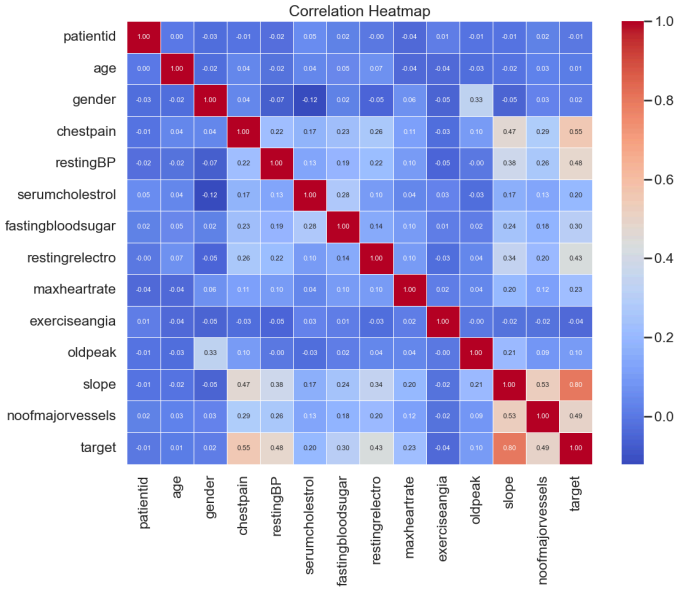

Fig 1. Correlation Heat map

B. Data preprocessing

Several preprocessing steps were performed before model training. The first step was handling missing or duplicate data. No missing data or duplicate rows were present, so no imputation or deduplication was required. Next, a feature–target separation was performed where all features except the target were assigned to the input matrix X, while the target column was assigned to y. The data set was then partitioned into an 80/20 split, producing training data of shape (800, 12) and testing data of shape (200, 12). Finally, feature scaling was then performed for models requiring normalization (e.g., Logistic Regression). Scaling was performed using StandardScaler, applied only to the training data, and later transformed on the test data for consistency.

C. Data analysis/mining methods

Four models were trained to find the best, most simplified, and reliable model to detect a cardiovascular disease:

1. Logistic Regression
   - This was the baseline linear classifier model. It predicts probabilities yielding outputs between 0 and 1. It is fast, simple, and highly interpretable, which is favored in fields such as healthcare. This model was trained on standardized features.
2. Random Forest
   - A 200-tree ensemble with balanced class weighting, capable of capturing non-linear feature interactions. Random forests significantly reduce the risk of overfitting. The algorithm performs effectively in both classification and regression tasks, making it particularly useful for large datasets. Feature importance was derived from this model.
3. Multi-Layer Perceptron (MLP)
   - A neural network with two hidden layers (32 and 16 units) using ReLU activation was trained. It is important to take note that MLP can be architecturally simplified to mimic logistic regression under appropriate constraints, such as one layer or both input and output. In this study, additional layers were included to evaluate performance relative to a more advanced model architecture.
4. Support Vector Machine (RBF Kernel)
   - A kernel-based classifier designed to capture nonlinear separability; trained with balanced class weights and probability estimation, which enables ROC analysis.

D. Evaluation and interpretations

The performance of all classification models was evaluated using standard metrics for binary classification, including accuracy, precision, recall, and F1-score [4]. In addition, confusion matrices were computed to analyze the distribution of true positives, true negatives, false positives, and false negatives for each model, which is essential in the medical field when predicting diseases, seeing as a misclassification can cause fatal results. The dataset was split into an 80/20 train–test hold-out, with 800 samples used for training and 200 for testing. All evaluation metrics reported in this study are calculated on the independent test set. For models that require feature scaling (e.g., Logistic Regression), the scaler was fit on the training data only and applied to the test data using the same parameters. The F1-score on the test

set was used as the primary criterion for selecting the best-performing model, due to its balance between precision and recall. This is particularly important in healthcare applications, where minimizing both false negatives (missed diagnoses) and false positives (incorrect disease predictions) is critical. Additional model behavior, such as feature importance and probability outputs, was examined but did not affect the primary model selection criterion.

III. RESULTS

This section presents the outcomes of the exploratory analysis, model training, evaluation, and feature interpretation. Each result corresponds directly to the analysis steps performed in the study.

A. Dataset Exploration Results

The target variable showed a mild class imbalance with 58% disease (1) and 42% no disease (0). Figure 2 illustrates the distribution of the target classes.

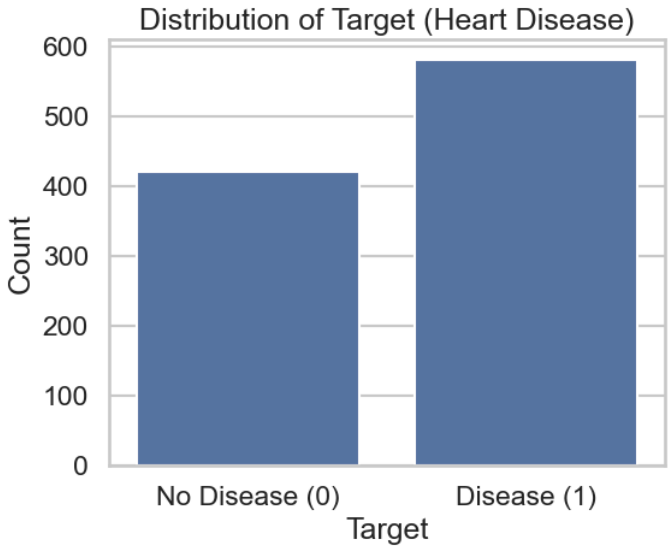

Fig 2. Distribution of Target classes

Histograms of numeric features (Figure 3) demonstrated approximately normal or slightly skewed distributions for variables such as age, restingBP, and serumcholestrol, maxheartrate, oldpeak, and noofmajorvessels. Categorical feature countplots stratified by the target (Figure 4) revealed visible differences between disease and non-disease groups, especially for chest pain, slope, and exercise.

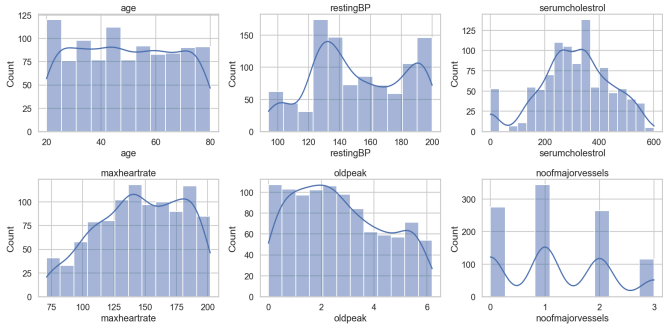

Fig 3. Histogram of numerical features

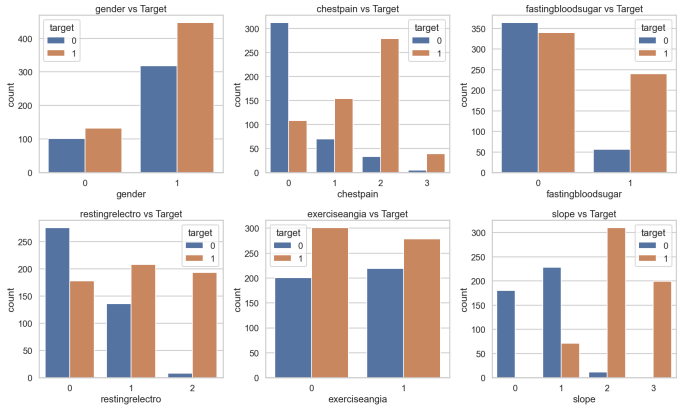

Fig 4. Count plot of categorical features.

Correlation analysis (Fig. 1) identified slope, chestpain, and noofmajorvessels as the top three features most strongly associated with cardiovascular disease.

B. Model Performance Results

Four models were trained: Logistic Regression, Random Forest, Multi-Layer Perceptron (MLP), and Support Vector Machine (SVM). Their evaluation on the test set is summarized in Table I.

Table I — Performance of Classification Models

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	**0.985**	0.9829	0.9914	**0.9871**
Random Forest	**0.985**	0.9829	0.9914	**0.9871**
Multi-Layer Perceptron	0.980	0.9746	0.9914	0.9829
Support Vector Machine	0.970	0.9741	0.9741	0.9741

C. Confusion Matrices

Confusion matrices (Figures 5–8) demonstrated strong

classification capability with very few misclassifications. Logistic Regression misclassified only 3 out of 200 test samples, along with Random Forest. There were high misclassifications with the MLP and SVM, classifying three cases as False negatives, which is a crucial mistake to make in health care when predicting the presence of a cardiovascular disease.

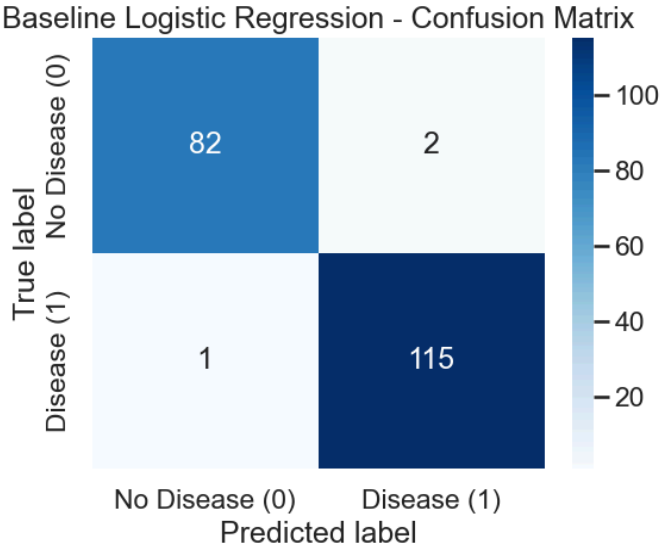

Fig. 5. Confusion Matrix for logistic regression

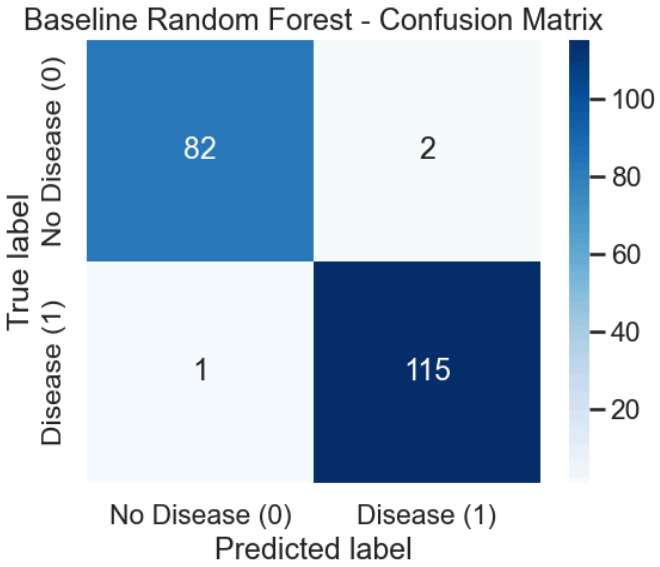

Fig. 6. Confusion Matrix for Random Forest.

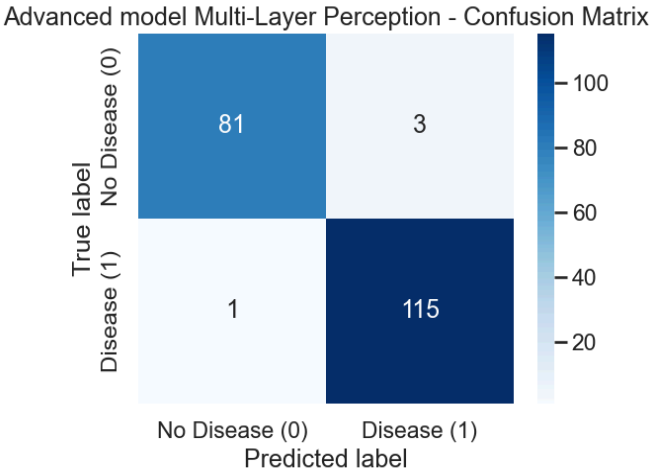

Fig. 7. Confusion Matrix for MLP

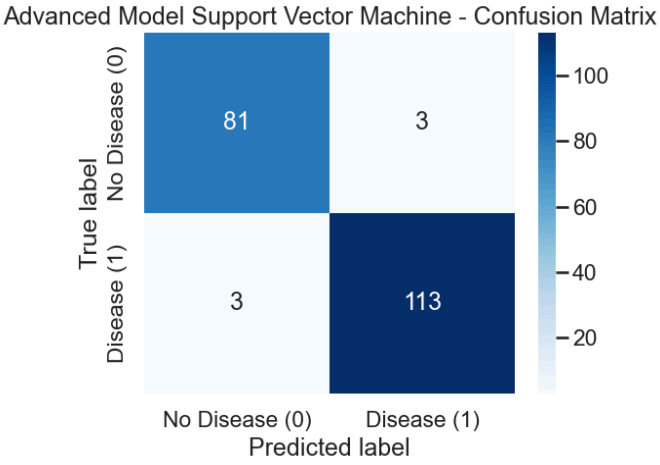

Fig. 8. Confusion Matrix for SVM

D. Feature Importance Results

Figure 9 shows the relevance of features using the Random Forest model. The barplot confirmed slope as the most influential predictor with a weight of 0.384. Other meaningful contributors included chestpain and restingBP, which were different results as seen in the correlation heat map.

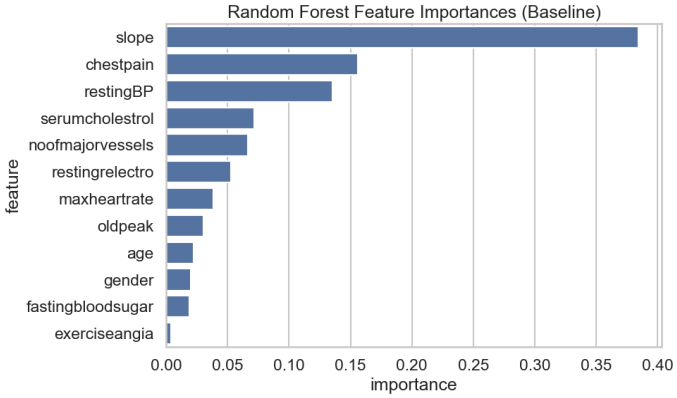

Fig 9. Feature importance using Random Forest(baseline)

E. ROC–AUC Performance Comparison

To further evaluate the discriminative ability of the models, Receiver Operating Characteristic (ROC) curves were generated for all four classifiers. The Area Under the Curve (AUC) scores provide an aggregate measure of model performance across all classification thresholds. As shown in Fig. 10, all models demonstrated excellent separability, with AUC values exceeding 0.99 for Logistic Regression, Random Forest, and SVM, and 0.994 for the MLP classifier. These near-perfect ROC curves indicate strong predictive power and consistent model behavior across thresholds.

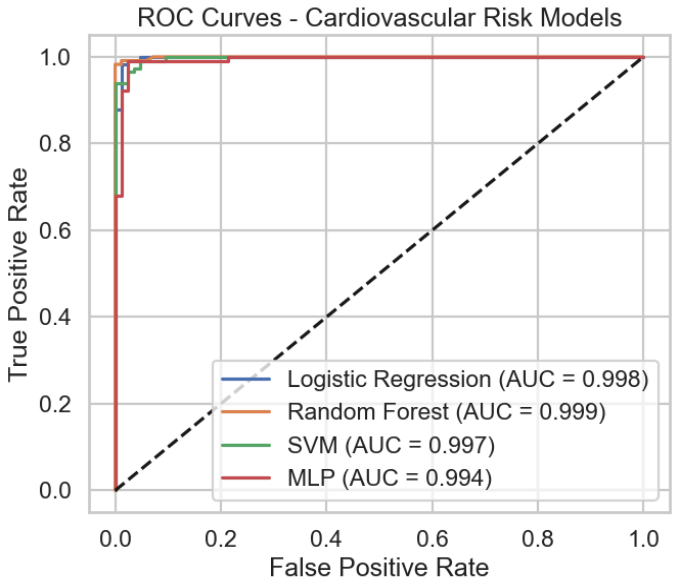

Fig. 10. ROC curves for all cardiovascular risk prediction models.

IV. DISCUSSION AND CONCLUSION

A. Interpretation of Results

The results indicate that both Logistic Regression and Random Forest achieved exceptionally high performance, with identical accuracy and F1-scores. The consistency across models, combined with the high recall (0.9914), suggests that the dataset is highly predictive and contains features that strongly differentiate between patients with and without cardiovascular disease.

A noteworthy finding is that the simplest model, Logistic Regression, performed just as well as the more complex Random Forest and neural network models. Random Forest, while accurate,
is less suited to this dataset because it learns complex nonlinear patterns that do not meaningfully exist in the data. The outcome of logistic regression is beneficial for medical applications, where explainability is essential. Coefficients from Logistic Regression can be interpreted by clinicians, making the model more suitable for real-world decision support. Feature analysis aligns with clinical knowledge: Slope of ST segment, Chest pain type, and Number of major vessels are established indicators of cardiac abnormalities, supporting the validity of the model's learned patterns.

However, the extremely high performance across all models, especially with minimal feature engineering, is somewhat surprising, suggesting that the dataset is clean, well-structured, and contains a strong signal relative to noise. The agreement between the rankings of the correlation and Random Forest feature importance also reinforces the robustness of the findings. From this study, insight is given to show that interpretability does not reduce performance; a simple Logistic Regression model can achieve near-perfect results. A few key clinical features carry most of the predictive power, which can guide future feature selection or clinical screening criteria.

B. Prediction Function Output

To demonstrate the practical application of the trained models, a prediction module was developed to automatically select the best-performing classifier (Logistic Regression, in this study) and generate both a binary prediction and a disease probability score for a new patient input. The function preprocesses the input features according to the model's requirements, such as applying standard scaling when Logistic Regression is selected, and then computes predictions using the trained classifier.

An example patient taken from the test set is shown in Table II, illustrating the feature values passed into the prediction function.

Table II — Example Patient Features

Feature	Value
age	20.0
gender	0.0
chestpain	1.0
restingBP	143.0
serumcholestrol	432.0
fastingbloodsugar	0.0
restingrelectro	1.0
maxheartrate	113.0
exerciseangia	0.0
oldpeak	1.8
slope	1.0
noofmajorvessels	0.0

When evaluated using the prediction function, the best model returned:

- Predicted label: 0 (No Disease)
- Predicted probability of disease: 0.0674

Indicating that the patient, given the following characteristics, has a low risk of high disease. This demonstrates that the model, when deployed, can not only provide a classification output but also produce a risk probability, making it suitable for clinical decision environments.

C. Limitation and Future Work

This study has several limitations. Logistic Regression assumes linear relationships between features and risk, which may oversimplify complex physiological interactions. Additionally, he dataset is limited to the features provided and may not capture all relevant clinical risk factors, limiting generalizability as the model has not been validated on a diverse or external population. Future research to address these limitations includes conducting cross-validation and hyperparameter optimization (GridSearch, RandomizedSearch, Bayesian optimization), which may improve model performance. Second, a larger and more diverse dataset to improve generalizability across populations. Third, incorporating additional clinical variables (e.g., lab results, family history, imaging data) to enhance predictive power and find more complex relationships. Finally, deploying the selected model into a clinical decision-support interface for testing in a real environment and prospective evaluation.

D. Overall Summary

This study successfully evaluated machine learning models for cardiovascular disease prediction using a structured clinical dataset. Logistic Regression emerged as the top model due to its exceptional predictive performance and explainability, making it ideal for clinical adoption. Feature analysis identified clinically relevant predictors, and the results demonstrated the effectiveness of machine-learning–based screening tools for heart disease. While the study has limitations regarding dataset size and generalizability, it still provides a strong foundation for more advanced modeling and real-world implementation.

REFERENCES

[1] Centers for Disease Control and Prevention, “Heart disease facts & statistics,” CDC, Oct. 24, 2024. [Online]. Available: https://www.cdc.gov/heart-disease/data-research/facts-stats/index.html

[2] S. Wan, F. Wan, and X. Dai, “Machine learning approaches for cardiovascular disease prediction: A review,” Archives of Cardiovascular Diseases, vol. 118, no. 10, pp. 554–562, 2025, doi: 10.1016/j.acvd.2025.04.055.

[3] Y. Ning, et al., “Advancing ethical AI in healthcare through interpretability,” Patterns, vol. 6, no. 6, p. 101290, 2025.

[4] S. A. Hicks, I. Strümke, V. Thambawita, et al., “On evaluation metrics for medical applications of artificial intelligence,” Scientific Reports, vol. 12, p. 5979, 2022, doi: 10.1038/s41598-022-09954-8.

[5] A. R. Ilyas, S. Javaid, and I. L. Kharisma, “Heart disease prediction using ML,” Engineering Proceedings, vol. 107, no. 1, p. 124, 2025, doi: 10.3390/engproc2025107124.

[6] Jocelyn Dumlao. Heart Health Insights: Cardiovascular Disease Dataset [Internet]. Kaggle; [cited 2021 April 16]. Available from: https://www.kaggle.com/datasets/jocelyndumlao/cardiovascular-disease-dataset/data