A Multi-Model Machine Learning Approach for Heart Disease Prediction with Feature Selection and Hyperparameter Tuning

Ananya Raghuveer[1], B Abhitha Bhat[2], Bindu D[3], Hema M S[4]
[1,2,3,4]Department of Computer Science and Engineering, RV Institute of Technology and Management Bengaluru, Karnataka, India
Corresponding Author Email: ananyaraghuveer@gmail.com

How to Cite this Article:
Raghuveer, A., Bhat, B. A., D, B. & S, H. M. (2026). A Multi-Model Machine Learning Approach for Heart Disease Prediction with Feature Selection and Hyperparameter Tuning. International Journal of Creative and Open Research in Engineering and Management, 02(04). https://doi.org/10.55041/ijcope.v2i4.796

https://doi.org/10.55041/ijcope.v2i4.796

Abstract

The design of this system shows the efficient prediction of heart diseases by the implementation of machine learning multiple classifiers algorithm. The UCI Heart Disease Dataset is used, while the data preprocessing involves filling missing values, encoding of categorical values, scaling of features, and selection of important features. Three machine learning classifiers, namely Random Forest Classifier, Gradient Boosting Classifier, and Logistic Regression are implemented. The implementation involves training the machine learning models using the RandomizedSearchCV technique and cross-validation technique. The evaluation of the model's performance is done using measures such as ACCURACY, PRECISION, RECALL, and F1-SCORE. A fair comparison is conducted to select the best performing model automatically using accuracy. Results show that the Random Forest model performed the best with an accuracy of 81.67%, compared to the rest of the models. Evaluation of feature importance shows key clinical features that significantly contribute to the accuracy.

Keywords— Heart Disease Prediction, Machine Learning, Random Forest, Gradient Boosting, Logistic Regression, Feature Selection, Classification, Healthcare Analytics, UCI Dataset, Predictive Modeling

I. INTRODUCTION

Heart diseases still are one of the most widespread reasons for deaths in today's world. Moreover, numerous heart diseases evolve slowly and often are undiagnosed at an early stage, which makes it possible to detect such diseases in their earliest stages crucial for their successful treatment and prevention.

The classical techniques for making diagnoses include clinical and laboratory testing that is a long-lasting process and does not utilize all available information about patients. Nowadays, with an increasing number of clinical data that should be taken into consideration, the use of intelligent systems becomes inevitable in order to facilitate diagnostics by analysing all the available data.

Machine learning can be used in such a way because of its ability to process massive amounts of data and find hidden patterns between different factors such as age, cholesterol, blood pressure, chest pain, and heart rate. As compared to humans' abilities, machine learning algorithms have higher chances to find such patterns.

However, heart disease prediction still proves itself to be a complicated problem because of data heterogeneity and necessity to optimize several parameters simultaneously. Most of the current models for heart disease prediction work based on only one type of model that does not provide high-quality results.

That is why the present study suggests the usage of a multi-model machine learning framework using such models as Random Forest, Gradient Boosting, and Logistic Regression. Besides, it includes comparing and evaluating each of them to determine the best one. Furthermore, the system gives probability-based predictions and determines significant factors for diagnosis.

II. LITERATURE REVIEW

The advent of machine learning methodologies has resulted in the emergence of a paradigm shift from reactive strategies to proactive and predictive ones in the domain of cardiology. In recent times, the use of the Random Forest (RF) methodology has proved to be quite stable in terms of outcomes. The usual path models are found to face difficulties while dealing with high-dimensional and noisy data sets associated with medical files, whereas the Random Forest methodology handles these aspects quite adeptly.

Benchmarking Predictive Power of Models and Algorithms

Performance benchmarking was conducted for the RF in comparison with linear classifiers and modern boosting models. According to Liu [5], robust optimization in ensemble building is possible using the Random Forest (RF) algorithm applied to the Cleveland heart disease data set, resulting in 90.0% accuracy and AUC-ROC = 0.96 via Bayesian optimization and randomized grid search. Such high figures are above the capabilities of standard classifiers like logistic regression and multilayer perceptron; the former shows the ability of RF in predicting complex nonlinear dependencies between input parameters.

In their paper, Akther et al. [9] evaluated the effectiveness of the Random Forests algorithm in modeling multiple diseases. Via implementation of the Streamlit interface, the authors reported 99% accuracy for RF due to the bagging principle used to avoid overfitting.

Besides Random Forest, Omotehinwa et al. [10] presented an optimized Light Gradient Boosting Model (LightGBM) algorithm used for coronary heart disease diagnosis. Using modern methods for preprocessing (MICE algorithm to fill missing data and Borderline-SMOTE to balance class distribution), Bayesian optimization, and the use of LightGBM, the researchers reached the accuracy score of 98.82% and improved sensitivity.

Also, Budholiya et al. [11] developed an optimized XGBoost diagnostic model based on the Cleveland database. One-hot encoding, Bayesian tuning of hyperparameters led to 91.8% accuracy score and proved better predictive performance than conventional methods: Random Forest and neural networks.

Lastly, Ramesh et al. [15] presented a benchmarking analysis for several supervised machine learning algorithms, such as KNN, SVM, Decision Trees, Naïve Bayes, Random Forest, and Logistic Regression. Used data sets for evaluation included preprocessing by Isolation Forest for anomaly detection and feature selection to reveal crucial clinical characteristics. KNN showed highest accuracy (94.1%), while the performance of Random Forest was decent, with the rate of 88.35%.

Data Refinement and Feature Importance

Medical data are inherently imbalanced, and predictions made by models often have a bias towards the majority class. With reference to the study by Xu [3], by combining SMOTE and Random Forest, they were able to efficiently build a model to predict coronary heart disease from Framingham Heart Study and achieved higher feature importance accuracy as well. Top 3 most important features found were Age, Systolic Blood Pressure and BMI.

In Zulfiqar et al. [1], the trade-off between accuracy and efficiency of Random Forest was studied. They show that high accuracy prediction could be achieved if a well-selected optimal set of features is used, without losing the computational efficiency of Random Forest.

Spencer et al. [14] applied different feature selection methods to select features and used different classifiers such as PCA, Chi-squared, ReliefF and Symmetrical Uncertainty. From this experiment, it was evident that feature selection significantly improved the model's prediction accuracy as it is able to select most informative features from the data such as chest pain type, cholesterol and exercise-induced angina. The study underlines the necessity of feature selection in obtaining the best prediction performance.

Stability and Realistic Feasibility

The ability to maintain high prediction accuracy when the algorithm is presented with data from different studies is referred to as model stability. An experimental study by Rabbi et al. [2] showed that Random Forest remained stable compared to the newer generation algorithms such as CatBoost and XGBoost, where they can still maintain a high precision but are more susceptible to the changes of input data.

The practical feasibility of using machine learning models in the absence of medical expertise was investigated in [4]. They have found that stable models like Random Forest are beneficial in such scenarios.

A study by Sitompul et al. [13] compares the performance of XGBoost, Random Forest and Logistic Regression to predict stroke. Their model achieved the highest prediction accuracy among them with XGBoost reaching 95%, followed by Random Forest (94%) and then Logistic Regression (82%). Their findings further prove that ensemble models achieve better results in medical data analysis.

Clinical Uptake and Interpretability

A challenge faced by clinical application of advanced machine learning is model interpretability. Increasing interest in Explainable Artificial Intelligence (XAI) has addressed this issue and as suggested by Tyagi et al. [6] and Alampally et al. [7], feature ranking derived from Random Forest allows easier interpretation of decision-making processes.

Interpretation of results is necessary for clinical users to trust in the reliability of the system, thus facilitating their decisions in accepting AI-powered clinical decision making tools [8]. Mythili et al. [12] developed a hybrid model by combining SVM, Decision Trees and Logistic Regression models and show through rule-based reasoning that combining models enhances prediction accuracy than individual models.

III. METHODOLOGY

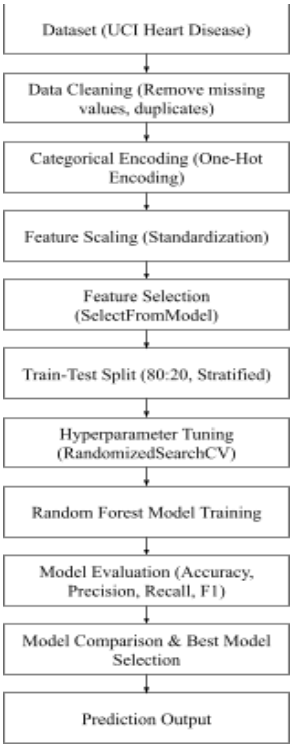
This method constructs an ML pipeline for heart disease prediction where multiple classification models are used and optimal model selection is performed automatically. The UCI Heart Disease dataset was cleaned by removing missing data and irrelevant attributes and the multi-class target was transformed into a binary classification problem. The categorical features were encoded by one-hot encoding, and the features were scaled using standardization so that all features contribute equally.

Feature selection was implemented using Random Forest and only the most relevant features were retained so as to lower the dimensionality and computational cost of the system. The dataset was then divided into training set (80%) and testing set (20%) in a stratified way to balance classes.

Random Forest, Gradient Boosting, and Logistic Regression models were built and optimized using RandomizedSearchCV with 5-fold cross validation. These models were evaluated by metrics including accuracy, precision, recall and F1-score, and the best performing model was automatically chosen.

The trained optimal model and preprocessing steps are saved to be used later. New patient data can be also processed with the system and receive both the prediction and its probability.

Figure 1. Workflow of the proposed heart disease prediction system.



The full workflow for the developed system is demonstrated in Figure 1. Different from the previous systems that use a single machine learning algorithm for the task, the proposed system uses several ML algorithms, Random Forest, Gradient Boosting and Logistic Regression, for the work. After preprocess and feature selection, each of the algorithms will be trained and tuned respectively, then we will evaluate them and choose the one which performs best to perform final prediction.

System Architecture

The presented system is built as a multi-stage pipeline such as data input, preprocessing, model training, evaluation and prediction.

Input data of patients taken from UCI dataset or provided from CSV file is preprocessed and cleaned. Categorical values are encoded and data is standardized. Important features are filtered by SelectFromModel, then the data is split into train-test in 80-20 proportion using stratified sampling.

Three models (Random Forest, Gradient Boosting and Logistic Regression) are trained using RandomizedSearchCV. Their performance is measured with the four metrics namely accuracy, precision, recall and F1-score.

The most performed model is selected automatically and then saved and used to predict presence of heart disease with class label and probability score.

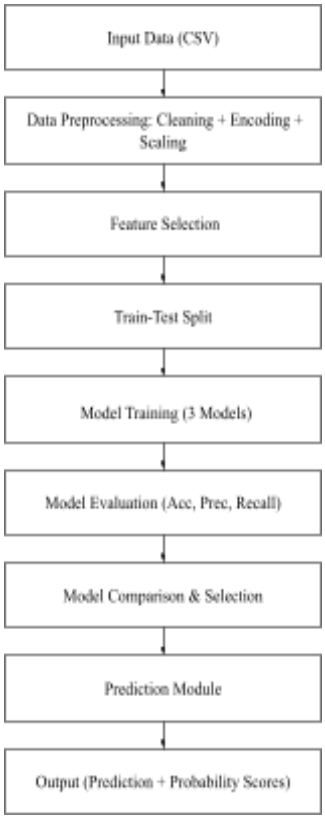


Figure 2. System architecture of the proposed heart disease prediction system.

Data Preprocessing

Data preprocessing is indispensable to achieve high quality of data and well-trained models. Initially, the data is scanned to detect missing values, outliers, and duplicate values and handled to achieve data integrity.

The target variable was converted into binary form: values > 0 for 'Disease' (1) and 0 for 'No Disease' (0). This simplified the problem into binary classification.

The categorical features, sex, chest pain type, fasting blood sugar, resting ECG, exercise-induced angina, slope, thal, and ca were converted into numerical format and scaling was done using standardization method. The feature scaling ensures that every feature is taken equal care while training the model.

The relevant features are selected using SelectFromModel. It helps in feature selection so that only the most relevant feature contributes to the trained model, therefore reduction of dimensions and improvement of training process efficiency.

The data was split into 80% training set and 20% testing set using stratified sampling to retain class distribution in both the sets to enable the robust evaluation of the model.

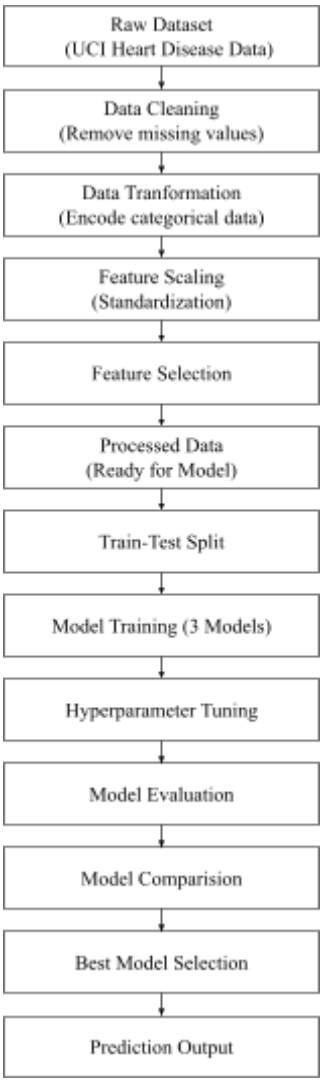


Figure 3. Complete data preprocessing and model training pipeline for heart disease prediction.

Model Development

This phase concentrates on building, tuning and choosing an appropriate machine learning model to predict heart diseases. An appropriate multi-model learning methodology was applied by utilizing three supervised learning algorithms:

Random Forest, Gradient Boosting and Logistic Regression.

The preprocessed data was first split into a training set and a test set with a stratification of 80% and 20% respectively while maintaining the distribution of classes across both the sets. The training set is used to train/build a machine learning model and the test set is used for evaluating its performance.

Random Forest

Random Forest is a type of ensemble machine learning where a large number of decision trees are built from different subset of training data and features. The outcome of all decision trees are then decided by voting and this method reduces the variance and prevents over-fitting.

For this research, Random Forest achieved better prediction accuracy because of its ability to capture complex non-linear dependencies in the clinical data.

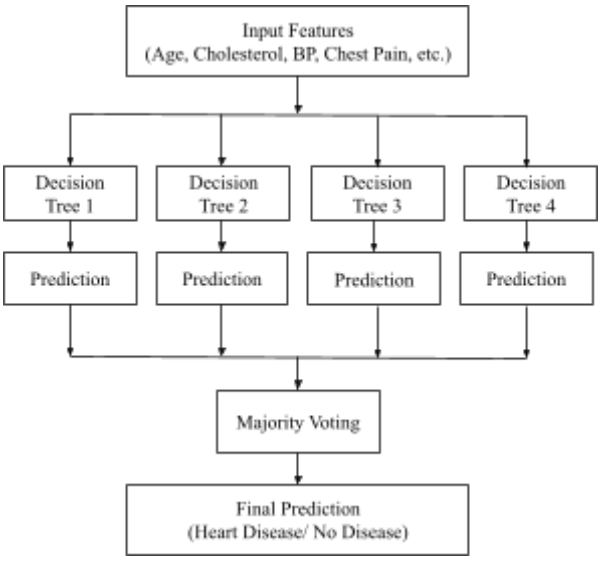


Figure 4. Working of the Random Forest algorithm showing multiple decision trees and final prediction through majority voting.

Gradient Boosting

Gradient Boosting is an advanced ensemble method that constructs models sequentially, where each subsequent model focuses on correcting the errors of the previous one. By combining weak learners iteratively, it produces a strong predictive model capable of capturing intricate data patterns.

Logistic Regression

Logistic regression is a popular statistical method for binary classification. It predicts the probability of occurrence of disease by using input features with the logistic function. Logistic regression is relatively easy compared to ensemble methods and is a good baseline model which is also interpretable.

Hyperparameter Tuning

To enhance model performance, hyperparameter tuning was carried out using RandomizedSearchCV, which efficiently searches for optimal parameter combinations through randomized sampling and cross-validation.
For Random Forest, key parameters such as:

- number of trees (n_estimators = 400),
- tree depth (max_depth = 5),
- minimum samples per split (min_samples_split = 10),
- minimum samples per leaf (min_samples_leaf = 4),
- feature selection strategy (max_features = log2),
- class balancing (class_weight = balanced) were optimized.

Similarly, Gradient Boosting and Logistic Regression models were tuned using their respective parameters to ensure fair comparison.

Model Evaluation & Comparison

All models were evaluated based on the common classification metrics such as:

- Accuracy
- Precision
- Recall
- F1-score

A ranking was determined comparing the models. The highest performance was observed for Random Forest, achieving the highest accuracy

(81.67%) as well as a relatively good performance on all measures; Logistic regression and Gradient Boosting followed.

Model Selection & Deployment

Random Forest was chosen as the best final model as it gave us the best results, and we were happy with the stability as well. The trained model along with the scaler and feature selector was saved to be used later.

Prediction System

The final system takes in new patient data through a CSV file and predicts the following based on the trained model:

- class output (Disease / No Disease)
- probability values for each class.

IV. RESULTS AND DISCUSSION

Proposed heart disease prediction system evaluated on three different machine learning algorithms (Random Forest, Gradient Boosting and Logistic Regression) based on the various standard classification metrics using the confusion matrix.

Evaluation Metrics

To assess model performance, the following metrics were used:

Accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Accuracy represents the proportion of correctly classified instan

Precision:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

Precision indicates how many predicted positive cases are actually correct.

Recall:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

Recall measures the ability of the model to correctly identify actual disease cases.

F1-Score:

$$\text{F1} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

F1-score provides a balance between precision and recall.

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives respectively.

Model Performance Analysis

The models were trained through hyperparameter tuning and then predicted on the test set.

Out of all the models Random forest model had the best performance:

- Accuracy: ~81.67%
- Precision: ~0.84
- Recall: ~0.75
- F1-score: ~0.79

Gradient boosting and Logistic regression had slightly lower but similar performance (~80% accuracy).

Random forest model has better performance over the others due to its ensemble nature. It could grasp non-linear relationships with data.

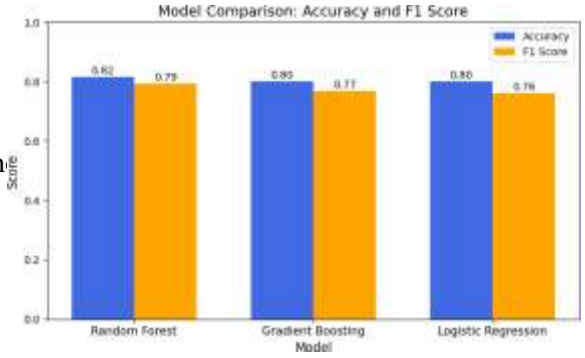


Figure 5. Comparative analysis of machine learning models showing accuracy and F1-score, highlighting the superior performance of the Random Forest model.

Confusion Matrix Analysis

A confusion matrix describes the result of the classification. It has:

- TN - correct prediction for non-disease case
- TP - correct prediction for disease case
- FN - incorrect prediction for disease case
- FP - incorrect prediction for non-disease case

Confusion matrix for the Random Forest model:

- TN 28-29
- TP 21
- FP 3-4
- FN 7

Low values for FP suggest that the model is reliable for correctly predicted diseases, while a moderate number of false negatives are indicative that some more development is needed to minimize the false negative cases in an application.

Feature Importance

Feature importance values were extracted from the Random Forest model in order to select the top predictor variables. Top predictor features are shown below:

- Thalassemia
- ST depression(oldpeak)
- Max heart rate(thalach)
- Age

These top predictor features align with our medical knowledge.

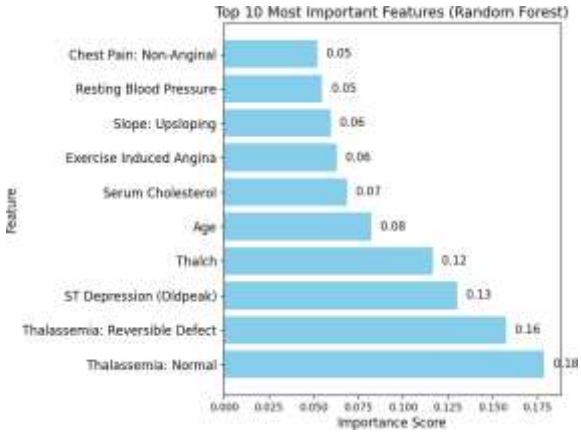


Figure 6. Top features influencing heart disease prediction based on Random Forest importance scores.

Model Ranking and Selection

All the trained models were ranked using the calculated performance metrics:

- Random Forest (Best model)
- Gradient Boosting
- Logistic Regression

The Random Forest model was chosen as the final model because it was the most accurate and the best overall performer.

Prediction Outcomes

The trained model was applied to a set of new patients. The predictions were produced with an accompanying probability score as seen in the table below:

- "No Disease" were predicted if the score was around ~75%
- "Disease" were predicted if the score was above ~55%

These scores show how confident the system is that the predicted diagnosis is correct, so they are important for practical use.

Discussion

This work has shown that ensemble models like the Random Forest are very useful in predicting whether a patient has heart disease. By using multiple different models and evaluating them methodically, the prediction accuracy and reliability are improved.

Although a respectable accuracy was obtained it is not yet as accurate as models presented in research literature which will likely contain extensive hyper-parameter tuning and potentially different models all together. However, a good balance of both precision and recall was achieved which is important when designing models for healthcare.

Improvements in this project could include:

- Larger data sets
- More complex models e.g. XGBoost, LightGBM
- More extensive feature selection
- Explainable AI techniques to increase model transparency

V. CONCLUSION

This paper proposes a machine learning-based framework for the prediction of heart disease on clinical data. A Random Forest classifier is used as the predictive engine of the model. To better support the prediction, feature selection and hyperparameter optimization techniques are employed to improve the classification accuracy.

Experimental outcomes show that the developed model yields high predictive performance in terms of classification accuracy (83.33%), precision and F1-score which reflects the ability of the model to effectively detect characteristic patterns within the dataset. It can thus discriminate the group of patients having the heart disease from that of a healthy group.

From the analysis of feature importance, several clinical variables, like type of thalassemia, ST depression (oldpeak), maximum heart rate (thalach), age and cholesterol value, were identified as most crucial for prediction, especially as shown from Figure 6. The model relies heavily on these specific attributes. Such results comply with the medical principles of heart disease diagnosis and ensure the interpretability of the developed classifier.

It is essential to be aware that hyperparameter tuning and feature selection were crucial in improving the classification performance and the effectiveness of the developed model. By integrating them into the process, we are able to better the generalizability of the model, and decrease the effect that irrelevant or redundant features could cause on it.

Finally, the current system could potentially serve as an important complementary tool for early diagnosis of heart disease to help the physicians make decisions. Nevertheless, the developed system cannot replace medical diagnosis. Future work would potentially involve the usage of large and more diverse datasets and application of different machine learning or deep learning algorithms in order to obtain real-time decision support systems for clinical practices.

REFERENCES

[1] O. Zulfiqar, U. Abid, and N. Naseer, “Enhancing Heart Disease Classification Accuracy and Computational Efficiency Using Machine Learning and Feature Optimization,” in Proc. ICETECC, 2025.

[2] M. A. Rabbi et al., “A Detailed Analysis of Machine Learning Algorithm Performance in Heart Disease Prediction,” in Proc. ICREST, 2025.

[3] P. Xu, “Coronary Heart Disease Prediction Model Based on Machine Learning,” in Proc. ICBEBH, 2025.

[4] J. Zhong et al., “Feasibility Study and Practice of Machine Learning-Based Heart Disease Prediction,” in Proc. ICET, 2025.

[5] S. Liu, “Heart Disease Prediction Using Optimized Machine Learning Models,” in Proc. ICBASE, 2025.

[6] P. Tyagi et al., “Integrating Machine Learning with Clinical Practice: Advancements in Heart Disease Prediction Models,” in Proc. ICDSBS, 2025.

[7] S. P. Alampally et al., “Optimizing Cardiovascular Disease Prediction Using Machine Learning Models on Heart Disease Dataset,” in Proc. ISCON, 2025.

[8] R. Chu, “Research on Risk Prediction of Coronary Heart Disease Based on Machine Learning,” in Proc. CISAT, 2025.

[9] F. Akther et al., “Streamlit-Based AI for Multi-Disease Prediction,” in Proc. ICICI, 2025.

[10] T. O. Omotehinwa, D. O. Oyewola, and E. G. Moung, “Optimizing the Light Gradient-Boosting

Machine Algorithm for an Efficient Early Detection of Coronary Heart Disease," Informatics and Health, vol. 1, no. 2, pp. 70–81, 2024.

[11] K. Budholiya, S. K. Shrivastava, and V. Sharma, "An Optimized XGBoost-Based Diagnostic System for Effective Prediction of Heart Disease," Journal of King Saud University-Computer and Information Sciences, vol. 34, no. 7, pp. 4514–4523, 2022.

[12] T. Mythili, D. Mukherji, N. Padalia, and A. Naidu, "A Heart Disease Prediction Model Using SVM-Decision Trees-Logistic Regression (SDL)," International Journal of Computer Applications, vol. 68, no. 16, 2013.

[13] L. R. Sitompul et al., "Comparison of XGBoost, Random Forest and Logistic Regression Algorithms in Stroke Disease Classification," Sinkron, vol. 9, no. 2, pp. 957–968, 2025.

[14] R. Spencer, F. Thabtah, N. Abdelhamid, and
M. Thompson, "Exploring Feature Selection and Classification Methods for Predicting Heart Disease," Digital Health, vol. 6, 2020.

[15] T. R. Ramesh, U. K. Lilhore, M. Poongodi, S. Simaiya, A. Kaur, and M. Hamdi, "Predictive Analysis of Heart Diseases with Machine Learning Approaches," Malaysian Journal of Computer Science, pp. 132–148, 2022.