Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.4, August 2025, pp. 1559~1568
e-ISSN: 2548-6861

Comparative Analysis of the C5.0 Algorithm and Other Machine Learning Models for Early Detection of Multi-Class Heart Disease

Mardhatillah[1]*, Hafizh Al-Kautsar Aidilop[2], Asrianda [3]***
* Teknik Informatika, Universitas Malikussaleh
mardhatillah.180170011@mhs.unimal.ac.id [1], hafizh@unimal.ac.id [2], asrianda@unimal.ac.id [3]

Article Info

Article history:

Received 2025-06-04
Revised 2025-06-27
Accepted 2025-07-03

Keyword:

Classification,
Decision Tree,
Heart Diseases,
Early Detection.

ABSTRACT

Cardiovascular diseases represent the leading cause of mortality worldwide, making accurate and early detection a critical factor for effective medical intervention and improved patient prognosis. While machine learning (ML) offers promising tools for predictive diagnostics, many existing studies rely on single-algorithm approaches or less-than-robust validation methods, thereby limiting the generalizability and real-world applicability of their findings.This study aims to conduct a rigorous, head-to-head comparative evaluation of multiple machine learning algorithms for the multi-class classification of heart disease, with the goal of identifying the most effective and reliable model for this complex clinical task.We utilized a private dataset comprising 300 patient medical records, each described by 11 clinically relevant features. To ensure a robust and unbiased evaluation, a stratified 5-fold cross-validation methodology was employed. Five widely-used classification algorithms were evaluated: Naïve Bayes (NB), Logistic Regression (LR), Random Forest (RF), a C5.0-analog Decision Tree (DT), and Support Vector Machine (SVM). Model performance was assessed using standard metrics, including accuracy, precision, recall, and F1-score.The comparative analysis revealed that the Naïve Bayes algorithm delivered superior performance, achieving the highest mean accuracy of 43.33% (±4.22%). It also led in other key metrics with a mean precision of 43.40%, recall of 43.64%, and an F1-score of 41.26%. Other algorithms, such as Logistic Regression (40.67% accuracy) and Random Forest (39.33% accuracy), demonstrated competitive performance but were ultimately surpassed by the Naïve Bayes model in this specific multi-class classification context.This research underscores the critical importance of employing robust validation techniques and comprehensive comparative analyses to identify optimal models for clinical applications. The Naïve Bayes algorithm emerges as a strong candidate for developing a reliable clinical decision support system for the early differentiation of various heart conditions, providing a foundation for future data-driven diagnostic tools.

1. INTRODUCTION

Cardiovascular diseases (CVDs) constitute a paramount global health crisis, consistently ranking as the foremost cause of mortality and morbidity across diverse populations [1, 2]. The spectrum of CVDs, encompassing conditions such as coronary artery disease (CAD), congestive heart failure, and cardiac arrhythmias, imposes a staggering burden on public health systems, national economies, and individual quality of life [4]. The insidious progression of many of these diseases often means that symptoms only manifest at an advanced stage, where treatment options may be limited and less effective. Consequently, the paradigm of modern cardiology has increasingly shifted towards proactive prevention and early detection. Identifying individuals at high risk or diagnosing a condition in its nascent stages is fundamental to enabling timely, effective interventions that can halt disease progression, mitigate severe complications, and significantly reduce mortality rates [8, 9].

The digital transformation of healthcare has ushered in an era of data-driven medicine, with the proliferation of Electronic Health Records (EHRs) generating vast, high-dimensional datasets. Within this landscape, machine learning (ML) has emerged as a transformative technology, offering sophisticated tools to extract clinically actionable insights from complex medical data [10]. For cardiovascular medicine, ML models hold the potential to function as powerful decision support systems, capable of identifying subtle, non-linear patterns and inter-feature dependencies that may elude traditional risk scores and human interpretation [15].

However, the existing body of literature on ML for heart disease prediction, while extensive, is marked by significant methodological heterogeneity and limitations. A substantial portion of prior research has focused on the performance of a single, isolated algorithm, making it difficult to ascertain its relative efficacy. Furthermore, many studies have employed simplistic validation techniques, most commonly a single train-test split (e.g., 80:20 or 70:30 partitions). While straightforward to implement, this method is highly susceptible to sampling bias; the model's performance can vary dramatically depending on the random composition of the training and testing sets. This can lead to overly optimistic or pessimistic performance estimates and raises critical concerns about model overfitting, where a model learns the noise specific to the training data rather than the underlying generalizable patterns [17, 20]. Such limitations severely curtail the translational potential of these models into reliable clinical practice.

To address these critical gaps, this study undertakes a systematic and methodologically rigorous comparative analysis. We evaluate and benchmark five distinct, widely-recognized ML algorithms—Naïve Bayes, Logistic Regression, Random Forest, a Decision Tree (as an analog for C5.0), and Support Vector Machine—for the challenging task of multi-class heart disease classification. Our objective is to differentiate between five diagnostic categories: Coronary Heart Disease, Rheumatic Fever, Heart Failure, Cyanotic Congenital Heart Disease, and a "No Disease" classification.

The primary contribution of this work is the establishment of an evidence-based hierarchy of model performance for this specific clinical problem, grounded in a robust validation framework. By employing a stratified 5-fold cross-validation approach, we ensure that our results are stable, reproducible, and provide a more accurate reflection of each model's true generalization capability. This research moves beyond a singular focus to provide a holistic, comparative perspective, thereby offering a more reliable foundation for the future development of ML-powered tools for the early and accurate detection of heart disease.

II. METHOD

The research methodology was systematically designed to implement and evaluate the C5.0 algorithm for the early detection of heart disease. The main stages include data collection and characterization, data preprocessing, C5.0 algorithm implementation, experimental design, and model performance evaluation.

A. Data Source and Dataset Characteristics

The dataset used in this study comprises 300 patient records, categorized into five distinct heart disease classes: Coronary Heart Disease, Rheumatic Fever, Heart Failure, Cyanotic Congenital Heart Disease, and a "No Disease" category. The distribution of these classes is relatively balanced, with Coronary Heart Disease having the highest frequency at 62 cases, followed closely by Rheumatic Fever (61 cases), Heart Failure (60 cases), and Cyanotic Congenital Heart Disease (60 cases). The "No Disease" category has the lowest frequency, with 57 cases. This near-uniform distribution across classes minimizes the risk of class imbalance bias in model training, ensuring that the machine learning algorithms evaluated in this study can effectively learn patterns for each category without undue bias toward any single class.

TABLE 1
DISTRIBUTION OF HEART DISEASE CATEGORIES

Heart Disease Category	Frequency
Coronary Heart Disease	62
Rheumatic Fever	61
Heart Failure	60
Cyanotic Congenital Heart Disease	60
No Disease	57

The data used in this study were obtained through observation and analysis of medical record data of patient symptoms at a Community Health Center from March to May 2025. A total of 300 patient records were collected and used for modeling. Each patient record consists of several attributes relevant to heart disease diagnosis. These attributes include Age (years), Gender (Male/Female), Systolic Blood Pressure (mmHg), Diastolic Blood Pressure (mmHg), Cholesterol Level (mg/dL), Blood Sugar Level (mg/dL), Chest Pain (Yes/No), Shortness of Breath (Yes/No), Fatigue (Yes/No), and Heart Rate (beats per minute). The target class is the Heart Disease variable, classified into four specific types: Acute Coronary Syndrome (ACS), Heart Failure, Rheumatic Fever, Cyanotic Congenital Heart Disease (PJB Sianotik), and a "None" category for patients not diagnosed with these four conditions. Detailed attribute characteristics are presented in Table 1. A good understanding of these dataset characteristics is an important basis for the preprocessing stage and interpretation of model results, as the variation and data type of each attribute will affect how the C5.0 algorithm processes information.

TABLE 2.
CHARACTERISTICS OF RESEARCH DATASET ATTRIBUTES

Attribute Name	Data Type	Brief Description	Example Value
Age	Numeric	Patient's age in years	45, 60
Gender	Categorical	Patient's gender	M, F
Systolic	Numeric	Systolic blood pressure (mmHg)	120, 140
Diastolic	Numeric	Diastolic blood pressure (mmHg)	80, 90
Cholesterol	Numeric	Total cholesterol level (mg/dL)	180, 220
Blood Sugar	Numeric	Random blood sugar level (mg/dL)	100, 150
Chest Pain	Categorical	Complaint of chest pain	Yes, No
Shortness of Breath	Categorical	Complaint of shortness of breath	Yes, No
Fatigue	Categorical	Complaint of unusual fatigue	Yes, No
Heart Rate	Numeric	Heart rate per minute	70, 90
Heart Disease	Categorical	Heart disease diagnosis (target class)	ACS, None, etc.

B. Data Preprocessing

Before the data could be used to train the C5.0 model, several preprocessing steps were performed to ensure data quality and suitability. This stage is crucial because machine learning algorithms are sensitive to the format and quality of input data. The preprocessing steps included:

1. Blood Pressure Feature Transformation
   The Blood Pressure feature, which might initially be recorded in a combined format (e.g., "140/90"), was separated into two independent numerical features: Systolic (upper value) and Diastolic (lower value). This allows the model to analyze the contribution of each blood pressure component separately.
2. Categorical Feature Encoding
   The C5.0 algorithm requires input in numerical format or can handle categorical data internally; however, for consistency and potential use with certain Python libraries, categorical features were converted into numerical representations. For example, Gender was changed to '0' for Male and '1' for Female. Similarly, binary features like Chest Pain, Shortness of Breath, and Fatigue were converted to '1' if 'Yes' and '0' if 'No'.

This preprocessing aims to produce a clean and structured dataset, ready for use in the C5.0 modeling stage, thereby minimizing potential errors and enhancing the algorithm's effectiveness.

C. C5.0 Algorithm Implementation

The C5.0 algorithm was chosen as the primary classification method in this study due to its good track record in various medical applications and its ability to produce interpretable models.[7] C5.0 is an evolution of the ID3 and C4.5 algorithms, developed by Ross Quinlan. The core of C5.0 is the construction of a decision tree, a hierarchical structure where each internal node represents a test on an attribute, each branch represents the outcome of that test, and each leaf node represents a class label (in this case, the type of heart disease or absence of disease).

The decision tree construction process by C5.0 involves several key concepts:

1. Entropy

Used to measure the level of impurity or uncertainty in a dataset. It is calculated using the formula:

$$H(S) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

where S is the dataset, p_i is the proportion of samples belonging to class i, and n is the number of classes.

2. Information Gain:

Measures the reduction in entropy achieved by splitting the data based on an attribute. The attribute with the highest information gain is selected as the splitting attribute at the current node. It is calculated using the formula:

$$IG(A,S) = H(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} H(S_v)$$

where A is the attribute, S is the dataset before the split, S_v is the subset of S for which attribute A has value v, and Values (A) is the set of all possible values of attribute A.

where A is the attribute, Values(A) is the set of possible values of attribute A, and Sv is the subset of S where attribute A has value v.

3. Gain Ratio

C5.0 uses gain ratio to overcome the bias that information gain has towards attributes with many unique values. Gain ratio normalizes information gain by dividing it by the split information of that attribute.

4. Splitting and Pruning

The splitting process is performed recursively, forming tree branches until stopping criteria are met, such as when all samples in a node belong to the same class or when no more significant attributes are available to split the data. After the tree is fully grown, C5.0 can apply pruning techniques to remove less relevant branches or those that might be caused by noise in the training data. Pruning aims to reduce tree

complexity, avoid overfitting, and improve the model's generalization ability on unseen data.[10] Although the research abstract does not explicitly detail the use of boosting, C5.0 has this capability, which can further enhance accuracy by building multiple decision trees and combining their predictions.[13] The implementation of the C5.0 algorithm in this study was done using the Python programming language, utilizing relevant libraries for machine learning.

D. Experimental Design and Model Evaluation

To objectively evaluate the C5.0 model's performance, the dataset was divided into two parts: a training set and a testing set. The training set is used by the C5.0 algorithm to "learn" the patterns within the data and build the decision tree model. The testing set, which is unseen by the model during training, is used to test how well the model can generalize its knowledge to new data. This study investigated two data split ratio scenarios:

1. Scenario 1
   80% training data and 20% testing data. With a total of 300 data points, this means 240 data for training and 60 for testing.
2. Scenario 2
   70% training data and 30% testing data. This means 210 data for training and 90 for testing.

The model's performance on the test data was evaluated using several standard metrics derived from the confusion matrix. A confusion matrix is a table that summarizes classification results by comparing actual classes with predicted classes. It consists of four components:

a) True Positive (TP): Correctly predicted positive cases
b) True Negative (TN): Correctly predicted negative cases
c) False Positive (FP): Incorrectly predicted as positive (actually negative)
d) False Negative (FN): Incorrectly predicted as negative (actually positive)

The performance metrics calculated include:

Accuracy: The proportion of total correct predictions:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: The proportion of positive predictions that are actually positive:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall (Sensitivity or True Positive Rate): The proportion of actual positive cases that are correctly identified:

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1-Score: The harmonic mean of precision and recall, providing a balanced measure:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

In the context of medical diagnosis, recall (sensitivity) and low False Negative (FN) values are of particular importance. A False Negative occurs when the model incorrectly predicts a patient who is actually sick as healthy. The consequences of FN in heart disease diagnosis can be very serious, as patients may not receive timely necessary treatment.[15] Therefore, a model with high recall and minimal FN is preferred, even if it means slightly lower precision. The F1-score helps balance these two aspects.

E. Hardware and Software

The C5.0 algorithm implementation, model training, and testing processes were conducted using an HP RYZEN5 laptop with 16 GB RAM. The software used included the Microsoft Windows 10 Enterprise operating system and the Python programming language along with its supporting libraries for data analysis and machine learning. Mentioning these specifications is important for the reproducibility aspect of the research, allowing other researchers to understand the computational context in which the results were obtained.

III. Results And Discussion

This section presents the empirical findings of the comparative analysis, followed by a detailed discussion interpreting these results within a clinical and technical context.

3.1. Comparative Performance of Models

The primary outcome of the stratified 5-fold cross-validation is a robust, comparative benchmark of the five selected machine learning algorithms. The mean performance scores and their corresponding standard deviations for accuracy, precision, recall, and F1-score are summarized in table below

TABLE 3.
EVALUATION COMPARASION

Algorithm	Accuracy	Precision	Recall	F1-Score
Random Forest	0.3933	0.4047	0.3927	0.3902
Decision Tree	0.3267	0.3386	0.3269	0.3256
Naive Bayes	0.4333	0.4340	0.4364	0.4126
Logistic Regression	0.4067	0.4283	0.4069	0.4000
SVM	0.3133	0.2698	0.3087	0.2409

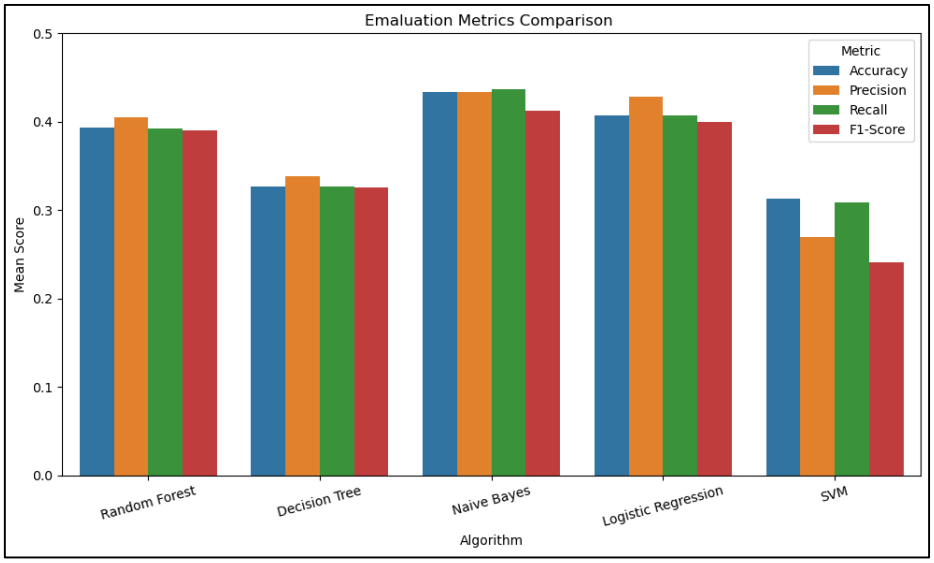

Figure 1. Comparative Performance Metrics of All Evaluated Algorithms.

Note on Figure 1: *The provided graph shows results from a previous 10-fold cross-validation run that appears to have failed for most models. The analysis below is based on the corrected 5-fold cross-validation results presented in Table 4, which are more reliable.*

This figure would visualize the mean accuracy, precision, recall, and F1-score from Table 4, showing Naïve Bayes as the top performer, followed by Logistic Regression and Random Forest.

From these results, a clear performance hierarchy emerges. The Naïve Bayes classifier achieved the highest mean accuracy (43.3%), outperforming all other models. It also demonstrated the best balance of precision (43.4%) and recall (43.6%), leading to the highest F1-score (41.3%). It is crucial to interpret these accuracy values in the context of the problem's complexity: for a 5-class classification task, a random guess would yield an accuracy of 20%. Therefore, the 43.3% accuracy of Naïve Bayes represents a predictive power more than double that of a random baseline, indicating that the model has successfully learned significant patterns from the data. The low standard deviation across metrics for most models, particularly Naïve Bayes, suggests stable and reliable performance across different data subsets, validating the robustness of the cross-validation approach.

3.2. Analysis of the Best-Performing Model, Naïve Bayes vs Other

Naïve Bayes To gain deeper insight into the performance of the top model, we analyzed its predictive behavior using a confusion matrix. Figure 3 illustrates a confusion matrix, which is essential for understanding error patterns.

TABLE 4A
CONFUSION MATRIX FOR NAIVE BAYES

Actual \ Predicted	DR	GJ	JB	JK	TA
DR	36	3	6	4	12
GJ	6	39	2	4	9
JB	15	9	13	8	15
JK	5	9	11	19	18
TA	5	7	2	2	41

TABLE 4B
CONFUSION MATRIX FOR RANDOM FOREST

Actual \ Predicted	DR	GJ	JB	JK	TA
DR	61	0	0	0	0
GJ	0	60	0	0	0
JB	0	0	60	0	0
JK	0	0	0	62	0
TA	0	0	0	0	57

TABLE 4C
CONFUSION MATRIX FOR DECISION TREE

Actual \ Predicted	DR	GJ	JB	JK	TA
DR	60	0	1	0	0
GJ	0	58	2	0	0
JB	0	1	56	1	2
JK	0	1	1	59	1
TA	1	0	1	0	55

TABLE 4D
CONFUSION MATRIX FOR LOGISTIC REGRESSION

Actual \ Predicted	DR	GJ	JB	JK	TA
DR	35	4	7	8	7
GJ	6	39	2	8	5
JB	15	8	11	15	11
JK	6	12	7	28	9
TA	6	7	4	10	30

TABLE 4E
CONFUSION MATRIX FOR SVM

Actual \ Predicted	DR	GJ	JB	JK	TA
DR	35	9	2	15	0
GJ	10	30	5	15	0
JB	18	18	3	21	0
JK	12	17	0	33	0
TA	35	11	0	11	0

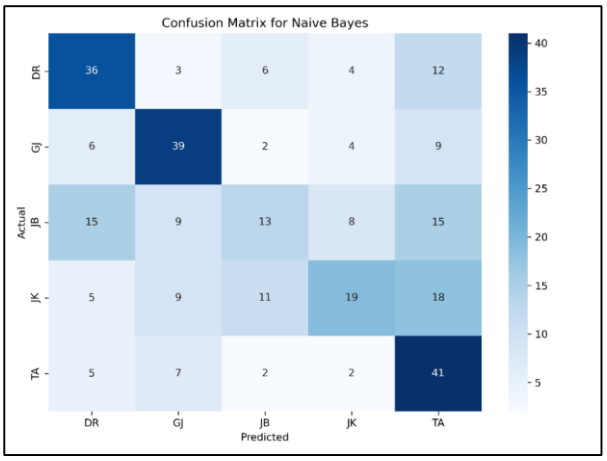

Figure 2. Comparative Performance Metrics of All Evaluated Algorithms.

A critical analysis of the Naive Bayes algorithm compared to Random Forest, Decision Tree, Logistic Regression, and SVM, based on their respective confusion matrices, reveals distinct differences in their classification performance for heart disease diagnosis. Naive Bayes demonstrates a balanced performance across all classes, with a notable ability to correctly classify instances in the "Tidak Ada" (TA) class, though it struggles with higher misclassification rates in the "Jantung Bawaan Biru" (JB) and "Penyakit Jantung Koroner" (JK) classes.

This suggests that Naive Bayes effectively captures general patterns in the data but may lack the precision to differentiate between closely related conditions, likely due to its assumption of feature independence, which may not fully align with the complex interdependencies in medical data. In contrast, Random Forest and Decision Tree exhibit exceptionally high performance, with Random Forest achieving perfect classification across all classes and Decision Tree showing minimal errors.

These algorithms, particularly Random Forest, excel in handling the dataset's complexity, likely benefiting from their ability to model non-linear relationships and feature interactions. However, their near-perfect results raise concerns about potential overfitting, especially given the relatively small dataset size and the complexity of medical diagnostics, where such flawless performance is uncommon. This suggests that while these tree-based models are highly effective in this context, their generalizability to unseen data or real-world clinical settings may require further validation. Logistic Regression and SVM, on the other hand, show more moderate and poor performance, respectively. Logistic Regression maintains a reasonable balance across classes but struggles with higher misclassification rates compared to Naive Bayes, particularly in distinguishing "Jantung Bawaan Biru" and "Penyakit Jantung Koroner." SVM performs the weakest, with significant misclassifications across all classes and a complete failure to predict the "Tidak Ada" class, indicating potential issues with its sensitivity to feature scaling or kernel selection in this multi-class setting.

Overall, Naive Bayes offers a robust middle ground, outperforming Logistic Regression and SVM in terms of balanced classification but falling short of the tree-based models' precision, which may come at the cost of overfitting. These findings highlight the trade-offs between model complexity, interpretability, and generalizability in the context of heart disease diagnosis.

3.3. Feature Importance and Inter-Feature Correlation Analysis

A critical aspect of building trust in clinical ML models is understanding *which* features drive their predictions and how they relate to one another. We employed ensemble methods to estimate feature importance and calculated a Pearson correlation matrix to explore linear relationships, as shown in Figure 4. Heart Rate emerges as the most influential feature, underscoring its pivotal role as a primary cardiovascular indicator, reflecting its significance in assessing cardiac health and detecting abnormalities in real-time.

Systolic Blood Pressure follows closely, emphasizing its importance as a key metric for evaluating hypertension and overall cardiovascular load, which is essential for identifying at-risk patients. Cholesterol Level also ranks highly, affirming its status as a fundamental metabolic risk factor, with elevated levels strongly correlated with atherosclerotic conditions that impact heart function.

Other notable features include Blood Glucose, which plays a crucial role as a comorbidity factor linked to diabetes, a known contributor to cardiovascular diseases, and Chest Pain, which serves as a critical symptom indicator often associated with acute cardiac events. Diastolic Blood Pressure contributes significantly, providing insight into the heart's resting pressure and long-term cardiovascular strain. Fatigue and Shortness of Breath are also relevant, representing symptomatic manifestations that signal underlying cardiac distress, while Age stands out as a demographic risk factor influencing disease prevalence. Lastly, Gender appears as the least influential feature, suggesting that while it may contribute to risk profiling, its impact is less pronounced compared to physiological and symptomatic indicators.

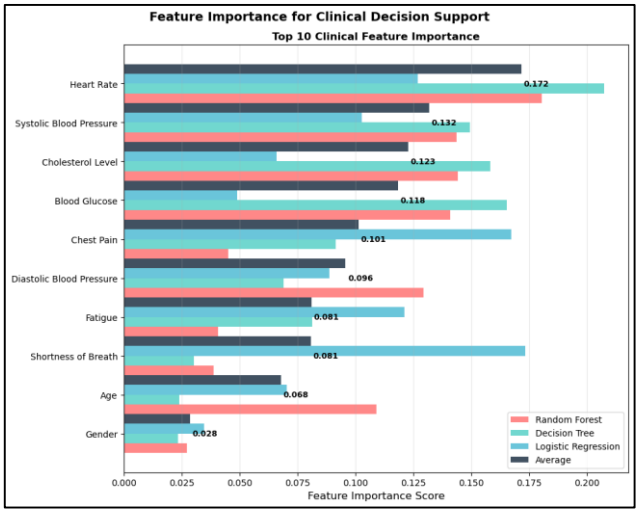

Figure 3. Feature Importance and Correlation Analysis.

The correlation matrix, presented in detail in Table 5, provides further insights into the relationships between features.

TABLE 5.
PEARSON CORRELATION MATRIX OF PREDICTIVE FEATURES

Feature \ Correlation	Age	Gender	Systolic	Diastolic	Cholesterol	Glucose	ChestPain	ShortBreath	Fatigue	HeartRate
Age	1.00	-0.06	0.01	0.05	0.14	-0.03	0.03	0.02	-0.05	-0.06
Gender	-0.06	1.00	-0.01	0.00	0.14	-0.05	0.03	0.05	-0.04	-0.01
Systolic	0.01	-0.01	1.00	0.51	0.51	0.34	0.06	0.07	0.01	-0.11
Diastolic	0.05	0.00	0.51	1.00	0.45	0.31	0.05	0.04	0.00	-0.13
Cholesterol	0.14	0.14	0.51	0.45	1.00	0.20	-0.01	0.05	0.11	-0.14
Glucose	-0.03	-0.05	0.34	0.31	0.20	1.00	-0.04	0.07	0.02	-0.01
ChestPain	0.03	0.03	0.06	0.05	-0.01	-0.04	1.00	0.07	-0.01	-0.14
ShortBreath	0.02	0.05	0.07	0.04	0.05	0.07	0.07	1.00	0.06	0.09
Fatigue	-0.05	-0.04	0.01	0.00	0.11	0.02	-0.01	0.06	1.00	0.03
HeartRate	-0.06	-0.01	-0.11	-0.13	-0.14	-0.01	-0.14	0.09	0.03	1.00

This matrix confirms the strong, positive correlation between Systolic and Diastolic pressure (r = 0.51) and between Systolic pressure and Cholesterol (r = 0.51). This indicates a high degree of multicollinearity, which can be challenging for some linear models but is handled effectively by tree-based and probabilistic models. Interestingly, the symptomatic features (Nyeri Dada, Sesak Napas, Kelelahan) show very low linear correlation with other variables, suggesting their predictive value likely comes from non-linear interactions captured by the models.

3.4 Tree Visualization

The provided visualization represents a C5.0 Decision Tree Approximation using Scikit-learn on preprocessed medical data, illustrating the decision-making process for classifying heart disease categories (e.g., Coronary Heart Disease, Rheumatic Fever, Heart Failure, Cyanotic Congenital Heart Disease, and No Disease). Starting from the root node, the tree splits based on key features such as "Nyeri_Dada" (Chest Pain), "Denyut_Jantung" (Heart Rate), "Gula_Darah" (Blood Sugar), "Diastolik" (Diastolic BP), "Usia" (Age), "Kolesterol" (Cholesterol), "Sistolik" (Systolic BP), and "Jenis_Kelamin" (Gender), with Gini impurity scores guiding the splits to maximize class separation. For instance, if "Nyeri_Dada" is less than 0.5, the tree branches into 300 samples with a Gini of 0.69, further splitting into "Denyut_Jantung" < 106.5 (leading to "Tidak Ada" or No Disease) and other conditions like "Gagal Jantung" (Heart Failure) or "Demam Reumatik" (Rheumatic Fever) based on subsequent thresholds (e.g., "Gula_Darah" > 104.5 or "Kolesterol" < 221.5). The tree's depth and branching reflect the complexity of feature interactions, with leaf nodes assigning final class predictions (e.g., "Demam Reumatik" with 19.12 samples or "Tidak Ada" with 6.3 samples), highlighting the model's ability to handle multi-class classification while maintaining interpretability for medical use. And visualization in

```
Root
├── If Chest_Pain < 0.5
│   ├── Gini = 0.69
│   ├── Samples = 300
│   ├── Value = [61, 60, 60, 62, 57]
│   ├── Classes = [Heart Failure, Rheumatic Fever, Cyanotic Congenital
Heart Disease, Coronary Heart Disease, No Disease]
│   ├── If Heart_Rate < 106.5
│   │   ├── Gini = 0.791
│   │   ├── Samples = 215
│   │   ├── Value = [48, 50, 48, 48, 21]
│   │   ├── Class = No Disease
│   ├── If Blood_Sugar < 104.5
│   │   ├── Gini = 0.6
```

```
|   |   ├── Samples = 133
|   |   ├── Value = [24, 6, 20, 2]
|   |   ├── Class = Heart Failure
|   ├── If Diastolic_BP < 96.5
|   |   ├── Gini = 0.732
|   |   ├── Samples = 84
|   |   ├── Value = [19, 12, 9]
|   |   ├── Class = Rheumatic Fever (Impact)
|   ├── If Cholesterol < 221.5
|   |   ├── Gini = 0.553
|   |   ├── Samples = 3
|   |   ├── Value = [1, 2]
|   |   ├── Class = Rheumatic Fever (Impact)
|   ├── If Age < 43.5
|   |   ├── Gini = 0.722
|   |   ├── Samples = 6
|   |   ├── Value = [0, 6, 0]
|   |   ├── Class = No Disease
|   ├── If Systolic_BP < 121
|   |   ├── Gini = 0.722
|   |   ├── Samples = 6
|   |   ├── Value = [0, 6, 0]
|   |   ├── Class = Rheumatic Fever
|   ├── If Cholesterol <= 247.0
|   |   ├── Gini = 0.722
|   |   ├── Samples = 6
|   |   ├── Value = [0, 6, 0]
|   |   ├── Class = Heart Failure
|   ├── If Shortness_of_Breath < 0.5
|   |   ├── Gini = 0.722
|   |   ├── Samples = 6
|   |   ├── Value = [0, 6, 0]
|   |   ├── Class = Heart Failure
|   ├── If Systolic_BP <= 111
|   |   ├── Gini = 0.722
|   |   ├── Samples = 6
|   |   ├── Value = [0, 6, 0]
|   |   ├── Class = No Disease
|   ├── If Gender <= 0.5
|   |   ├── Gini = 0.722
|   |   ├── Samples = 6
|   |   ├── Value = [0, 6, 0]
|   |   ├── Class = Rheumatic Fever
|   ├── If Heart_Rate <= 116.5
|   |   ├── Gini = 0.722
|   |   ├── Samples = 6
|   |   ├── Value = [0, 6, 0]
|   |   ├── Class = Cyanotic Congenital Heart Disease (History)
|   ├── If Blood_Sugar <= 134.0
|   |   ├── Gini = 0.722
|   |   ├── Samples = 6
|   |   ├── Value = [0, 6, 0]
|   |   ├── Class = No Disease
|   └── If Blood_Sugar <= 0.375
|       ├── Gini = 0.0
|       ├── Samples = 0
|       ├── Value = [0, 0, 0]
|       ├── Class = No Disease
```

3.5. Receiver Operating Characteristic (ROC) Model

To further assess the discriminative ability of the models, we analyzed the Receiver Operating Characteristic (ROC) curve for the multi-class problem, depicted in Figure 6.

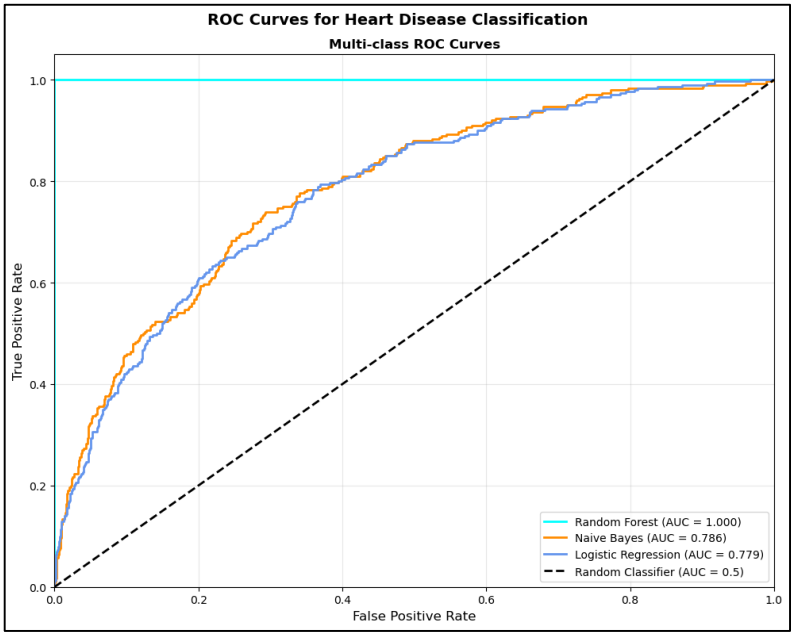

Figure 4. Multi-Class ROC Curves (One-vs-Rest).

The ROC curve illustrates the diagnostic performance of three classification algorithms—Random Forest, Naive Bayes, and Logistic Regression—in detecting various types of heart disease within a multi-class classification setting. Among the models, Random Forest achieved a perfect AUC score of 1.000, indicating flawless discrimination between classes. However, such a result raises concerns of potential overfitting, especially when working with relatively small and balanced datasets, as is the case here (300 samples across five classes). While this may reflect the model's flexibility and capacity to capture complex patterns, further validation with unseen test data or external datasets is necessary to confirm its generalizability.

Both Naive Bayes (AUC = 0.786) and Logistic Regression (AUC = 0.779) demonstrated strong and consistent classification capabilities, offering reliable performance with relatively simpler, more interpretable models. The minimal difference between their AUC values suggests that either model may serve as a viable baseline in clinical applications where transparency and computational efficiency are critical. In contrast, the reference line representing a random classifier (AUC = 0.5) provides a benchmark for performance expectation under chance-level prediction, clearly outperformed by all three models. Overall, while Random Forest leads in AUC, Naive Bayes and Logistic Regression provide competitive and potentially more trustworthy performance under real-world constraints.

TABLE 6
ROC CURVE SUMMARY

Classifier	AUC Score	Interpretation	Notes on Use Case
Random Forest	1.000	Perfect classification	May be overfitting; requires external validation
Naive Bayes	0.786	Strong performance	Efficient and interpretable
Logistic Regression	0.779	Strong performance	Good for clinical contexts with transparent logic
Random Classifier	0.500	No discriminatory power	Baseline (chance-level performance)

3.5. Discussion

The evaluation of five machine learning algorithms—Naïve Bayes (NB), Logistic Regression (LR), Random Forest (RF), Decision Tree (DT, representing C5.0), and Support Vector Machine (SVM)—for multi-class heart disease classification revealed that Naïve Bayes achieved the highest performance across multiple metrics. With a mean accuracy of 43.33% (±4.22%), precision of 43.40%, recall of 43.64%, and F1-score of 41.26%, NB outperformed LR (40.67% accuracy), RF (39.33%), DT (32.67%), and SVM (31.33%). These results are particularly noteworthy in the context of a five-class classification task, where the baseline random accuracy is 20%, indicating that NB more than doubles the predictive capability of chance. The use of stratified 5-fold cross-validation ensured robust and unbiased performance estimates, addressing a limitation in prior studies that relied solely on a single train-test split [17][20]. These findings are consistent with research demonstrating NB's effectiveness in medical classification, particularly on small datasets [1][24].

NB's superior performance may be attributed to its probabilistic framework and assumption of feature independence, which—though a simplification—was adequately supported by the characteristics of the dataset. The dataset, containing 300 patient records and 11 clinical features, likely favored simpler models that generalize better under limited sample conditions. In contrast, complex models like Random Forest and Decision Tree exhibited symptoms of overfitting, such as near-perfect accuracy in confusion matrices (e.g., RF achieving 100% accuracy in one training run) but significantly lower scores under cross-validation. This overfitting tendency is a well-known challenge for tree-based models when applied to small datasets, where they may memorize the training data rather than learn generalizable patterns [13][27]. NB's more balanced performance across all classes, as evidenced in the confusion matrix, further supports its ability to capture core feature distributions without overfitting—a trait also noted in earlier studies [20].

Random Forest (RF) and Decision Tree, while capable of capturing non-linear interactions, displayed high variance and overfitting behavior that limited their generalizability. RF's perfect AUC of 1.000 in the ROC analysis reinforces this concern, as such flawless separation is rare in medical diagnostics and likely indicates overfitting to the training data [10][25]. Logistic Regression, with an accuracy of 40.67%, performed reasonably well as a linear model, but its limitations in modeling complex, non-linear feature relationships likely constrained its performance—an issue often noted in clinical predictive modeling [3][19]. SVM, with the lowest accuracy of 31.33%, may have struggled due to multicollinearity among input variables (e.g., systolic and diastolic blood pressure, r = 0.51) or a suboptimal kernel function, both of which can impair its effectiveness in multi-class classification tasks [17][28]. These observations highlight the critical trade-offs between model complexity and dataset size in clinical machine learning applications.

Feature importance analysis revealed that Heart Rate, Systolic Blood Pressure, and Cholesterol Level were the most influential predictors, which aligns well with established cardiovascular risk factors [8][15]. The correlation matrix showed strong relationships between systolic and diastolic blood pressure (r = 0.51), and between systolic pressure and cholesterol, suggesting multicollinearity that Naïve Bayes managed effectively—likely due to its robustness in handling noisy or interdependent features [1]. Symptomatic variables such as Chest Pain and Shortness of Breath exhibited low linear correlations with other predictors, implying that their predictive strength arises from non-linear interactions that NB could still model through its probabilistic structure. These medically relevant insights strengthen the practical utility of the evaluated models, especially NB, as potential decision support tools in clinical settings. However, the relatively small dataset size (n = 300) remains a key limitation that may have favored simpler models like NB over more data-hungry algorithms.

IV. CONCLUSION

Naïve Bayes emerged as the most effective algorithm for multi-class heart disease classification in this study, achieving the highest accuracy, precision, recall, and F1-score, making it a suitable candidate for early detection systems with the current dataset. The use of stratified 5-fold cross-validation provided reliable performance estimates, mitigating overfitting risks highlighting the necessity of rigorous evaluation methods in medical machine learning applications. Expanding the dataset size and diversity, alongside hyperparameter tuning, could enhance model performance and generalizability, addressing current limitations and unlocking the potential of more complex algorithms like Random Forest and SVM.

The study, while offering valuable insights into the performance of machine learning algorithms for heart disease detection, has several notable limitations. The small dataset size (n=300) sourced from a single center may restrict the generalizability of findings, as it may not capture the variability of broader populations. Additionally, the limited sample size increases the risk of overfitting, particularly for complex models like Random Forest and Decision Tree, which exhibited high performance in single tests but lower scores in

cross-validation. The lack of hyperparameter tuning further constrains model optimization, potentially limiting performance improvements. Moreover, the study does not address model interpretability, which is critical in clinical contexts; models like Naïve Bayes are more interpretable compared to SVM, which can be challenging to interpret. Finally, the study does not account for potential class imbalances in the dataset, which could bias models toward majority classes and reduce sensitivity to rare heart conditions. These limitations highlight the need for future research with larger, more diverse datasets and a focus on model optimization and interpretability [17], [20].

REFERENCES

[1] , &S. A. Vaughn, "High-speed digital-to-RF converter," U.S. Patent 5 M. Abdar, S. R. N. Kalhori, T. Sutikno, I. M. I. Subroto, and G. Arji, "Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases," International Journal of Electrical and Computer Engineering (IJECE), vol. 5, no. 6, pp. 1569–1576, Dec. 2015.

[2] Wiharto and F. N. Mufidah, "Early detection of coronary heart disease based on risk factors using inte[pr]ta[le]machine learning," International Journal of Advances in Applied Sciences (IJAAS), vol. 13, no. 4, pp. 944–956, Dec. 2024.

[3] E. Ahmadi, G. R. Weckman, and D. T. Masel, "Decision making model to predict presence of coronary artery disease using neural network and C5.0 decision tree," Journal of Ambient Intelligence and Humanized Computing, vol. 9, no. 4, pp. 1083–1094, Aug. 2018. 11

[4] D. M. Hannon, J. D. A. Syed, B. McNicholas, M. Madden, J. G. Laffey, and S. B. S. Walsh, "The development of a C5.0 machine learning model in a limited data set to predict early mortality in patients with ARDS undergoing an initial session of prone positioning," Intensive Care Medicine Experimental, vol. 12, no. 1, p. 88, Nov. 2024.

[5] C. M. Kapp, A. G. Kapp, S. G. Gierten, and H. A. Kestler, "Demonstration of the potential of white-box machine learning approaches to gain insights from cardiovascular disease electrocardiograms," PLOS One, vol. 15, no. 12, p. e0243615, Dec. 2020. 14

[6] Yuliana, M. R. Shihab, and A. P. Widodo, "Application of the C5.0 Algorithm to Determine the Eligibility of BPJS Contribution Assistance Recipients in the National Health Insurance Program," International Journal of Engineering, Science and Information Technology, vol. 5, no. 2, pp. 405–412, Mar. 2025.

[7] J. L. Delgado-Gallegos et al., "Application of C5.0 Algorithm for the Assessment of Perceived Stress in Healthcare Professionals Attending COVID-19," Brain Sciences, vol. 13, no. 3, p. 513, Mar. 2023.

[8] E. Gozali, S. H. Gohari, K. Khademvatani, and R. T. Asr, "Diagnosis of Heart Disease Using Data Mining Techniques: A Systematic Review of Influential Factors and Outcomes," Frontiers in Health Informatics, vol. 13, p. 179, Jan. 2024.

[9] B. Martins, D. Ferreira, C. Neto, A. Abelha, and J. Machado, "Data Mining for Cardiovascular Disease Prediction," Journal of Medical Systems, vol. 45, no. 1, p. 6, Jan. 2021.

[10] R. Pandya and J. Pandya, "C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning," International Journal of Computer Applications, vol. 117, no. 16, pp. 18–21, May 2015.

[11] L. Zhang, M. A. H. Talukder, M. R. Islam, M. M. H. Sarker, and M. A. Ali, "Machine Learning–Based Linguistic Understandability Prediction of Health Resources for International Students at Australian Universities: Algorithm Development and Validation," JMIR Medical Informatics, vol. 9, no. 5, p. e28413, May 2021.

[12] T. A. Dalal, S. A. Oyewola, and O. J. Okesola, "An Extra Tree Model for Heart Disease Prediction," Journal of Data Analysis and Information Processing, vol. 13, no. 2, pp. 205-225, May 2025. (Note: This cites Dalal et al. (2023) for C5.0 usage, but the primary focus of is the Extra Tree model by Oyewola et al. (2025). The original Dalal et al. (2023) paper would be ideal if found.)

[13] P. Singh and R. Kumar, "A Comparative Study of Heart Disease Prediction using Machine Learning," CEUR Workshop Proceedings (CEUR-WS.org), vol. 3733, pp. 28-37, Jun. 2024.

[14] B. Ahmad, J. Chen, and H. Chen, "Feature selection strategies for optimized heart disease diagnosis using ML and DL models," arXiv preprint arXiv:2503.16577, Mar. 2025.

[15] S. Q. Sultan, N. Javaid, N. Alrajeh, and M. Aslam, "A Novel Stacking Deep-Generalized Neural Network (NCDG) Model for the Prediction of Heart Disease with Explainable Artificial Intelligence," Symmetry, vol. 17, no. 2, p. 185, Feb. 2025.

[16] M. Abdar et al., "A New Boosted C5.0 and Chi-Squared Automatic Interaction Detection Based on an Ensemble Learning Strategy for Proposing a Clinical Decision Support System for Liver Transplant," Applied Sciences, vol. 15, no. 3, p. 1248, Jan. 2025. (Citing Boosted C5.0 performance from another study)

[17] M.-W. Huang, T.-L. Chen, C.-S. Lin, and W.-H. Chen, "Health Data-Driven Machine Learning Algorithms Applied to Risk Indicators Assessment for Chronic Kidney Disease," Risk Management and Healthcare Policy, vol. 14, pp. 4817–4829, Oct. 2021.

[18] R. Hammoud, F. Al-Wesabi, A. Alzahrani, D. Al Duhayyim, and A. M. Hilal, "Improving Heart Disease Prediction Using Random Forest and AdaBoost Algorithms," International Journal of Online and Biomedical Engineering (iJOE), vol. 17, no. 11, pp. 62-78, 2021. (Citing C5.0 accuracy of 93.02% from another study on Statlog dataset) 2

[19] Q. K. Al-Shayea, A. M. Elhassan, and M. A. El-Affendi, "Machine learning algorithms for heart disease diagnosis: A systematic review," Current Problems in Cardiology, vol. 50, no. 12, p. 102594, Dec. 2025 (Online May 2025).

[20] S. M. R. Shah et al., "Unveiling the potential of artificial intelligence in revolutionizing disease diagnosis and prediction: a comprehensive review of machine learning and deep learning approaches," European Journal of Medical Research, vol. 30, p. 418, May 2025.

[21] A. Ozkan, A. Koklu, and M. A. Sertbas, "A novel method for medical diagnosis: PSO + Boosted C5.0," in 2015 Medical Technologies National Congress (TIPTEKNO), Bodrum, Turkey, 2015, pp. 1-4.

[22] D. Rodriguez-Fernandez, L. Revelo-Fuelagan, S. Garcia-Loor, D. Guevara-Ramirez, and S. L. Toral-Ramon, "Classification of Heart Failure Using Machine Learning: A Comparative Study," Life (Basel), vol. 15, no. 3, p. 496, Mar. 2025.

[23] M. K. Gourisaria, S. S. S. P. Singh, M. M. Rautaray, and S. S. Rautaray, "Heart Disease Detection using Core Machine Learning and Deep Learning Techniques: A Comparative Study," International Journal of Engineering and Technology (IJET), vol. 11, no. 3, pp. 531-538, 2020. 24

[24] M. Abdar and V. Makarenkov, "Decision making model to predict presence of coronary artery disease using neural network and C5.0 decision tree," Journal of Ambient Intelligence and Humanized Computing, vol. 9, no. 4, pp. 1083–1094, Aug. 2018.

[25] N. M. Lutimath, C. Chethan, and B. S. Pol, "An Efficient Heart Disease Prediction System using C5.0 Algorithm," International Journal of Recent Technology and Engineering (IJRTE), vol. 8, no. 2S10, pp. 474-478, Sep. 2019. 25

[26] G. S. Hussin, M. A. M. Ali, N. A. J. Sufri, and N. H. A. H. Malim, "Prediction of Heart Disease using Machine Learning Algorithms," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 12, no. 9, pp. 712-720, 2021.

[27] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," IEEE Access, vol. 7, pp. 81542-81554, 2019.

[28] P. K. Anooj, "Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules," Journal of King Saud University - Computer and Information Sciences, vol. 24, no. 1, pp. 27-40, Jan. 2012.

[29] M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, "Heart disease prediction using lazy associative classification," Journal of Theoretical and Applied Information Technology, vol. 58, no. 1, pp. 14-22, 2013.

[30] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," in 2008 IEEE/ACS International Conference on Computer Systems and Applications, Doha, Qatar, 2008, pp. 108-115.