Improving Cardiovascular Disease Prediction Accuracy with Three-Way Decisions

Arshad Aziz[1], Hasham Khan[1], M. Attaullah[2], M.K. Afridi[1]
[1]Dept. of Computing & Tech. Abasyn Univ. Peshawar Peshawar, Pakistan
[2]Dept. of Computing & Tech. UET Lahore, Lahore, Pakistan
*__Correspondence__: gullsher210@gmail.com, khanhasham554@gmail.com, muhammadattaullah.pk@gmail.com, mohammad@abasyn.edu.pk

Citation | Aziz. A, Khan. H, Attaullah. M, Afridi. M. K, "Improving Cardiovascular Disease Prediction Accuracy with Three-Way Decisions", IJIST, Vol. 07 Special Issue. pp 127-142, May 2025 DOI:https://doi.org/10.33411/IJIST/1308
Received | April 0, 2025 **Revised |** May 01, 2025 **Accepted |** May 03, 2025 **Published |** May 09, 2025.

Cardiovascular Disease (CVD) is a leading cause of death worldwide, making accurate and early risk prediction crucial for better patient outcomes. Traditional CVD prediction models often rely on binary decision-making, which struggles with uncertain or borderline cases, leading to misclassification and ineffective treatment strategies. This research proposes an advanced predictive model that combines machine learning algorithms with a three-way decision approach to improve the accuracy and reliability of CVD risk assessment. The three-way decision model, based on rough set theory, divides decisions into three categories acceptance, rejection, and deferment—allowing for more detailed and informed predictions. Using the Cleveland Heart Disease dataset, this study applies machine learning techniques such as Random Forest (97.14% accuracy), Logistic Regression (91.30% accuracy), Naïve Bayes (88.24% accuracy), and Support Vector Machine (89.74% accuracy) to evaluate the model's effectiveness. The results show that integrating three-way decisions with machine learning improves predictive performance, especially for unclear cases, enhancing clinical decision-making. However, the model's reliance on dataset quality and threshold selection poses some limitations that need further investigation. This research introduces an intelligent and flexible approach to CVD prediction, which could reduce diagnostic errors and support early interventions for high-risk patients.

Keywords: Cardiovascular Disease Prediction, Machine Learning, Three-Way Decisions, Rough Set Theory, Medical Diagnostics, Predictive Modeling

Introduction:

Cardiovascular Disease (CVD) is one of the most pressing health concerns worldwide. According to World Health Organization reports, CVDs—including heart attacks, strokes, and related conditions—cause about 18 million deaths annually, making them the leading cause of death globally [1]. Notably, the causes of CVD-related mortality vary between countries [2]. These differences are mainly due to regional disparities in risk factor prevalence and the availability and quality of healthcare services in recent years [3], [4], [5]. This growing number of cases not only affects individuals and their families but also places significant pressure on healthcare systems. Early and accurate identification of people at risk for CVD is crucial to preventing severe outcomes, reducing treatment costs, and improving public health overall [6]. Traditional CVD risk prediction methods typically rely on binary decision-making, evaluating the likelihood of cardio- vascular events based on specific thresholds, such as cholesterol levels, blood pressure, and other health indicators [7]. However, these models often oversimplify medical diagnostics, particularly when dealing with uncertain or incomplete patient data. As a result, binary predictions can misclassify individuals with borderline or atypical risk factors, leading to inaccurate diagnoses and insufficient preventive measures [4].

Recent advancements in Machine Learning (ML) offer promising solutions to improve CVD prediction. ML algorithms can process large and diverse datasets, uncovering complex patterns that traditional methods might miss [8]. By leveraging historical data, ML models can adapt to intricate, nonlinear relationships within the data. Techniques such as Random Forest, Logistic Regression, Support Vector Machine (SVM), and Naïve Bayes have shown potential for accurate CVD prediction [9]. However, most ML models still rely on binary classification and lack mechanisms to address ambiguous cases.

This paper is organized as follows: Section II provides a detailed description of the Cleveland Heart Disease dataset and exploratory data analysis. Section III reviews related literature on CVD prediction. Section IV outlines the proposed methodology, describing the integration of machine learning algorithms with the Three-Way Decision model, including classifier selection, thresholding, and model architecture. Section V presents the experimental setup, performance evaluation metrics, and results from applying various algorithms. Finally, Section VI discusses the findings, highlights the advantages and limitations of the proposed approach, and suggests directions for future research.

Objectives:

The primary objective of this study is to enhance the accuracy and reliability of CVD prediction by integrating classical machine learning models with a Three- Way Decision (3WD) framework. Specifically, this research aims to:

1) Develop and evaluate four well-established classifiers Logistic Regression, Naïve Bayes, Support Vector Machine (SVM), and Random Forest on the Cleveland Heart Disease dataset.
2) Incorporate a probability-based three-way classification approach that introduces an additional "deferment" class alongside traditional accept/reject outcomes, allowing the system to abstain from uncertain predictions.
3) Compare the performance of the proposed framework with existing literature to assess improvements in classification metrics, particularly overall accuracy, recall, and F1-score.

Novelty Statement:

The novelty of this study lies in the integration of the 3WD model with standard machine learning classifiers to improve diagnostic precision in heart disease prediction. While prior studies have predominantly employed binary classification models, this research introduces a third decision option deferment which allows the model to withhold predictions when confidence is low, thereby reducing the risk of false positives and false negatives.

Additionally, the study conducts a grid search over and threshold values combined with 10-fold stratified cross-validation, which has not been extensively applied in earlier CVD prediction frameworks. This approach not only optimizes model performance but also aligns with clinical decision-making practices, where uncertain cases are referred for further evaluation rather than forced into premature diagnostic categories. Thus, the proposed framework enhances both predictive robustness and decision-making reliability in healthcare settings.

Literature Review:

CVD is a leading global cause of mortality, highlighting the urgent need for effective predictive models to identify high-risk individuals early. Traditional clinical models for CVD prediction typically rely on binary decision systems, which may not fully capture the complexity and variability of patient data. As a result, ML techniques have emerged as promising alternatives, offering robust, flexible, and accurate models that can adapt to diverse patient populations and complex data patterns. This literature review explores recent advancements in CVD prediction using machine learning, from traditional statistical methods to hybrid and ensemble ML approaches [10]. A study [11], introduced a hybrid model combining Logistic Regression with Support Vector Machines (SVM) and Decision Trees to enhance heart disease prediction accuracy. While traditional approaches often use individual algorithms based on preset rules, these methods struggle to detect complex, nonlinear patterns in patient data.

This hybrid model integrated multiple algorithms, boosting prediction accuracy to 88.7% by leveraging the strengths of each model. The combination of Decision Trees, SVM, and Logistic Regression was strategic—Decision Trees are effective for managing categorical variables, SVM excels at finding optimal decision boundaries, and Logistic Regression offers robust linear separation. By merging these methods, the study balanced interpretability and predictive power, providing clinicians with a more reliable tool for early CVD detection. Future work suggested integrating real-time data, allowing the hybrid system to dynamically update predictions as patient information evolves, making the model adaptable to clinical settings. In 2023, a study [12], presented at an IEEE conference tackled the limitations of traditional diagnostic methods, particularly their inability to manage ambiguous or borderline cases.

This study focused on an ensemble model that combined multiple machine learning approaches to improve CVD diagnosis re- liability. By utilizing techniques like boosting and bagging, the ensemble model reduced overfitting and improved generalization across diverse datasets, achieving an accuracy of 91%. Ensemble models aggregate predictions from multiple algorithms, providing a "voting" system that enhances prediction robustness, especially in cases where individual models might disagree. This is particularly useful in healthcare, where variability in patient data can complicate classification. The study emphasized the potential for future improvements by using larger datasets and deep learning techniques, which could capture subtler patterns and improve accuracy [13].

Another 2023 study focused on early-stage CVD detection using advanced ML techniques such as Random Forest and Gradient Boosting. These ensemble methods, known for their high accuracy and ability to manage both structured and unstructured data, achieved an accuracy of 92%, showcasing their potential to enhance robustness in CVD detection. Random Forest, based on decision trees, resists overfitting and handles large feature spaces well, which is critical for CVD prediction that involves a wide array of physiological and demographic variables. Gradient Boosting, on the other hand, constructs models sequentially to correct the errors of previous models, progressively improving accuracy. This method is particularly effective for imbalanced datasets, where positive cases of CVD are less frequent than negative cases [14]. The study proposed implementing these models in clinical settings for

real-time diagnostics, which could significantly improve patient care by providing immediate risk assessments and allowing for timely interventions.

In a study [15], researchers created a machine learning model that integrated conventional risk factors—such as age, blood pressure, and cholesterol levels—with modern ML techniques, achieving an accuracy of 85%. This approach highlights the effectiveness of combining established risk factors with ML algorithms to enhance predictive accuracy. Common techniques like SVM, Random Forest, Logistic Regression, and Naïve Bayes classifiers have been successfully applied in CVD prediction. However, while these models handle both linear and nonlinear data, they lack methods to manage uncertainty and intermediate decisions. The three-way decision model, proposed by Yao within rough set theory, addresses this gap by classifying decisions into three categories: acceptance, rejection, and deferment.

The three-way decision model offers a novel framework for integrating machine learning in CVD prediction. By in- cooperating a "deferment" category, the model enables predictions to be postponed when uncertainty is high or data is ambiguous, allowing for additional testing or analysis before making a final decision. This feature is especially useful in medical settings, where misclassification can have serious consequences. Combining this model with existing ML algorithms could improve CVD prediction by adding flexibility in decision-making, particularly in borderline cases, and enhancing interpretability and caution—important aspects of medical diagnostics [16].

Collectively, these studies demonstrate the transformative potential of machine learning in CVD prediction. Hybrid and ensemble techniques, in particular, have proven to offer increased accuracy and robustness over traditional single- algorithm models. By integrating multiple algorithms or com- bining ML with established risk factors, these models over- come many limitations of binary or standalone approaches. Ensemble methods like Gradient Boosting and Random Forest have shown promise in managing diverse patient data, improving diagnostic accuracy. As the field evolves, future research must focus on integrating these models into clinical practice through real-time data processing and exploring the inclusion of more varied datasets. The introduction of the three-way decision model adds a new dimension for handling ambiguity in medical diagnostics, offering more reliable predictions and ultimately improving patient outcomes while alleviating the burden of CVD on healthcare systems worldwide.

Methodology:

Dataset Description and Exploratory Data Analysis:

The Cleveland Heart Disease dataset [17], sourced from Kaggle, is the primary dataset used in this study. It contains 303 instances, with 14 features, where each instance represents a patient, and each attribute corresponds to a specific health- related characteristic. The key features of this dataset are listed below:

Attributes Description:

- **Age**: Age plays a significant role in heart disease risk. As people age, the likelihood of developing cardiovascular issues increases due to factors like arterial stiffening, plaque buildup, and reduced heart function. Older individuals are at higher risk, making age an essential feature for prediction models.
- **Sex**: Gender influences the development and manifestation of heart disease. Men tend to develop heart disease earlier than women, while postmenopausal women experience an increased risk due to hormonal changes. This feature helps models understand gender-based differences in disease occurrence.
- **Chest Pain Type (CP)**: Chest pain is a common symptom of heart disease. This feature categorizes chest pain into four types: typical angina, atypical angina, non-anginal pain, and asymptomatic. Typical angina is often linked to coronary artery disease, while

asymptomatic cases may indicate silent heart disease.

- **Resting Blood Pressure (Trestbps)**: This feature represents a patient's blood pressure in mmHg at rest. High blood pressure is a strong indicator of hypertension, a major risk factor for cardiovascular diseases. Consistent high blood pressure increases the strain on the heart and raises the risk of heart disease.
- **Serum Cholesterol (Chol)**: This feature measures total cholesterol in the blood, in mg/dL. High cholesterol levels, especially elevated LDL (bad cholesterol), can cause plaque buildup in the arteries, increasing the risk of heart disease and heart attacks.
- **Fasting Blood Sugar (Fbs)**: This binary feature indicates whether the patient's fasting blood sugar level exceeds 120 mg/dL. High fasting blood sugar levels are associated with diabetes, a significant risk factor for heart disease, due to insulin resistance and blood vessel damage.
- **Resting Electrocardiographic Results (Restecg)**: Electrocardiograms (ECG) measure the heart's electrical activity. This feature categorizes ECG results into normal, abnormal ST-T wave changes, and left ventricular hyper- trophy. Abnormal readings can suggest heart conditions such as ischemia or arrhythmias.
- **Maximum Heart Rate Achieved (Thalach)**: This feature records the highest heart rate achieved by the patient during physical exertion. A lower heart rate may indicate poor cardiovascular fitness, while a high value suggests a healthy heart response to activity.
- **Exercise-Induced Angina (Exang)**: This binary feature shows whether the patient experiences chest pain during physical activity. A positive result indicates reduced blood flow to the heart, often due to blockages in the coronary arteries.
- **ST Depression Induced by Exercise (Oldpeak)**: ST depression refers to changes in the ST segment of an ECG during exercise, indicating myocardial ischemia (insufficient blood flow to the heart). A higher ST depression value suggests a higher likelihood of coronary artery disease.
- **Slope of the Peak Exercise ST Segment (Slope)**: This feature describes the shape of the ST segment during peak exercise, categorized as upsloping, flat, or downsloping. A downsloping ST segment is linked to a higher risk of ischemic heart disease.
- **Number of Major Vessels Colored by Fluoroscopy (Ca)**: Fluoroscopy is a diagnostic imaging technique used to view blood vessels. This feature counts the number of major coronary arteries (from 0 to 4) showing narrowing due to blockages. More blockages increase the risk of heart disease.
- **Thalassemia (Thal)**: Thalassemia is a genetic blood disorder affecting hemoglobin production. This feature is categorized as normal, fixed defect, or reversible defect. A fixed defect suggests permanent heart damage, while a reversible defect indicates a treatable abnormality.

Table 1. Dataset Attributes and Characteristics

ID	Attribute	Type	Values / Description
1	**Sex/Gender**	Discrete	1 = Male, 0 = Female
2	**Age**	Continuous	Age in years
3	**Cp (Chest Pain)**	Discrete	1 = Typical angina, 2 = Atypical angina, 3 = Non-anginal pain, 4 = Asymptomatic
4	**RestBP** (Resting Blood Pressure)	Continuous	90–200 mmHg
5	**Chol** (Cholesterol Level)	Continuous	126–564 mg/dL
6	**Fbs** (Fasting Blood Sugar)	Discrete	1 = True (FBS > 120 mg/dL), 0 = False
7	**Restecg** (Resting ECG Results)	Discrete	0 = Normal, 1 = ST-T wave abnormality, 2 = Left ventricular hypertrophy
8	**Thalach** (Max Heart Rate)	Continuous	71–202 bpm
9	**Exang** (Exercise-Induced Angina)	Discrete	1 = Yes, 0 = No
10	**Oldpeak ST** (ST Depression)	Continuous	0 to 6.2 (depression induced by exercise relative to rest)
11	**Slope** (Peak Exercise Slope)	Discrete	1 = Upsloping, 2 = Flat, 3 = Downsloping
12	**Ca** (No. of Major Vessels)	Discrete	0 to 3 (number of major vessels colored by fluoroscopy)
13	**Thal** (Thallium Stress Test)	Discrete	3 = Normal, 6 = Fixed defect, 7 = Reversible defect
14	**Target** (Heart Disease Presence)	Discrete	1 = Yes (disease), 0 = No (no disease)

These features form the basis of our heart disease prediction model. Through exploratory data analysis, we examined patterns and relationships among these attributes. The analysis revealed that the age of patients ranges from 29 to 77 years, with an average age of 54.36 years (see Figure 1 for age distribution). This suggests that most patients are in the middle-aged to elderly category.

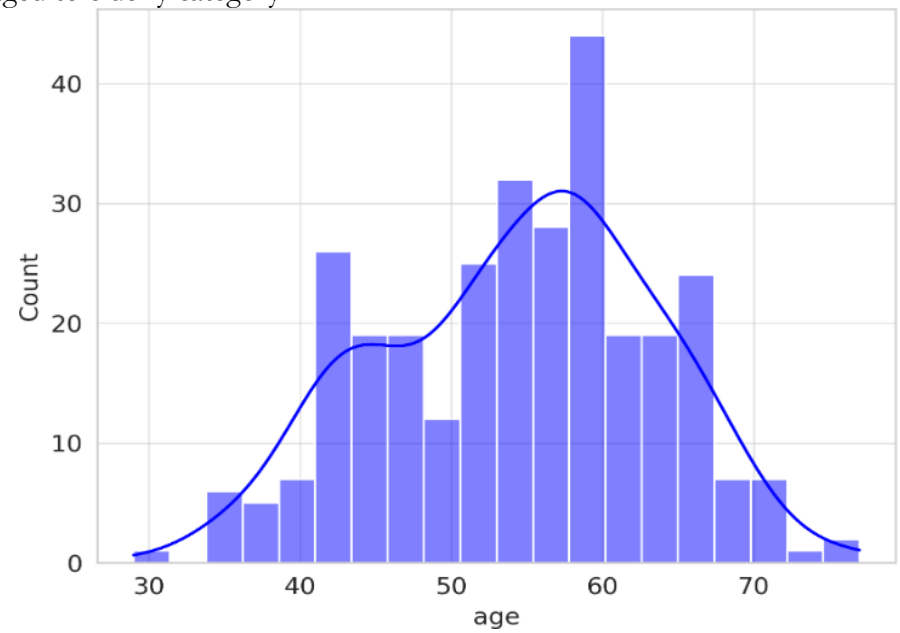


Figure 1. Age Distribution of Patients

The sex distribution, shown in Figure 2, is skewed towards females, with 68.3% of the dataset comprising female patients. This supports the general medical observation that females are more susceptible to cardiovascular diseases.

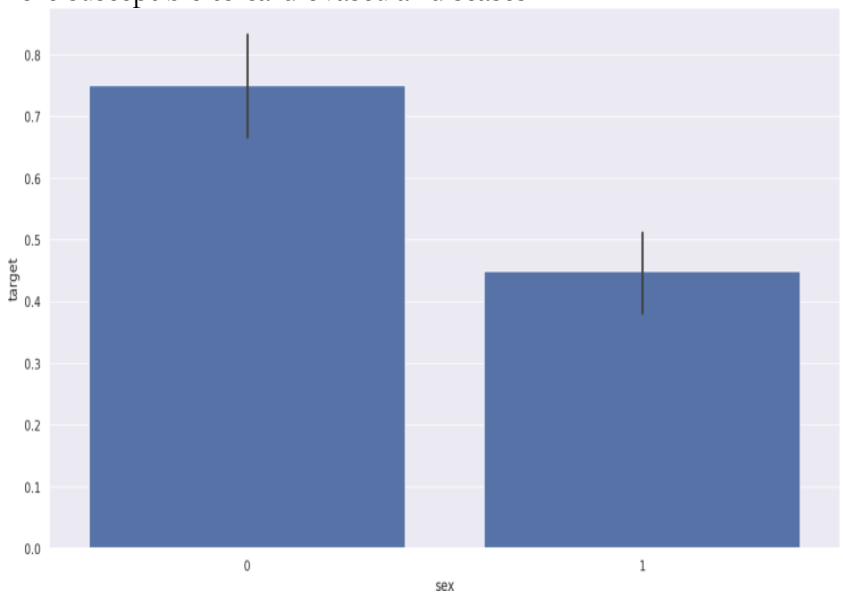


Figure 2. Gender Distribution of Patients

In terms of clinical variables, the chest pain type (CP) ranges from 0 to 3, where higher values indicate more severe chest pain. Most patients experience mild to moderate chest pain, a common symptom of cardiovascular disease. The resting blood pressure (Trestbps) values range from 94 to 200 mm Hg, with an average of 131.6 mm Hg, as shown in Figure 3. This suggests that many patients have elevated blood pressure, a known risk factor for heart disease.

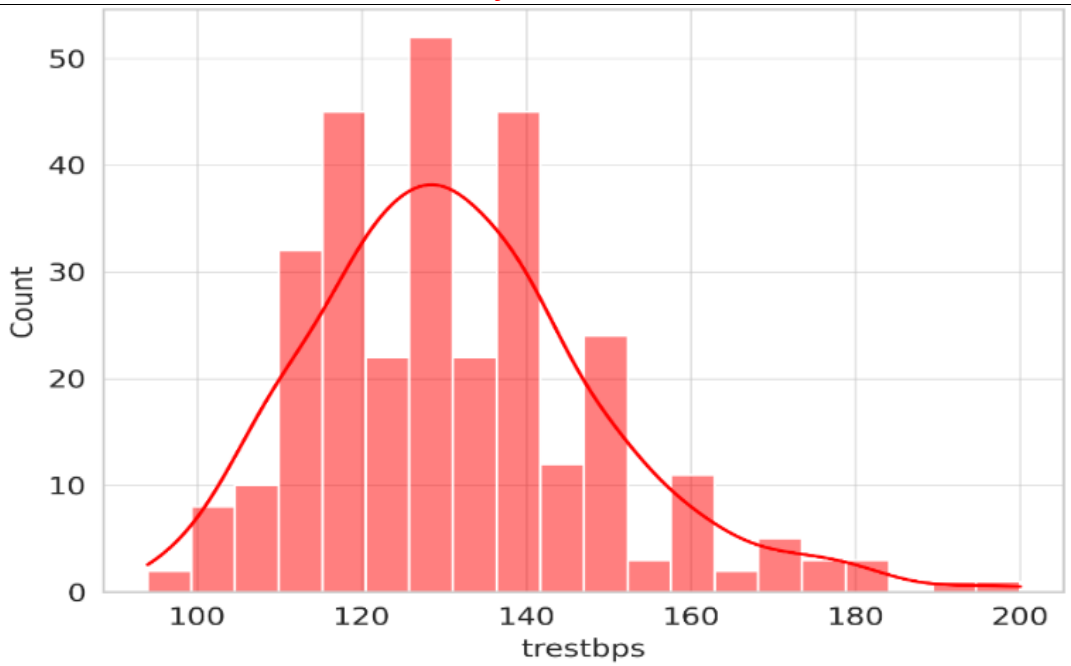


Figure 3. Resting Blood Pressure Distribution

Similarly, cholesterol levels (Chol) vary significantly, ranging from 126 mg/dL to 564 mg/dL, with an average of 246.26 mg/dL (see Figure 4). While some patients have dangerously high cholesterol levels, others fall within normal ranges, reflecting the diverse health profiles within the dataset.

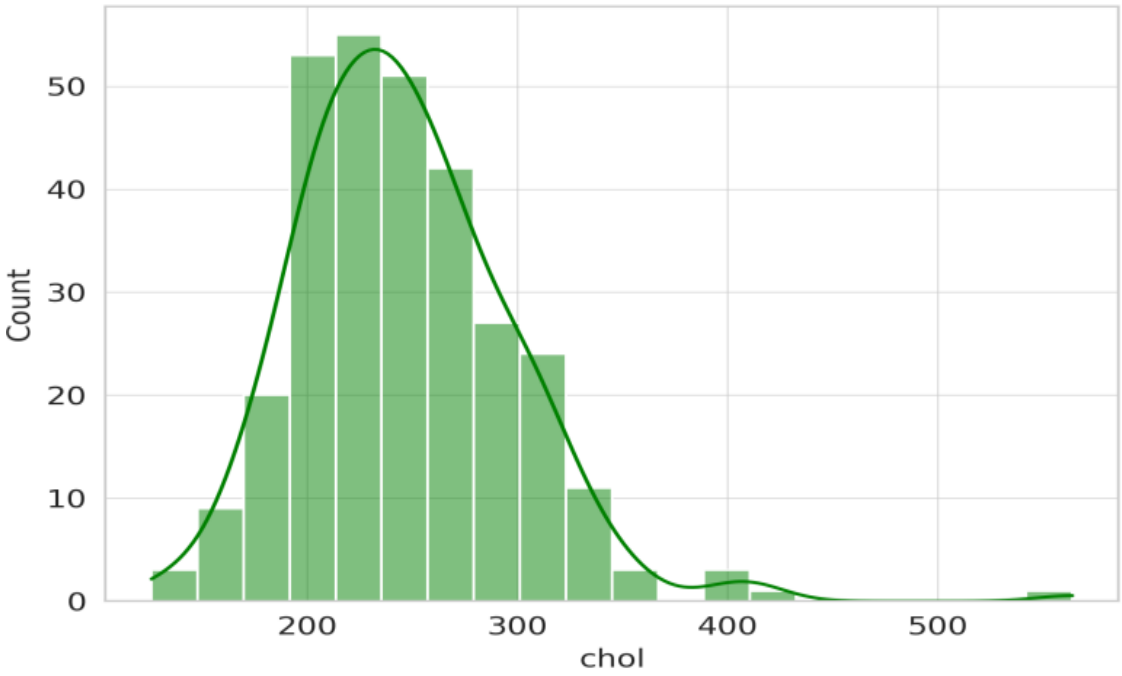


Figure 4. Cholestrol Level Distribution

The maximum heart rate achieved (Thalach) ranges from 71 to 202 beats per minute, with higher rates often linked to increased cardiovascular stress. About 32.7% of patients experienced exercise-induced angina (Exang = 1), indicating restricted blood flow during physical exertion. The ST depres- sion (Oldpeak) values range from 0 to 6.2, reflecting changes in heart function due to exercise stress, with higher values typically suggesting a higher likelihood of heart disease.

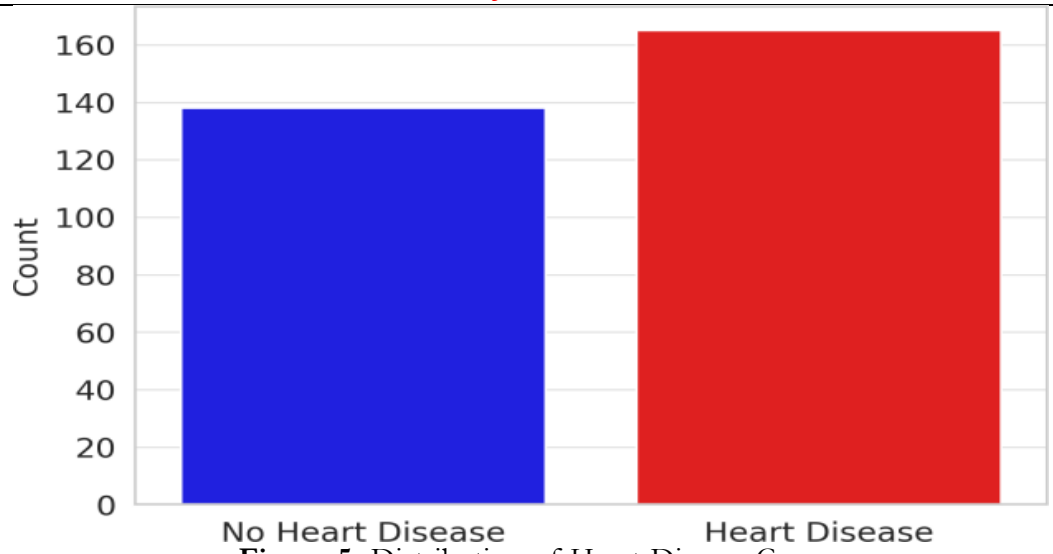


Figure 5. Distribution of Heart Disease Cases

Finally, the target variable (heart disease diagnosis), as shown in Figure 5, is relatively balanced, with 54.45% of patients diagnosed with heart disease (Target = 1) and 45.55% classified as non-CVD patients (Target = 0). This balance makes the dataset suitable for machine learning applications without the need for oversampling or under sampling techniques.

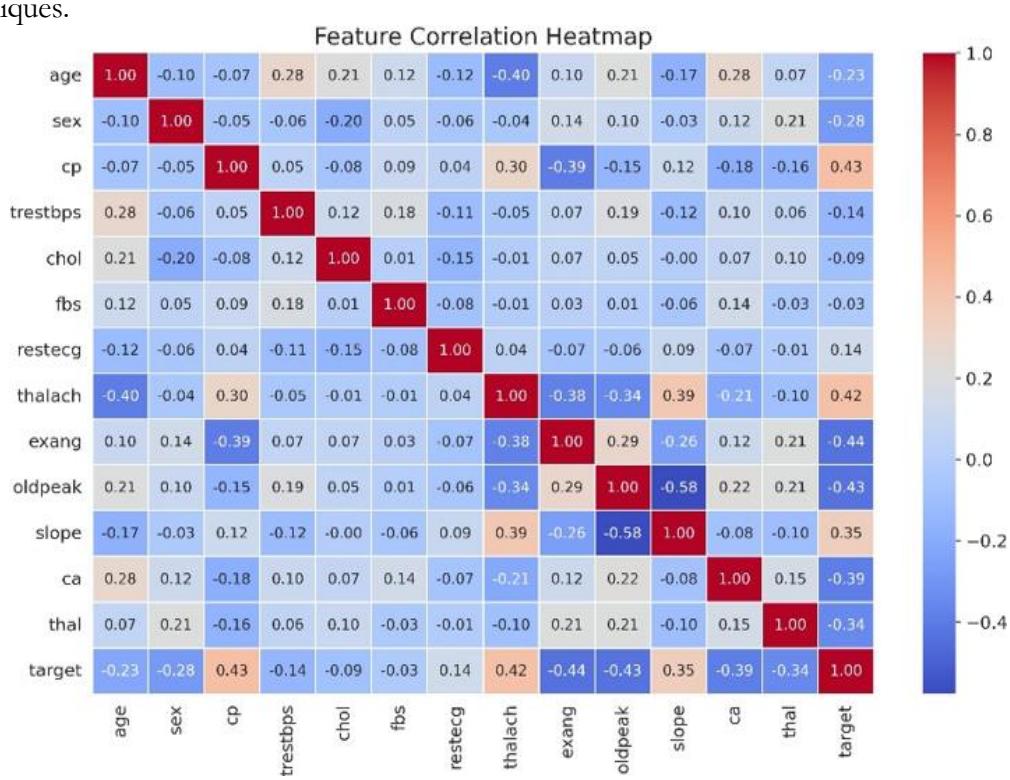


Figure 6. Correlation Analysis of Dataset

Figure 6 displays the correlation heatmap, which shows the direction and strength of relationships between key numerical variables in the cardiovascular disease dataset. The heatmap highlights the connections between important cardiovascular health indicators and the presence of disease. Among the features, chest pain type (CP) exhibits the strongest positive correlation with the target variable (0.43), indicating its strong link to heart disease. Similarly, maximum heart rate achieved (Thalach) shows a notable positive correlation (0.42), suggesting its potential as a predictive feature. On the other hand,

exercise-induced angina (Exang), ST depression (Oldpeak), and the number of major vessels colored by fluoroscopy (Ca) show strong negative correlations with the target (-0.44, - 0.43, and -0.39, respectively), suggesting a higher likelihood of disease as these values increase. Features like age, cholesterol, and resting blood pressure exhibit weaker correlations with the target, though age is moderately negatively correlated with Thalach (-0.40) and slightly positively correlated with Ca and Trestbps (0.28). Overall, the heatmap helps identify the most informative features for predicting cardiovascular disease, supporting effective feature selection in the implementation of the three-way decision model.

Methodological Approaches:

The predictive models used in this study are based on well-established machine learning algorithms that have proven efficacy in medical predictions. Specifically, Naïve Bayes, Support Vector Machine, Logistic Regression, and Random Forest models were selected on the pretext of their simplicity, interpretability, and proven accuracy in classification tasks. These models were trained on the preprocessed training dataset.

- **Logistic Regression (LR)**: Chosen for its ability to output probabilities, which is useful for the three-way decision model. It is well-suited for binary classification tasks and provides a solid baseline for evaluation.
- **Naïve Bayes (NB)**: A probabilistic model based on Bayes' Theorem, effective for categorical features and computationally efficient. It was included to assess the performance of a simpler, more intuitive model.
- **Support Vector Machine (SVM)**: An ensemble method that builds multiple decision trees and integrates their predictions to mitigate overfitting and enhance accuracy. This model was chosen for its robustness and high performance in real-world applications.
- **Random Forest (RF)**: Random Forest is an ensemble method which builds various decision trees and integrate their predictions, mitigating overfitting and enhancing predictive accuracy. This model was chosen for its robustness and high performance, especially in real-world applications.

The concept of three-way classification stems from the theory of three-way decisions, which provides an alternative to conventional binary classification. Traditional ensemble and hybrid machine learning approaches mainly aim to enhance predictive accuracy by combining multiple models, yet they still follow a binary framework, labeling instances as either positive or negative. In contrast, the Three-Way Decision (3WD) model introduces a third category: a deferment region for uncertain cases. This allows the system to refrain from making forced decisions when confidence is low, aligning with real-world medical practices where ambiguous cases often require additional testing or expert review before classification. The three-way decision model enhances interpretability, reduces misclassification in borderline cases, and promotes more responsible decision-making. It introduces flexibility by incorporating a middle-ground category, which contrasts with conventional ensemble methods that only focus on binary out- comes. In the context of CVD prediction, this model addresses situations where uncertainty arises, offering the possibility of further analysis before a definitive classification is made. Mathematically, for a patient o_i under evaluation, the relationship to a category C_k is defined by an evaluation function $e(C_k, o_i)$, which quantifies the degree of association between the patient's attributes and a particular risk group. Two thresholds—α and β—determine category membership:

- **Confirmed category** ($In(C_k)$): Patients confidently classified into a high-risk group based on their attributes exceeding a predefined threshold.
- **Excluded category** ($Out(C_k)$): Patients who clearly do

- not belong to the high-risk group, based on the same evaluation criteria.
- **Uncertain category** ($Pt(C_k)$): Instances where uncertainty exists, which may be further analyzed using a secondary classification model.

The three-way classification framework integrates with the models (Logistic Regression, Naïve Bayes, Decision Tree, and Random Forest), ensuring that ambiguous cases receive additional attention rather than being arbitrarily assigned to a category. Threshold tuning strategies, including careful selection of α and β, aim to balance the deferment rate and minimize misclassification. By adjusting these thresholds, we optimize the system's performance, enhancing reliability in real-world medical decision-making, where the cost of misclassification can be significant.

Incorporating the three-way decision model into our study allows us to address the challenge of uncertainty in machine learning-based heart disease prediction, providing more cautious and interpretable outcomes in cases of ambiguity.

$$\text{If } e(C_k, o_i) \geq \alpha \quad (1)$$

The patient is included in the high-confidence category (Con- firmed category).

$$\text{If } e(C_k, o_i) \leq \beta \quad (2)$$

The patient is excluded from the category (Excluded category).

$$\text{If } \beta < e(C_k, o_i) < \alpha \quad (3)$$

The patient is assigned to the Uncertain category, requiring further analysis. The patient was assigned to the Uncertain category, which required further analysis. This structured approach enhanced decision-making in heart disease prediction by ensuring that uncertain cases were systematically deferred, rather than being incorrectly classified. The threshold values α and β were crucial parameters that influenced model performance, and their optimization played a vital role in determining how patients were categorized. These thresholds defined the boundaries for confident classification, and the appropriate selection of these values significantly impacted the model's performance, particularly in balancing sensitivity and specificity.

To ensure the robustness of the model, the values of α and β were not arbitrarily chosen. Instead, a grid search strategy was employed. This study implemented a structured hyperparameter tuning strategy to determine the optimal values of the acceptance threshold (α) and rejection threshold (β) within the Three-Way Decision (3WD) framework. A comprehensive grid search was conducted over the range $\alpha \in [0.50, 1.00]$ and $\beta \in [0.05, 0.50]$ with a step size of 0.05, subject to the constraint that $\beta \leq \alpha$. For each (α, β) pair, a 10-fold stratified cross-validation method was employed to evaluate model performance.

A higher α resulted in stricter acceptance, increasing the number of instances classified as confident positive cases, which could reduce false negatives but may also increase false positives. Conversely, a lower β made the model more conservative in rejecting instances, but it could increase false negatives if not chosen carefully. The optimal threshold values were selected to strike a balance between sensitivity, specificity, and misclassification, ensuring that ambiguous predictions were handled with caution and that the model's performance was robust in real-world applications.

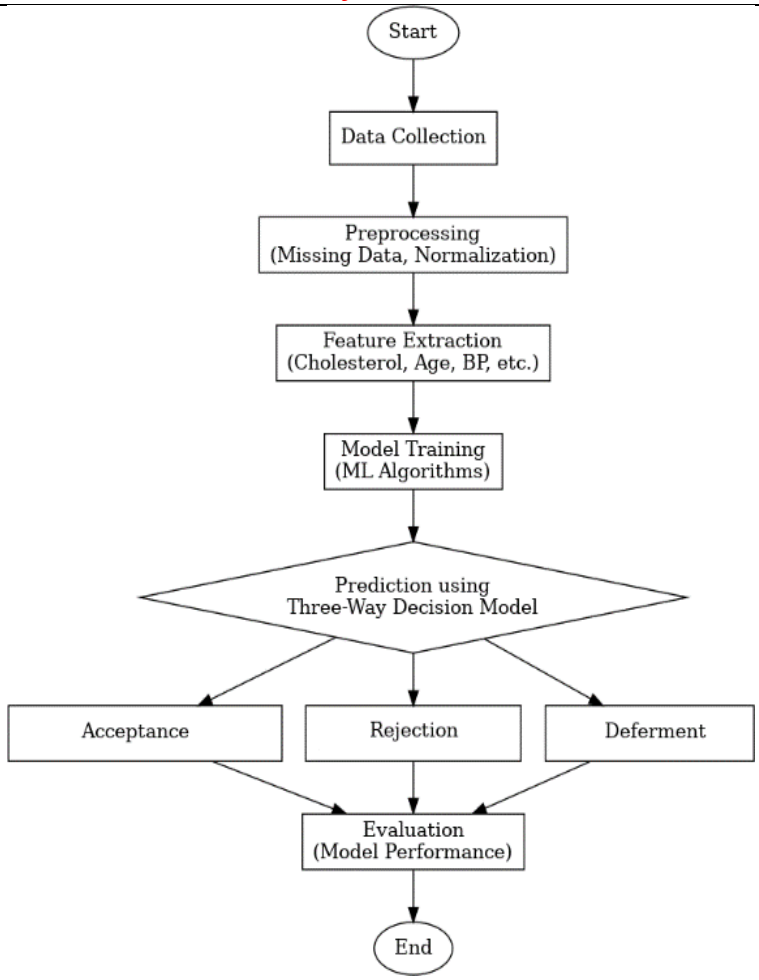


Figure 7. Flowchart of CVD prediction with 3-way Decisions

Results:

The comparative analysis of four machine learning models—Logistic Regression, Naïve Bayes, Support Vector Machine (SVM), and Random Forest—based on their performance in heart disease prediction. The evaluation considers key classification metrics: Precision, Recall, F1-score, and Accuracy, which assess each model's ability to correctly identify and classify heart disease cases. Among the models tested, Random Forest demonstrates the highest effectiveness, achieving 100% precision, 93% recall, an F1-score of 0.97, and an overall accuracy of 97%. These results indicate that Random Forest is highly capable of distinguishing between positive and negative cases, reducing false positives while maintaining a high recall. This superior performance is likely due to its ensemble learning technique, which leverages multiple decision trees to improve generalization and reduce overfitting.

Both Logistic Regression and SVM show comparable classification performance, with a precision of 0.94, an F1-score of 0.89, and accuracy levels of 91% and 90%, respectively. These results suggest that these models provide a balanced trade-off between precision and recall, making them viable options for heart disease prediction. However, their slightly lower recall values compared to Random Forest imply that they may have a marginally higher rate of false negatives, which could be a concern in medical applications where missing a diagnosis has serious implications. Naïve Bayes, in contrast, records the lowest performance, with a precision of 0.91, recall of 0.83, an F1-score of 0.87, and accuracy of 88%. This suggests that Naïve Bayes may struggle to capture complex patterns in the dataset, potentially due to its assumption that features are independent, which may not

be realistic in heart disease prediction where several risk factors interact with each other. The lower recall value indicates a higher likelihood of false negatives, which may limit its effectiveness in medical diagnoses.

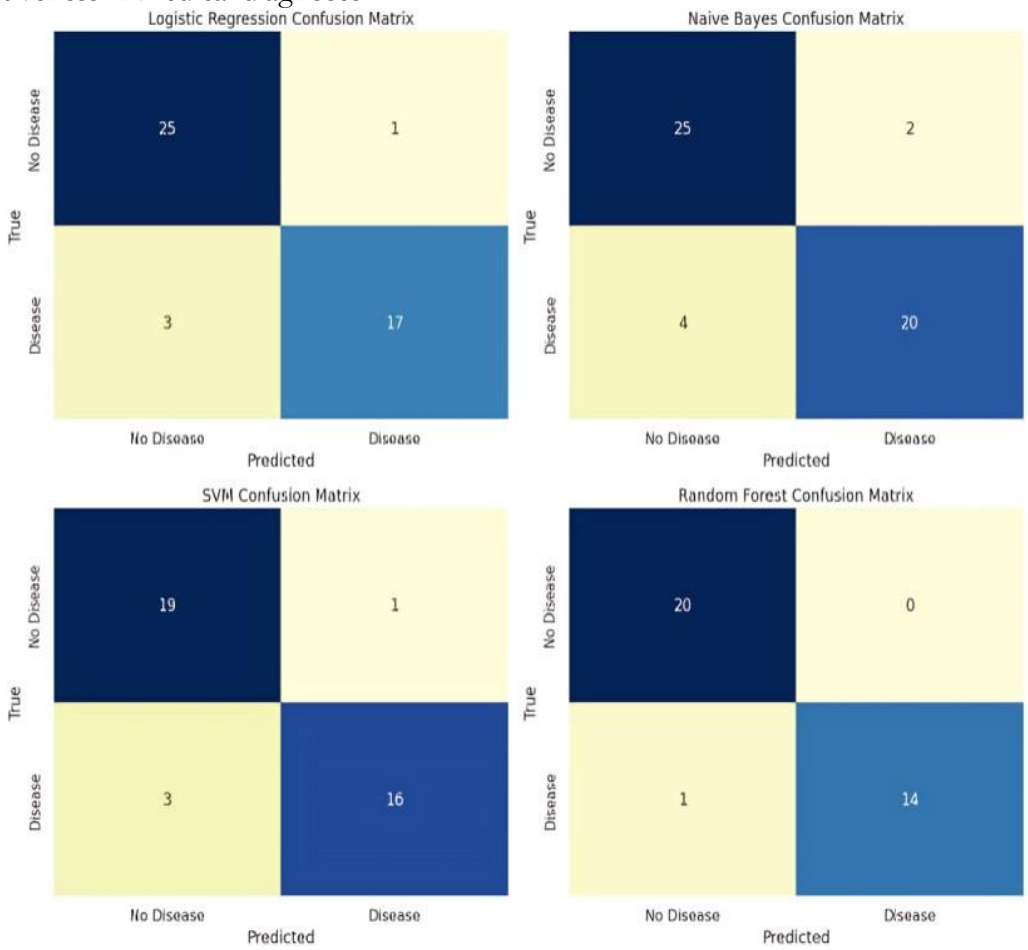


Figure 8. Confusion Matrix of Models

The findings emphasize the potential of machine learning in improving diagnostic accuracy in healthcare. Future research could explore hyperparameter tuning, feature selection, and hybrid modeling techniques to further refine these models and enhance their predictive capabilities in real-world clinical settings. After training and evaluation, the results from each model were compared to identify the best-performing algorithm for cardiovascular disease prediction. The Random Forest model showed the highest accuracy, followed by Logistic Regression and Naïve Bayes. The Decision Tree model, while still per- forming decently, had lower accuracy than the other models. Random Forest, with its ensemble approach, provided the most robust predictions, particularly for challenging cases where other models had difficulty. Despite this, the inclusion of the three-way decision model helped to significantly improve the handling of ambiguous cases, especially for the Logistic Regression and Naïve Bayes models.

Table 2. Classification Performance Comparison

Model	Precision	Recall	Accuracy
Logistic Regression	0.94	0.85	0.91
Naïve Bayes	0.91	0.83	0.88
Support Vector Machine	0.94	0.84	0.89
Random Forest	1.00	0.93	0.97

Discussion:

The results obtained in this study demonstrated improved performance over existing

machine learning approaches to cardiovascular disease (CVD) prediction. The proposed frame- work, which combines four classical classifiers Logistic Regression, Naïve Bayes, Support Vector Machine (SVM), and Random Forest with a Three-Way Decision (3WD) model, outperformed prior studies in terms of predictive accuracy, reliability, and handling of uncertainty in ambiguous cases. Notably, the Random Forest model attained a 97% accuracy, outperforming prior works such as Mohan et al. (88.7%), Islam et al. (91%), and Baghdadi et al. (92%). These improvements stem from the use of 3WD's deferment mechanism, which allows the model to abstain from uncertain predictions, thereby reducing misclassifications in borderline cases—an essential feature in medical applications. Additionally, our study showed better performance than Wang et al. and Kumar et al., whose reported accuracies ranged between 79% and 85%. Through stratified 10-fold cross-validation and optimization of probabilistic thresholds (and), our models, including Naïve Bayes (88%), SVM (89%), and Logistic Regression (91%), achieved more reliable and clinically aligned predictions.

In summary, the experimental outcomes presented in this work reflect consistent improvements over existing studies in the field. The integration of Three-Way Decision theory with standard ML classifiers, along with a rigorous cross-validation- based threshold tuning approach, enabled not only higher predictive accuracy but also introduced a layer of cautious decision-making, which is essential in real-world medical diagnosis scenarios

Conclusion and Future Work:

This study focused on improving cardiovascular disease prediction by applying a Three-Way Decision (3WD) model integrated with machine learning algorithms. By analyzing a well-structured dataset containing relevant clinical features, the research aimed to classify patients into three distinct decision regions: acceptance, rejection, and deferment, offering a more flexible, human-like approach to handling uncertainty in medical diagnosis. The results show that the 3WD framework enhances interpretability and introduces a cautious decision mechanism in borderline cases, potentially reducing misdiagnosis compared to traditional binary classifiers. The model demonstrated promising performance in identifying high-risk individuals, supporting early intervention and better resource allocation in healthcare settings.

However, the study has some limitations. One major constraint is the relatively small dataset, consisting of only 303 patient records. This limited sample size restricts the learning capacity of machine learning models and increases the risk of overfitting. Additionally, the dataset lacks demographic diversity, as it predominantly represents patients from a specific geographic and clinical context in the United States. Important factors such as ethnicity, socioeconomic status, and lifestyle habits are absent, limiting the applicability of the model to broader and more diverse populations, and affecting its real- world clinical relevance.

For future work, several improvements can be explored. Incorporating larger and more diverse datasets would enhance the model's ability to generalize across different populations. Additionally, integrating advanced techniques such as deep learning or hybrid ensemble methods with the 3WD model might further boost predictive accuracy. Feature selection and dimensionality reduction techniques could also be applied to remove irrelevant attributes and improve model efficiency. Finally, investigating the real-time implementation of the model in clinical decision support systems could help assess its practical utility in real-world healthcare environments. This research lays a strong foundation for developing intelligent, interpretable, and reliable diagnostic tools for cardio- vascular disease prediction.

Acknowledgment:

The authors hereby declare that the manuscript has not been published previously and is not under consideration for publication elsewhere. All authors have

made substantial contributions to the conception, design, execution, and interpretation of the research presented in this manuscript. Furthermore, all authors have reviewed the manuscript and are in full agreement with its content and the decision to submit it for publication

References:

[1] W.H.O, "Cardiovascular diseases (CVDs)," *World Heal. Organ.*, 2024, [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] T. O. Hirotsugu Ueshima, Akira Sekikawa, Katsuyuki Miura, Tanvir Chowdhury Turin, Naoyuki Takashima, Yoshikuni Kita, Makoto Watanabe, Aya Kadota, Nagako Okuda, Takashi Kadowaki, Yasuyuki Nakamura, "Cardiovascular Disease and Risk Factors in Asia: A Selected Review," *Circulation*, vol. 118, no. 25, 2008, doi: https://doi.org/10.1161/CIRCULATIONAHA.108.79004.

[3] W.H.O, "Noncommunicable diseases country profiles," *World Heal. Organ.*, 2024, [Online]. Available: https://www.who.int/teams/noncommunicable-diseases/surveillance/data/profiles-ncd

[4] G. L. Khor, "Cardiovascular epidemiology in the Asia–Pacific region," *Asia Pac. J. Clin. Nutr.*, vol. 10, no. 2, pp. 76–80, Jun. 2001, doi: 10.1111/J.1440-6047.2001.00230.X.

[5] H. I. Tetsuya Ohira, "Cardiovascular Disease Epidemiology in Asia," *Circ. J.*, vol. 77, no. 7, pp. 1646–1652, 2013, doi: https://doi.org/10.1253/circj.CJ-13-0702.

[6] S. M. Y. A. Jonayet Miah, Duc M Ca, Md Abu Sayed, Ehsanur Rashid Lipu, Fuad Mahmud, "Improving Cardiovascular Disease Prediction Through Comparative Analysis of Machine Learning Models: A Case Study on Myocardial Infarction," *arXiv:2311.00517*, 2023, doi: https://doi.org/10.48550/arXiv.2311.00517.

[7] M. Hajiarbabi, "Heart disease detection using machine learning methods: a comprehensive narrative review," *J. Med. Artif. Intell.*, vol. 7, no. 0, Jun. 2024, doi: 10.21037/JMAI-23-152/COIF.

[8] T. T. Dimitrios-Ioannis Kasartzian, "Transforming Cardiovascular Risk Prediction: A Review of Machine Learning and Artificial Intelligence Innovations," *Life*, vol. 15, no. 1, p. 94, 2025, doi: https://doi.org/10.3390/life15010094.

[9] K. G. Dinesh, K. Arumugaraj, K. D. Santhosh, and V. Mareeswari, "Prediction of Cardiovascular Disease Using Machine Learning Algorithms," *Proc. 2018 Int. Conf. Curr. Trends Towar. Converging Technol. ICCTCT 2018*, Nov. 2018, doi: 10.1109/ICCTCT.2018.8550857.

[10] C. T. and G. S. S. Mohan, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[11] A. E. & E. A. Nadiah A. Baghdadi, Sally Mohammed Farghaly Abdelaliem, Amer Malki, Ibrahim Gad, "Advanced machine learning techniques for cardiovascular disease early detection and diagnosis," *J. Big Data*, vol. 10, no. 144, 2023, doi: https://doi.org/10.1186/s40537-023-00817-1.

[12] T. Islam, A. Vuyia, M. Hasan, and M. M. Rana, "Cardiovascular Disease Prediction Using Machine Learning Approaches," *Proc. Int. Conf. Comput. Intell. Sustain. Eng. Solut. CISES 2023*, pp. 813–819, 2023, doi: 10.1109/CISES58720.2023.10183490.

[13] J. X. L. Wang, M. Zhao, "Prediction of Coronary Heart Disease Using Risk Factor Categories and ML Algorithms," *IEEE J. Transl. Eng. Heal. Med.*, vol. 10, no. 4, pp. 285–292, 2022.

[14] M. Nasiruddin, S. Dutta, R. Sikder, M. R. Islam, A. AL Mukaddim, and M. A. Hider, "Predicting Heart Failure Survival with Machine Learning: Assessing My Risk," *J.*

Comput. Sci. Technol. Stud., vol. 6, no. 3, pp. 42–55, Aug. 2024, doi: 10.32996/JCSTS.2024.6.3.5.

[15] E. A. Mohammad Khan Afridi, Nouman Azam, JingTao Yao, "A three-way clustering approach for handling missing data using GTRS," *Int. J. Approx. Reason.*, vol. 98, pp. 11–24, 2018, doi: https://doi.org/10.1016/j.ijar.2018.04.001.

[16] Yiyu Yao, "Three-way decisions with probabilistic rough sets," *Inf. Sci. (Ny).*, vol. 180, no. 3, pp. 341–353, 2010, doi: https://doi.org/10.1016/j.ins.2009.09.021.

[17] David Lapp, "Heart Disease Dataset," *Kaggle*, 2024, [Online]. Available: https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset



Copyright © by authors and 50Sea. This work is licensed under Creative Commons Attribution 4.0 International License.