Explainable Heart Disease Detection using Voting-Based Ensemble Methods, Experimenting in Data Imputation Methods

Asty Nabilah ‘Izzaturrahmah
School of Computing
Telkom University
Bandung, Indonesia
izzaturrahmah@student.telkomuniversity.ac.id

***Abstract*—Cardiovascular diseases (CVDs) are the leading cause of death worldwide, with coronary artery disease (CAD) being the most prevalent form. This study aims to compare different data imputation methods and identify the best parameters for classifying heart disease using ensemble methods, building on the work by Doppala et al. The research involves three experiments: removing missing values, mean imputation, and safe-region imputation, followed by training with a voting-based ensemble method. Results show that binary classification consistently outperforms 5-label classification, with safe-region imputation yielding the best performance. Visualization and explainable AI techniques are used to gain insights from the data. In the first experiment, removing missing values led to improved performance in binary classification (accuracy: 83.33%, F1-score: 83.33%) compared to 5-label classification (accuracy: 56.67%, F1-score: 26.81%). The second experiment, using mean imputation, showed better results for 5-label classification (accuracy: 62.50%, F1-score: 35.06%), while binary classification performance was slightly lower (accuracy: 82.87%, F1-score: 82.87%). The third experiment, using safe-region imputation, achieved the highest performance in binary classification (accuracy: 83.80%, F1-score: 83.80%). Explainable AI is employed to focus on the best-performing model, ensuring relevant and reliable insights. The correlation matrices indicate that binary classification benefits from strong relationships with key features like age, fbs, restecg, thalach, oldpeak, ca, and thal, which align with variable importance plots. Visualizations show that features like chest pain type, oldpeak, thalach, cholesterol, and thalassemia are crucial for predicting heart disease.**

Keywords—heart disease detection, ensemble method, data imputation, explainable AI

I. Introduction

Cardiovascular diseases (CVDs) are the top cause of death globally, responsible for about 17.9 million deaths each year. These conditions, which include coronary heart disease, cerebrovascular disease, and rheumatic heart disease, affect the heart and blood vessels. More than 80% of CVD deaths are due to heart attacks and strokes, with one-third of these fatalities occurring in individuals under the age of 70 [1]. Coronary artery disease (CAD) is the most frequently occurring form of cardiovascular disease (CVD) [2]. CAD is a significant health issue that arises from plaque accumulation in the coronary arteries, which obstructs the flow of oxygen-rich blood to the heart. Individuals with this condition might be asymptomatic, suffer from chest pain or discomfort (angina), or experience heart attacks [3], [4], [5].

Recently, medical data, including heart disease data, have been stored online and some are available as open source. However, these datasets often contain missing values. Medical data are crucial for accurate diagnosis and treatment, but missing values (MVs) or errors in records can obstruct complete patient information. This leads to incomplete records for medical professionals. Ethical guidelines mandate that patient information be shared only with consent. Patients may be reluctant to share certain details, increasing the occurrence of MVs. Therefore, imputing missing values is essential in medical research to achieve more complete and accurate data for analysis [6], [7]. While many imputation methods exist [8], [9], [10], not all provide optimal results. A common but limited approach is to delete records with missing values, suitable only when such records are few and their pattern is unknown. Proper estimation of missing values during data preprocessing is essential to minimize data loss and improve outcomes [7].

A study by Jian Ping Li et al. developed a heart disease detection model using various classification algorithms and feature selection techniques like minimum relevance maximum redundancy and Relief. They tested this model on the Cleveland Heart Disease dataset, where the Support Vector Machine, paired with the proposed feature selection method (FCMIM), achieved an accuracy of 92.37%. This approach outperformed deep neural networks in detecting heart disease [11]. Another study by Mohamed Elhoseny et al. developed an automated heart disease diagnostics system using a binary CNN and a multi-agent shell model (MAFW). Tested on the Cleveland Heart Disease database, the system achieved a maximum accuracy of 90.1%, a high accuracy of 88.9%, and a recall of 98.4%, outperforming other machine learning and conventional CNN models, which had accuracies between 72.3% and 83.8% [12]. Umarani Nagavelli et al. evaluated four machine learning models for heart disease detection. The XGBoost algorithm performed best in accuracy, precision, recall, and F1 scores, while the Naive Bayes and dual-optimized SVM models showed lower performance [13].

In 2023, Das et al. used BRFSS data, which includes 319,795 heart disease-related cases, to evaluate six machine learning models. XGBoost performed best with 91.30% accuracy, 92% sensitivity, 83% AUC, and a 95.40% F1-score. Random Forest also performed well with 90.20% accuracy, 92.50% sensitivity, 78% AUC, and a 94.78% F1-score. Models were evaluated using an 80/20 hold-out validation approach [14]. Doppala and colleagues proposed ensemble methods combining Naive Bayes, Random Forest, SVM, and XGBoost for heart disease detection, using three datasets: the Cleveland dataset (303 subjects), a cumulative cardiovascular dataset (1190 cases), and a heart illness dataset from an Indian hospital (1000 subjects). Their model achieved high accuracy rates: 96.75% on the cardiovascular dataset from Mendeley, 93.39% on the IEEE DataPort dataset, and 88.24% on the Cleveland dataset, outperforming existing models on all three datasets [15].

ALGORITHM 1 Safe-Region Imputation

```
c_ins: complete instance, i_ins: incomplete instance, k: the k
nearest points of dataset, and new_ins: imputed
instance.
# calculate the minimal distance
for ith in c_ins:
    for jth in c_ins:
        if ith != jth then
            dis[ith] = distance_complete(c_ins[ith], c_ins[jth])
set top_dis[k] as top k distance array;
top_dis = get_shortest(dis, k) # get the k minimal distance points
# impute missing value
set min_dis as min distance variable
set min_index as min distance data point
missTh = 0.05
failTh = 0
for ith in i_ins:
    for jth in top_dis:
        i_dis = distance_incomplete(i_ins[ith], c_ins[jth])
        if i_dis < min_dis or min_dis is null then
            min_dis = i_dis
            min_index = jth
for th in 0.01 to 3.00 step 0.01
    for ath in len(i_ins.arrt):
        for ith in i_ins:
            if i_ins[ith].attr[ath] is missing then/the attribute a is a missing value in the ith instance
                set score = 0
                for jth in top_dis:
                    if top_dis[jth].attr[ath] < th then
                        score++
                if score == 0 then
                    failTh++
                else
                    for s in score:
                        new_val = new_val + c_ins[top_dis[s]].attr(ath)
                    new_val = new_val/s
                new_ins[ith].attr(ath) = new_val
            else
                new_ins[ith].attr(ath) = i_ins[ith].attr(ath)
```

This study aims to compare different ways for data imputation and find the best parameters for classifying heart disease using ensemble methods. It builds on the previous work, using ensemble method proposed by Doppala et al. [15], who proposed an ensemble method applied to a dataset with the same features as the one used in this study.

II. LITERATURE REVIEW

A. Related Works

In a study by Jian Ping Li et al., an effective machine learning model for heart disease detection was developed, utilizing various classification algorithms. To refine the model, feature selection techniques such as minimum relevance maximum redundancy, Relief, and the Local Learning Least Absolute Shrinkage Selection Operator were applied to eliminate irrelevant and redundant features. The model was then evaluated using the Cleveland Heart Disease dataset and several evaluation metrics. The results showed that the Support Vector Machine algorithm, paired with the proposed feature selection method (FCMIM), achieved an accuracy of 92.37%. This machine learning-based method (FCMIMSVM) also demonstrated superior performance compared to deep neural networks in heart disease detection [11].

Mohamed Elhoseny et al. introduced an automated heart disease diagnostics (AHDD) system that combines a binary convolutional neural network (CNN) with a cutting-edge multi-agent shell model (MAFW). They evaluated the system using the Cleveland Heart Disease database, and the hybrid model achieved a peak accuracy of 90.1%, a high accuracy of 88.9%, and an impressive recall of 98.4%. In comparison, other machine learning models and standard CNN models typically showed accuracy levels between 72.3% and 83.8% on average [12].

Umarani Nagavelli et al. assessed four machine learning models for heart disease detection. The models were evaluated using precision, accuracy, recall, and F1 score metrics. These models included an enhanced prediction-based SVM (ISVM) utilizing a duality optimization (DO) system, a weighted approach-based prediction, two XGBoost-based prediction SVMs, and a prediction model based on XGBoost alone. The results indicated that the XGBoost algorithm performed exceptionally well, achieving high accuracy, precision, recall, and F1 scores. In comparison, the Naive Bayes model with a weighted approach had lower accuracy, and the dual-optimized (DO) SVM model exhibited lower precision, recall, and F1 scores [13].

In 2023, Das et al. used dataset from the Behavioral Risk Factor Surveillance System (BRFSS) to predict heart disease with six machine learning models. The dataset come from over 400,000 adult interviews annually across the US, collecting health-related data. XGBoost performed best, achieving an accuracy of 91.30%, sensitivity of 92%, AUC of 83%, and an F1-score of 95.40%. Random Forest also showed strong results, with an accuracy of 90.20%, sensitivity of 92.50%, AUC of 78%, and an F1-score of 94.78% [14].

A study by Doppala and colleagues proposed ensemble methods, combining Naive Bayes, Random Forest, SVM, and XGBoost for heart disease detection. They used three datasets: the Cleveland dataset with 303 subjects, a cumulative cardiovascular dataset with 1190 cases, and a heart illness dataset from an Indian hospital with 1000 subjects. The proposed model achieved high accuracy rates of 96.75% on the cardiovascular dataset from Mendeley, 93.39% on the comprehensive dataset from IEEE DataPort, and 88.24% on the Cleveland dataset from the UCI repository. The model outperformed existing models on all three datasets [15].

B. Safe-Region Data Imputation

Huang and Cheng, in 2020, proposed Safe-Region Imputation Method to handle medical dataset with missing values. Their study proposes a safe-region imputation method specifically for handling medical data with MVs. This method only imputes data points in the safe and boundary areas while discarding those in the sparse and outlier areas. The safe-region imputation uses k-nearest neighbors (kNN) where k = 5 for safe areas and k = 3 for boundary areas. The method identifies data points based on their nearest neighbors and class labels, categorizing them into safe, boundary, sparse, or outlier areas according to specific criteria [7]. The detailed algorithm is shown in Algorithm 1.

The imputation process involves calculating the minimal distance between data points and using the average of the nearest five data points to replace MVs. The effectiveness of this method is verified by comparing kNN and multiple imputation results before and after imputation. Decision Tree, Random Forest, REP Tree, and LMT classifiers were applied to evaluate performance data before and after imputation. The imputation method was proven to increase accuracy and AUC scores for datasets related to diabetes, audiology, thyroid

TABLE 1. DATA DESCRIPTION

Attribute	Description
age	Patient's age
sex	1 = male, 0 = female
cp	Chest Pain Type; 0 = typical angina, 1 = atypical angina, 2 = non-anginal pain, 3 = asymptomatic
trestbps	Resting Blood Pressure; Measured in mm Hg
chol	Serum Cholesterol; Measured in mg/dl
fbs	Fasting Blood Sugar; Levels greater than 120 mg/dl (1 = true, 0 = false)
restecg	Resting heart rate; 0 = normal, 1 = ST-T wave abnormality, 2 = probable or definite left ventricular hypertrophy
thalach	Maximum heart rate achieved
exang	Exercise Induced Angina; 1 = yes, 0 = no
oldpeak	ST Depression Induced by Exercise Relative to Rest
slope	Slope of the Peak Exercise ST Segment; 0 = upsloping, 1 = flat, 2 = downsloping
ca	Number of Major Vessels Colored by Fluoroscopy; Ranges from 0 to 3
thal	Thalassemia; 3 = normal, 6 = fixed defect, 7 = reversible defect
num	Target variable; 0 = no heart disease (< 50% diameter narrowing), 1 = heart disease (> 50% diameter narrowing)

disease, breast cancer, and stroke. Overall, the proposed imputation is better than kNN imputation [7].

C. Naïve Bayes

Naive Bayes is a simple yet potent classification technique. It is consistently employed due to its effectiveness as a learning algorithm. The classifier is represented by the following equation:

$$\mathrm{P(C|X)} = \frac{\mathrm{P(X|C)P(C)}}{\mathrm{P(X)}} \quad (1)$$

where P(C|X) is the posterior probability, P(X|C) is the likelihood, P(C) is the class prior probability, and P(X) is the predictor prior probability [16].

D. Random Forest

Random forest is a specialized bagging supervised machine learning technique commonly utilized for classification and prediction tasks [17]. This method constructs an ensemble of decision trees, each derived from bootstrap training samples. The bootstrap classifier closely mirrors the decision tree hyperparameter. For the forecasting model to perform optimally, it is important to fully grow the trees in both size and number. Additionally, an appropriate number of predictors must be selected at each node, and the number of observations at the terminal nodes should be kept minimal. This is the operational mechanism of the random forest ensemble method [18]. The benefits of Random Forest (RF) include its ability to manage missing data within the dataset, generate lower error rates, efficiently process large volumes of training data, deliver robust classification outcomes, and prevent overfitting [19].

E. SVM

Support Vector Machine (SVM) is a highly efficient machine learning method that assigns labels based on learned instances [20], [21]. It is widely recognized as the most reliable and efficient approach for various learning tasks, providing complete automation without the need for human parameter adjustment [22]. SVM enhances its capacity to classify unknown data by locating the Maximum Marginal Hyperplane (MMH), which is defined by the largest significant margin [22], [23]. SVM classification is based on four key concepts: the kernel function, the soft margin, the maximum-margin hyperplane, and the separating hyperplane [24]. A hyperplane that separates is represented by

$$\mathrm{W.X} + \mathrm{b} = 0 \quad (2)$$

W symbolizes the weight vector, and b serves as the bias scalar. Meanwhile, the two sides of the edge, or the dashed line is expressed as [25]

$$H1 = w_0 + w_1x_1 + w_2x_2 \geq 1 \quad for\ y_i = \ +1 \quad (2)$$

$$H2 = w_0 + w_1x_1 + w_2x_2 \leq -1 \quad for\ y_i = \ -1 \quad (3)$$

Formula (3) and (4) are combined, resulting in [25]

$$y_i(w_0 + w_1x_1 + w_2x_2) \geq 1,\ \forall_i \quad (4)$$

SVM faces difficulties when there is no distinct separation between classes, making it challenging to use a soft margin. To overcome this, the kernel function is employed to enhance data dimensionality. Commonly used kernels include linear, polynomial, Gaussian RBF, and sigmoid [23], [26].

F. XGBoost

Boosting is a technique that aggregates the outputs of several weak classifiers to create a single, stronger classifier. Extreme Gradient Boosting, commonly known as XGBoost, is an advanced version of the Gradient Boosting method. XGBoost was developed to enhance scalability, increase computation speed, and improve generalization performance. It has proven to be highly effective and is widely used in various machine learning and data mining tasks [27].

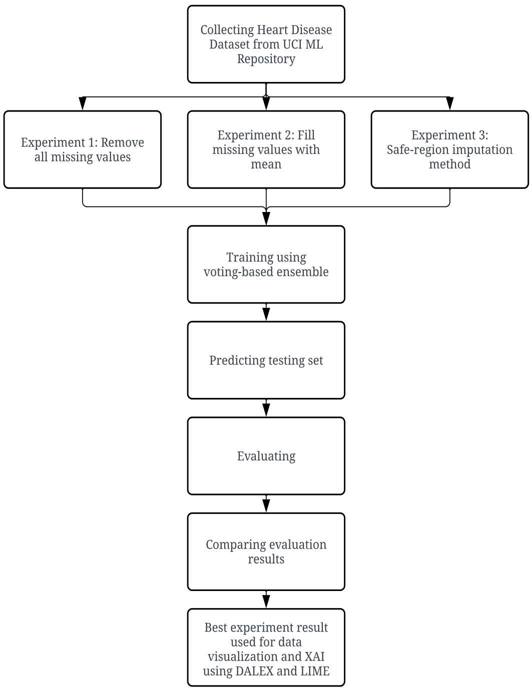
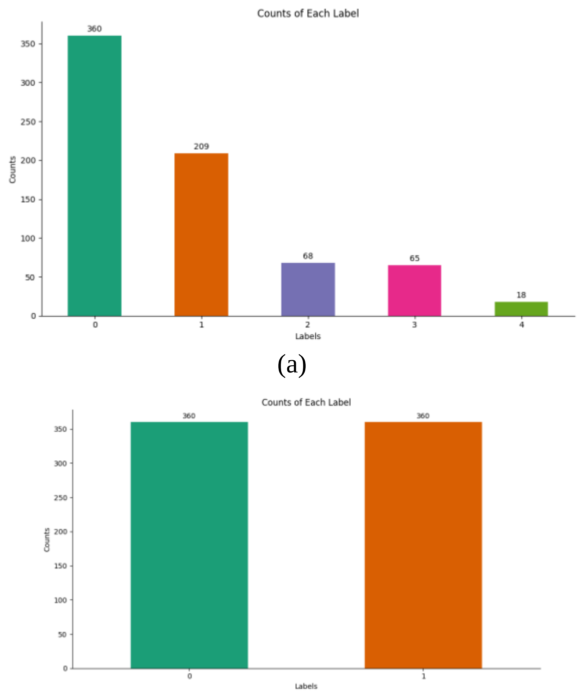

Collecting Heart Disease Dataset from UCI ML Repository
Experiment 1: Remove all missing values
Experiment 2: Fill missing values with mean
Experiment 3: Safe-region imputation method
Training using voting-based ensemble
Predicting testing set
Evaluating
Comparing evaluation results
Best experiment result used for data visualization and XAI using DALEX and LIME

Fig.1. Flowchart of this study



(a)

(b)

Fig. 2. Number of class for each label, (a) for 5-labels dataset and (b) for binary label dataset

G. *Voting-Based Ensemble*

Each model version produces a prediction for each test case, with the most frequent prediction being chosen. If no prediction receives more than half of the votes, the ensemble method might not yield a consistent forecast. Therefore, it is needed to determine the class$\hat{y}$ based on the majority vote of all classifiers (C_j):

$$\hat{y} = \mathrm{mode}\{C_1(x), C_2(x), \dots, C_m(x)\}. \qquad (5)$$

Majority voting is calculated by assigning a weight (w_j) to each classifier (C_j):

$$\hat{y} = \max_i \sum_{j=1}^{m} w_j \chi_A \big(C_j(x) = i\big) \qquad (6)$$

where (χ_A) is the characteristic function $(C_j(x) = i \in A)$. A is a unique label set of a class. The predicted probability of the classifier is given by:

$$\hat{y} = \max_i \sum_{j=1}^{m} w_j p_{ij}. \qquad (7)$$

The proposed algorithm is represented in Table 3 [15].

H. *DALEX*

Descriptive mAchine learning eXplanations **(**DALEX) offers a robust set of tools for the model-agnostic explanation and exploration of machine learning models. The package addresses the challenge of understanding machine learning models, which are often perceived as black boxes, making it hard to comprehend their workings and reasons for specific predictions. By measuring the contribution of each feature to the model's prediction and showing the relationship between a feature and the model's prediction while keeping all other features constant, DALEX provides valuable insights into model behavior [28], [29].

I. *LIME*

The Local Interpretable Model Agnostic Explanations (LIME) creates a new dataset by permuting feature values and assessing their impact on the model's outcome. The interpretative models generated are trained on this new dataset, providing a description of each feature's contribution to local predictions. The local surrogate model of LIME is represented by the following equation:

$$\text{explanation}(x) = \arg\min_{g \in G} L\,(f, g, \pi_x) + \Omega(g) \qquad (8)$$

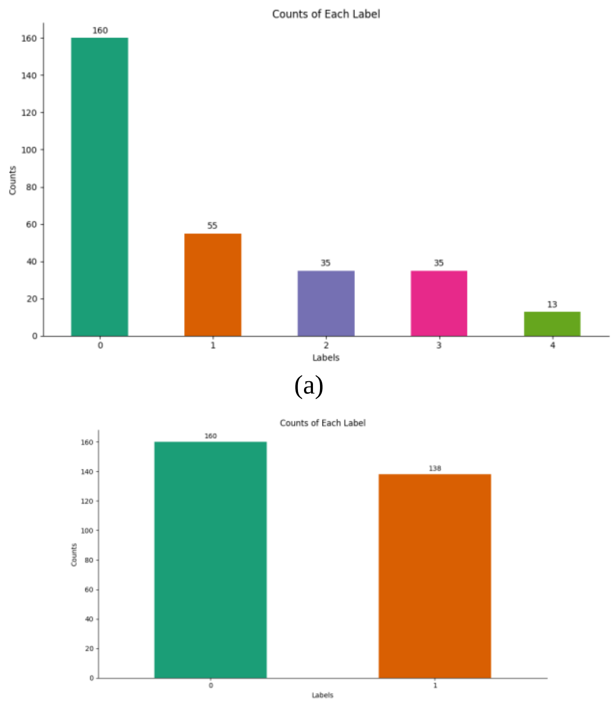
Here, f denotes the complex model and $g$ denotes the simple, interpretable model. $g$ represents the family of interpretable linear models. $L(f, g, \pi_x)$ is the loss term in the optimization function, which approximates the complex model f with the simple model $g$ in the neighborhood of the data point x. $\Omega(g)$ is used to regularize the complexity of the simple surrogate model. The goal is to minimize the loss term L, indicating the approximation accuracy of the complex model in the local area.

III. Methodology

As shown in flowchart in Fig.1., this research consists of several steps. First, data is collected from the UCI ML Heart Disease Dataset. The preprocessing step involves three experiments: removing all missing values, filling missing values with the mean, and using the safe-region imputation method. Next, a voting-based ensemble is trained, followed by predicting the testing set. The predictions are then evaluated, and the results are compared. Finally, the best experiment result is used for data visualization and explainable AI (XAI) using DALEX and LIME.

A. *Data Collection*

The dataset was collected from UCI Machine Learning Repository, a Heart Disease Dataset focusing on coronary artery disease (CAD). This dataset consists of 14 attributes plus the target variable. These attributes include Age, Sex, Chest Pain Type, Resting Blood Pressure (Trestbps), Serum



(a)

(b)

Fig. 3. Number of class for each label, (a) for 5-labels dataset and (b) for binary label dataset after missing values removed

TABLE 2. Results of Experiments

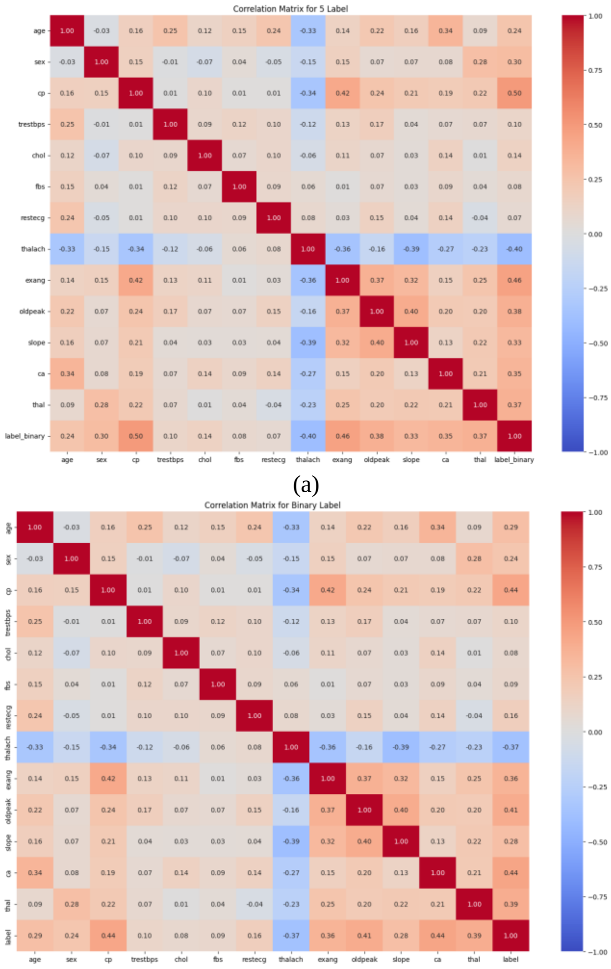
Dataset	Number of Labels	Accuracy	F1-score	Recall	Precision
First Experiment (removing all missing values)	5 labels	56.67%	26.81%	30.43%	25.16%
	Binary label	83.33%	83.33%	83.54%	83.48%
Second Experiment (replacing all missing value with mean)	5 labels	62.50%	35.06%	35.43%	38.69%
	Binary label	82.87%	82.87%	83.15%	83.19%
Third Experiment (applying safe-region imputation)	5 labels	59.26%	30.91%	31.83%	31.84%
	Binary label	83.80%	83.80%	84.03%	84.00%

Cholesterol (Chol), Fasting Blood Sugar (Fbs), Resting Electrocardiographic Results (Restecg), Maximum Heart Rate Achieved (Thalach), Exercise Induced Angina, ST Depression Induced by Exercise Relative to Rest (Oldpeak), Slope of the Peak Exercise, Number of Major Vessels Colored by Fluoroscopy, and Thalassemia (Thal). The details of this dataset is shown in Table 1 [30].

This data originally has 5 labels, ranged from 0 to 5, with 0 indicates no heart disease, and 1 to 5 indicates the degree of severity. In this experiment, the performance of this 5 labels dataset is compared with binary label dataset. The binary label is extracted with the criteria 0 for no heart disease and 1 for heart disease. The label ranged from 1 to 5 are merged into label 1. This data consists of 720 rows. The data then split into training and testing dataset, with ratio 7:3. The amount of data in each class for all datasets are shown in Fig. 2.

B. *Data Imputation*

Three scenarios are performed as experiment. In the first scenario, all missing values are removed. For the second experiment, all missing values are replaced by the mean values for each corresponding features. The third experiment utilizes safe-region imputation proposed by Huang and Cheng [7].


Correlation Matrix for 5 Label

(a)

Correlation Matrix for Binary Label

(b)

Fig. 4. Correlation plot for (a) 5-label dataset and (b) binary label dataset

C. *Data Visualization and XAI*

In this step, the data used for visualization is the one after safe-region imputation (third experiment) is done. After that, the explainable AI tools, DALEX and LIME, are applied. These are done because this dataset returns the best accuracy amongst all experiments.

D. *Voting-Based Classifier*

The library sklearn is used for Naive Bayes, Random Forest, and SVM [31], while XGBoost is applied from the xgboost library [32]. Each classifier is trained using a provided training dataset. Specifically, Gaussian Naive Bayes, Random Forest with 100 estimators, SVM with probability estimation, and XGBoost without label encoding are employed. These classifiers are then combined into an ensemble model using sklearn's VotingClassifier with soft voting. The ensemble model is subsequently trained on the same dataset to enhance overall performance evaluation.

E. *Model Validation*

To evaluate the performance and reliability of machine learning models, model validation is essential. This study employs accuracy, recall, precision, and F1-score as validation metrics. Calculating these metrics requires the use of specific formulas [28].

$$\text{accuracy} = \frac{True\ Positives + True\ Negatives}{Positives + Negatives} \quad (5)$$

$$\text{recall} = \frac{True\ Positives}{Positive\ Samples} \quad (6)$$

$$\text{precision} = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (7)$$

$$\text{F1} - \text{score} = \frac{2\ x\ precision\ x\ recall}{precision + recall} \quad (8)$$

IV. Result and Discussions

In the first experiment, all missing values were removed to ensure the reliability of the data. There were 422 data points with missing values and removing them resulted in a dataset of 298 complete data points. The detail of dataset is shown in Fig. 3. The second experiment is applied by simply filling missing values with the average value of the corresponding features. Finally, in the third experiment, safe-region imputation method is applied. Then, each dataset is trained using voting-based ensemble method, and the results are compared. The results are presented in Table 2.

In the first experiment, where all missing values were removed, the performance metrics for the 5-label classification were relatively low, with an accuracy of 56.67%, an F1-score of 26.81%, a recall of 30.43%, and a precision of 25.16%. However, for the binary classification, the results improved significantly, achieving an accuracy of

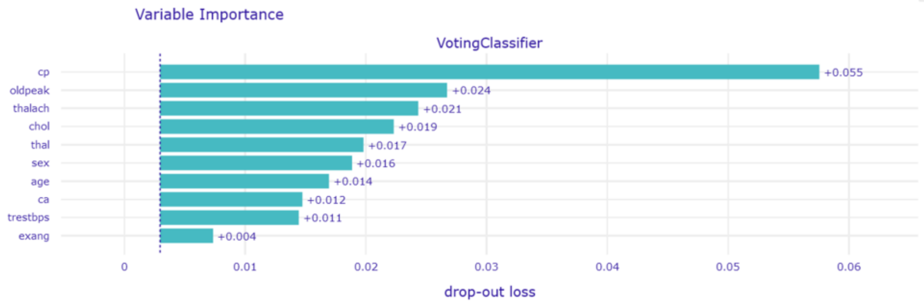


Fig. 5. Feature importance of VotingClassifier, generated using DALEX

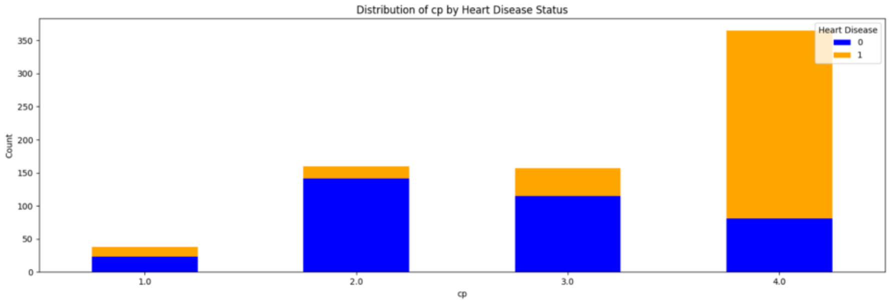


Fig. 6. Distribution of cp (chest pain type) by heart disease status

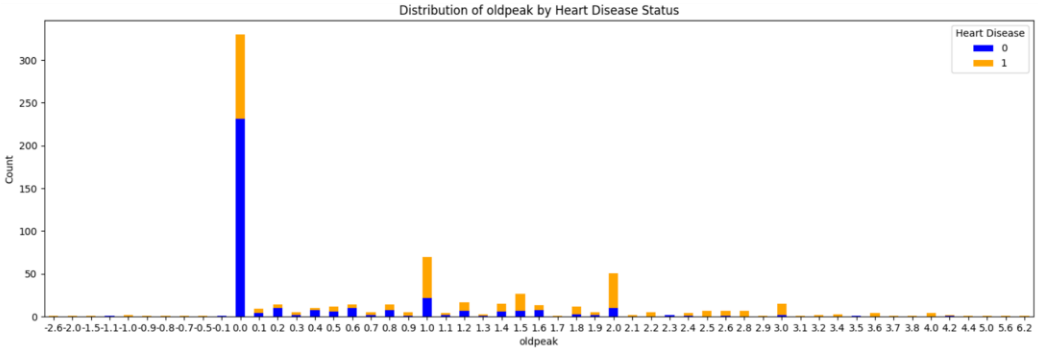


Fig. 7. Distribution of oldpeak (ST depression induced by exercise relative to rest) by heart disease status

83.33%, an F1-score of 83.33%, a recall of 83.54%, and a precision of 83.48%.

The second experiment involved replacing missing values with the mean. This method showed better performance than the first experiment for the 5-label classification, with an accuracy of 62.50%, an F1-score of 35.06%, a recall of 35.43%, and a precision of 38.69%. For the binary classification, the results were slightly lower than those of the first experiment, with an accuracy of 82.87%, an F1-score of 82.87%, a recall of 83.15%, and a precision of 83.19%.

In the third experiment, the safe-region imputation method was applied. This method resulted in an accuracy of 59.26%, an F1-score of 30.91%, a recall of 31.83%, and a precision of 31.84% for the 5-label classification. For the binary classification, it achieved the highest performance among the three experiments, with an accuracy of 83.80%, an F1-score of 83.80%, a recall of 84.03%, and a precision of 84.00%.

To conclude, binary classification generally outperformed the 5-label classification across all metrics. Among the three imputation methods, the safe-region imputation method yielded the best results for binary classification, with the slight difference with the first experiment. If all missing values are removed, several important data points are also removed, leaving the machine without enough data and patterns to learn from. Replacing missing values with the average allows the machine to have complete data but might dilute the unique patterns in the data, leading to less accurate predictions. The safe-region imputation works by identifying data points in the safe and boundary areas using k-nearest neighbors (kNN) and imputing missing values based on the average of the nearest neighbors, while discarding data points in sparse and outlier areas. This targeted approach ensures that only reliable data is used for imputation, preserving the dataset's integrity, and improving the overall model performance. As a result, the performance is better than the other two methods.

Explainable AI is performed after selecting the dataset with the highest performance to ensure that the interpretability and visualization efforts are focused on the most effective model. This approach is efficient because it avoids the need to repeatedly generate explanations and visualizations for less accurate models, thereby saving time and resources. By concentrating on the best-performing dataset, it can provide the most accurate and meaningful insights into the model's decision-making process, ensuring that the explanations are relevant and reliable.

The correlation matrices (Fig. 4) show that binary classification benefits from strong relationships with key features like age, fbs, restecg, thalach, oldpeak, ca, and thal. This matches the variable importance plot (Fig. 5), which shows that features like oldpeak, thalach, ca, and thal are very

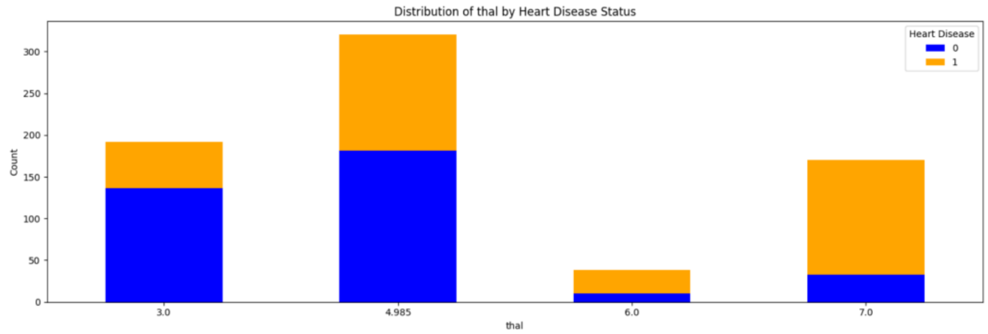
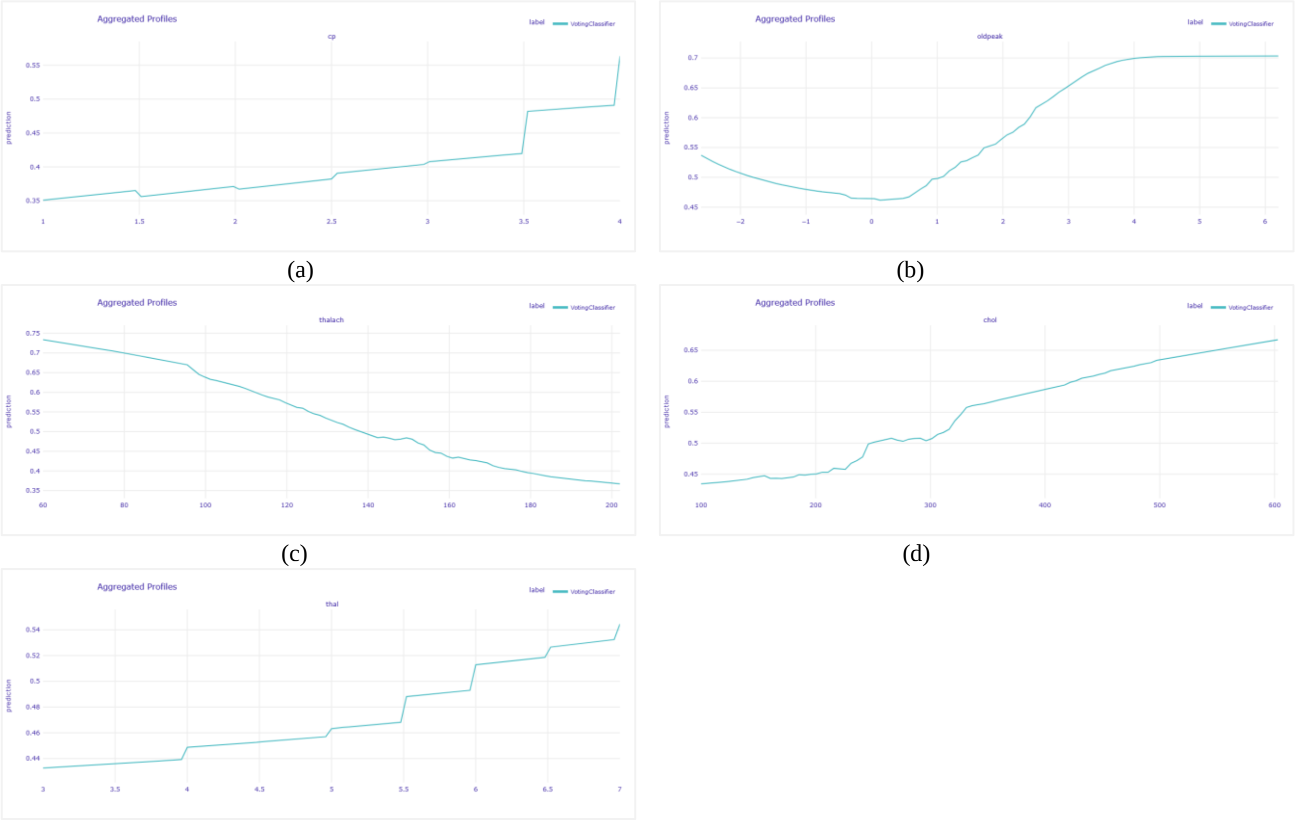


Fig. 8. Distribution of thalassemia (thal) by heart disease status



(a)

(b)

(c)

(d)

(e)

Fig. 9. Partial Dependency Plot for top 5 important features, (a) cp, (b) oldpeak, (c) thalach, (d) chol, and (e) thal

important for the model's predictions. Although fbs and restecg have strong correlations with the binary label, they are not in the top 10 important features. The strong correlations and high importance of key features help the model perform better in binary classification compared to multi-la
bel classification.

Naturally, when it comes to deadly disease, people want to see the types of people who are at potential risk for heart disease. For efficiency, top 5 important features (according to Fig. 5) are observed deeper to gain more insight.

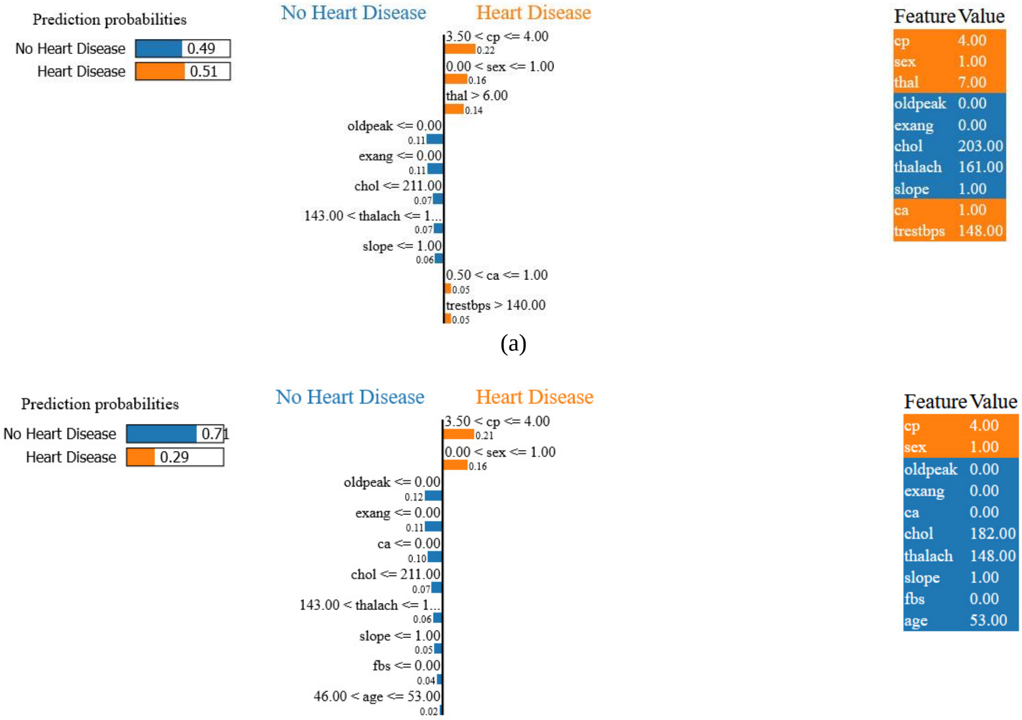
The plot in Fig. 6. shows that the distribution of chest pain types by heart disease status shows that individuals with asymptomatic chest pain (cp = 3) have a significantly higher prevalence of heart disease. In contrast, those with typical angina (cp = 0), atypical angina (cp = 1), and non-anginal pain (cp = 2) have a lower prevalence, with a more balanced or even higher number of individuals without heart disease. This highlights the importance of monitoring asymptomatic individuals closely, as they are at a higher risk of heart disease despite the lack of typical chest pain symptoms.

TABLE 3. THE MINIMUM, AVERAGE, AND MAXIMUM VALUES OF TALACH AND CHOL FOR PEOPLE WITH HEART DISEASE

Attributes	**Min**	**Avg**	**Max**
thalach	60	130.23	195
chol	117	254.35	603

The plot in Fig. 7. shows that an oldpeak value of 0.0 is common in both groups. However, higher oldpeak values (0.9, 1.0, 1.1, 2.0, 3.0) are more frequent in individuals with heart disease. This indicates that greater ST depression induced by exercise is a significant indicator of heart disease, emphasizing the importance of monitoring oldpeak values in heart disease assessment.

Due to the extensive range of values for thalach and chol, their visualization is presented in Table 3. This table details the minimum, average, and maximum values of thalach (maximum heart rate achieved) and chol (serum cholesterol in mg/dl) for individuals with heart disease. In people with heart disease, thalach ranges from 60 to 195, averaging 130.23. chol ranges from 117 to 603, averaging 254.35. For comparison, in



(a)

(b)

Fig. 10. LIME explanation of the feature importance in classification for (a) positive class and (b) negative class

healthy people, normal resting heart rate is 60-100 bpm, with a maximum typically below 195 [33]. Normal cholesterol levels are below 200 mg/dl [34]. The variability in heart rate and cholesterol among heart disease patients, highlighting the need to monitor these factors for effective assessment and management.

The plot shown in Fig. 8. shows how many people fall into different thalassemia categories (3 = normal, 6 = fixed defect, 7 = reversible defect), and whether they have heart disease (orange) or not (blue). There is an unusual value of 4.985, which considered as unknown or no thalassemia. The plot indicates that people with thalassemia values of 6 and 7 are more likely to have heart disease compared to those with a value of 3, suggesting a higher risk of heart disease for fixed and reversible defects.

Fig. 9. shows all dependency plot extracted using DALEX. Figure 9.a. shows how the prediction of heart disease varies with chest pain type (cp). As the chest pain type increases from 1 to 4, the probability of heart disease also rises, with a significant increase between types 3 and 4, indicating higher severity of chest pain is linked to a higher likelihood of heart disease. Plot 9.b. demonstrates that higher oldpeak values lead to higher predictions for heart disease. This suggests that greater ST depression during exercise is a significant indicator of heart disease risk. The plot for thalach (9.c.) shows a clear decreasing trend. Higher maximum heart rates are associated with lower predictions for heart disease, implying that better cardiovascular fitness (indicated by higher heart rates) reduces heart disease risk.

The cholesterol plot indicates that as cholesterol levels increase, so does the prediction probability for heart disease. Higher cholesterol levels are a well-known risk factor for heart disease, and the model's predictions align with this clinical understanding. The plot for thalassemia values shows that as the severity of thalassemia increases (from 3 to 7), the model predicts a higher likelihood of heart disease. This is consistent with clinical knowledge that more severe thalassemia conditions are linked to higher heart disease risk.

Lastly, this study uses LIME to explain feature importance in classification. LIME visualizes the contribution of each feature to the model's prediction for specific instances. The value associated with each feature represents its impact on heart disease classification. In both images shown in Fig. 10., the prediction probabilities indicate the model's confidence in predicting heart disease versus no heart disease. The bars next to each feature represent the contribution of that feature to the prediction. Orange bars indicate features that contribute to the prediction of heart disease, while blue bars indicate features that contribute to the prediction of no heart disease. The feature values listed on the right show the actual values of those features for the instance being explained. The first instance is predicted to have a slightly higher probability of heart disease (0.51), while the second instance is predicted to have a higher probability of no heart disease (0.71).

V. Conclussions

This study compares methods for data imputation to find the best parameters for classifying heart disease using ensemble methods, following Doppala et al. Three experiments were conducted: removing missing values, mean imputation, and safe-region imputation. Binary classification consistently outperformed 5-label classification, with the safe-region imputation yielding the best results. Data visualization method and explainable AI are applied to gain more insight from dataset.

From the analysis, it can be concluded that people who have the risk of getting heart disease are those who have asymptomatic chest pain, exhibit higher oldpeak values (0.9 and above), indicating significant ST depression induced by exercise, show high variability in maximum heart rate (thalach), ranging from 60 to 195, with an average of 130.23, have serum cholesterol (chol) levels ranging from 117 to 603 mg/dl, with an average of 254.35 mg/dl, have thalassemia values of 6 (fixed defect) or 7 (reversible defect).

For future research, it would be useful to try other advanced methods for handling missing data and see how they affect model performance. Testing different machine learning algorithms or combining multiple methods might also improve accuracy. Including more diverse patient data could make the results more applicable to different groups. Using more detailed explainable AI techniques can help us understand how the model makes decisions, which is important for clinical use. Additionally, long-term studies could show how well these models work in real-world healthcare, helping with better diagnosis and management of heart disease.

References

[1] World Health Organization, "Cardiovascular Diseases." [Online]. Available: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

[2] C. W. Tsao *et al.*, "Heart Disease and Stroke Statistics—2022 Update: A Report From the American Heart Association," *Circulation*, vol. 145, no. 8, Feb. 2022, doi: 10.1161/CIR.0000000000001052.

[3] E. Michniewicz, E. Mlodawska, P. Lopatowska, A. Tomaszuk-Kazberuk, and J. Malyszko, "Patients with atrial fibrillation and coronary artery disease–double trouble," *Advances in Medical Sciences*, vol. 63, no. 1, pp. 30–35, 2018.

[4] K. Okrainec, D. K. Banerjee, and M. J. Eisenberg, "Coronary artery disease in the developing world," *American Heart Journal*, vol. 148, no. 1, pp. 7–15, 2004.

[5] D. Arzamendi, B. Benito, H. Tizon-Marcos, and others, "Increase in sudden death from coronary artery disease in young adults," *American Heart Journal*, vol. 161, no. 3, pp. 574–580, 2011.

[6] F. F. Ozair, N. Jamshed, A. Sharma, and P. Aggarwal, "Ethical issues in electronic health records: A general overview," *Perspectives in Clinical Research*, vol. 6, pp. 73–76, 2015.

[7] S.-F. Huang and C.-H. Cheng, "A Safe-Region Imputation Method for Handling Medical Data with Missing Values," *Symmetry*, vol. 12, no. 11, p. 1792, Oct. 2020, doi: 10.3390/sym12111792.

[8] R. K. Bania and A. Halder, "R-Ensembler: A greedy rough set based ensemble attribute selection algorithm with kNN imputation for classification of medical data," *Computer Methods and Programs in Biomedicine*, vol. 184, p. 105122, 2020.

[9] J. A. C. Sterne *et al.*, "Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls," *BMJ*, vol. 338, p. b2393, 2009.

[10] A. Purwar and S. K. Singh, "Hybrid prediction model with missing value imputation for medical data," *Expert Systems with Applications*, vol. 42, pp. 5621–5631, 2015.

[11] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," *IEEE Access*, vol. 8, pp. 107562–107582, 2020, doi: 10.1109/ACCESS.2020.3001149.

[12] M. Elhoseny *et al.*, "A New Multi-Agent Feature Wrapper Machine Learning Approach for Heart Disease Diagnosis," *Computers, Materials & Continua*, vol. 67, no. 1, pp. 51–71, 2021, doi: 10.32604/cmc.2021.012632.

[13] U. Nagavelli, D. Samanta, and P. Chakraborty, "Machine Learning Technology-Based Heart Disease Detection Models," *Journal of Healthcare Engineering*, vol. 2022, pp. 1–9, Feb. 2022, doi: 10.1155/2022/7351061.

[14] R. C. Das, M. C. Das, Md. A. Hossain, Md. A. Rahman, M. H. Hossen, and R. Hasan, "Heart Disease Detection Using ML," in *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*, Las Vegas, NV, USA: IEEE, Mar. 2023, pp. 0983–0987. doi: 10.1109/CCWC57344.2023.10099294.

[15] B. P. Doppala, D. Bhattacharyya, M. Janarthanan, and N. Baik, "A Reliable Machine Intelligence Model for Accurate Identification of Cardiovascular Diseases Using Ensemble Techniques," *Journal of Healthcare Engineering*, vol. 2022, pp. 1–13, Mar. 2022, doi: 10.1155/2022/2585235.

[16] T. Bayu Adhi, I. Sun, and L. Seungual, "Improving an intelligent detection system for coronary heart disease using a two-tier classifier ensemble," *BioMed Research International*, vol. 2020, p. 10 pages, 2020, doi: 10.1155/2020/9816142.

[17] R. Prasad, R. C. Deo, Y. Li, and T. Maraseni, "Weekly soil moisture forecasting with multivariate sequential, ensemble empirical mode decomposition and Boruta-random forest hybridizer algorithm approach," *CATENA*, vol. 177, pp. 149–166, 2019.

[18] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[19] B. Boehmke and B. Greenwell, "Random Forests," in *Hands-On Machine Learning with R*, Chapman and Hall/CRC, 2019, pp. 203–219. doi: 10.1201/9780367816377-11.

[20] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," *COLT '92: Proceedings of the fifth annual workshop on Computational learning theory*, Jul. 1992.

[21] S. Karamizadeh, S. M. Abdullah, M. Halimi, J. Shayan, and M. javad Rajabi, "Advantage and drawback of support vector machine functionality," in *2014 International Conference on Computer, Communications, and Control Technology (I4CT)*, Langkawi, Malaysia: IEEE, Sep. 2014, pp. 63–65. doi: 10.1109/I4CT.2014.6914146.

[22] T. Joachims, "Text categorization with Support Vector Machines: Learning with many relevant features," in *Machine Learning: ECML-98*, vol. 1398, C. Nédellec and C. Rouveirol, Eds., in Lecture Notes in Computer Science, vol. 1398. , Berlin, Heidelberg: Springer Berlin Heidelberg, 1998, pp. 137–142. doi: 10.1007/BFb0026683.

[23] W. S. Noble, "What is a support vector machine?," *Nat Biotechnol*, vol. 24, no. 12, pp. 1565–1567, Dec. 2006, doi: 10.1038/nbt1206-1565.

[24] Y. Lin, "Support Vector Machines for Classification in Nonstandard Situations," *SUPPORT VECTOR MACHINES*, p. 12.

[25] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques 3rd Edition*, 3rd ed. Elsevier, 2011.

[26] Suyanto, *Machine Learning Tingkat Dasar dan Lanjut*. Informatika Bandung.

[27] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[28] S. Das, M. Sultana, S. Bhattacharya, D. Sengupta, and D. De, “XAI–reduct: accuracy preservation despite dimensionality reduction for heart disease classification using explainable AI,” *J Supercomput*, vol. 79, no. 16, pp. 18167–18197, Nov. 2023, doi: 10.1007/s11227-023-05356-3.

[29] M. B. Kursa and W. R. Rudnicki, “Feature Selection with the Boruta Package,” *Journal of Statistical Software*, vol. 36, no. 11, pp. 1–13, 2010.

[30] Janosi Andras, Steinbrunn, William, Pfisterer, Matthias and R. Detrano, “Heart Disease.” 1988.

[31] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[32] T. Chen and C. Guestrin, *XGBoost Documentation*. 2023. [Online]. Available: https://xgboost.readthedocs.io/en/stable/

[33] Heart Foundation, “How to Check Your Pulse (Heart Rate).” [Online]. Available: https://www.heartfoundation.org.nz/wellbeing/managing-risk/how-to-check-your-pulse-heart-rate

[34] Mayo Clinic, “High Cholesterol - Diagnosis and Treatment.” [Online]. Available: https://www.mayoclinic.org/diseases-conditions/high-blood-cholesterol/diagnosis-treatment/drc-20350806