**IJCCS (Indonesian Journal of Computing and Cybernetics Systems)**
Vol.19, No.3, July 2025, pp. 247~258
ISSN (print): 1978-1520, ISSN (online): 2460-7258
DOI: https://doi.org/10.22146/ijccs.92395


# Classifying Heart Disease through Fusion of Multi-Source Datasets: Integration of Feature Selection and Explainable Machine Learning Techniques

**Kasiful Aprianto*[1], Mila Desi Anasanti[2]**
[1,2]Computer Science Master's Study Program, Nusa Mandiri University, Jakarta, Indonesia
e-mail:***[1]14220022@nusamandiri.ac.id**, *[2]mila.mld@nusamandiri.ac.id

***Abstrak***

*Penelitian ini mengkaji klasifikasi penyakit jantung melalui seleksi fitur terintegrasi dan metodologi pembelajaran mesin, menggunakan tiga set data yang terdiri dari 4.728 partisipan dan 11 fitur, dengan 4,27% data yang hilang. Dengan menggunakan pembelajaran mesin, kami menggunakan XGBoost untuk mencapai akurasi 0,95 untuk satu fitur, sementara Random Forest (RF) menunjukkan akurasi 0,92 dan 0,99 untuk dua fitur yang tersisa. Dalam membandingkan 11 model klasifikasi, RF dan XGBoost mengklasifikasikan penyakit jantung dengan akurasi 0,97 dan 0,99, masing-masing, menggunakan semua fitur yang tersedia. Penerapan Eliminasi Fitur dengan Seleksi dan Peringkat Fitur Simultan Perturbation (SpFSR) mengungkapkan bahwa RF mencapai akurasi 0,99 dengan memilih hanya empat fitur (tingkat kolesterol, usia, pengukuran elektrokardiografi istirahat, dan denyut jantung maksimum), sementara XGBoost turun menjadi 0,91. Pembuatan model RF dengan empat fitur meningkatkan interpretabilitas tanpa mengorbankan akurasi. Teknik Pembelajaran Mesin yang Dapat Dijelaskan (XAI), termasuk Permutation Importance dan analisis SHAP Summary Plot, mengukur dampak fitur pada prediksi penyakit jantung. Fitur pengukuran elektrokardiografi istirahat memiliki nilai tertinggi (0,40 ± 0,01), diikuti oleh denyut jantung maksimum (0,32 ± 0,01), tingkat kolesterol (0,28 ± 0,01), dan usia (0,26 ± 0,005). Hasil ini menekankan pentingnya masing-masing fitur dalam mendiagnosis penyakit jantung melalui pembelajaran mesin.*

***Kata kunci**— Klasifikasi penyakit jantung, Penggabungan dataset, Pengisian nilai yang hilang, Pembelajaran mesin, Ekstraksi fitur, Machine Learning yang dapat dijelaskan, XGBoost, Random Forest.*

***Abstract***

*This study delves into heart disease classification through integrated feature selection and machine learning methodologies, utilizing three datasets comprising 4,728 participants and 11 features, with 4.27% missing data. Employing machine learning, we used XGBoost to achieve 0.95 accuracy for one feature, while Random Forest (RF) demonstrated accuracies of 0.92 and 0.99 for the remaining two features. Comparing 11 classification models, RF and XGBoost classified heart disease with 0.97 and 0.99 accuracy, respectively, using all available features. Applying Feature Elimination with Simultaneous Perturbation Feature Selection and Ranking (SpFSR) revealed that RF attained 0.99 accuracy by selecting only four features (cholesterol level, age, resting electrocardiographic measurements, and maximum heart rate), while XGBoost dropped to 0.91. Constructing an RF model with four features enhanced interpretability without compromising accuracy. Explainable Machine Learning (XAI) techniques, including Permutation Importance and SHAP Summary Plot analyses, gauged feature impact on heart disease prediction. The resting electrocardiographic measurements feature held the highest value (0.40 ± 0.01), followed by maximum heart rate (0.32 ± 0.01), cholesterol level (0.28 ± 0.01), and age*

*Received December 24th,2024; Revised January 22th, 2025; Accepted July 31th, 2025*

*(0.26 ± 0.005). These results underscore the significance of each feature in diagnosing heart disease via machine learning.*
***Keywords—*** *Heart disease classification, Dataset fusion, Imputation, Machine learning, Feature extraction, Explainable Machine Learning, XGBoost, Random Forest.*

## 1. INTRODUCTION

Noncommunicable diseases (NCDs) account for 75% of deaths worldwide, particularly prevalent in developing regions like South Asia and Sub-Saharan Africa [1]. NCD prevention involves managing risk factors for obesity, diabetes, and hypertension while encouraging healthy lifestyle practices such as physical activity, smoking cessation, balanced nutrition, and responsible alcohol consumption. According to WHO reports from 2020, Ischemic heart disease and stroke account for approximately 15% of global mortality and were the two primary causes of death [1]. Cardiovascular disease (CVD) continues to affect more than 500 million individuals globally and will account for nearly one-third of global deaths by 2021 [2]. Though CVD mortality rates have been declining over time, progress has slowed dramatically in low-income and middle-income nations [2]. Access and affordability remain critical barriers to heart disease diagnosis and could account for up to 17 million deaths [3]. CVD expenses account for 25-30% of annual medical costs within an organization [4]. Early diagnosis is critical to mitigating both physical and financial burdens associated with heart disease; WHO estimates project 23.6 million CVD deaths globally by 2030 [5].

Machine learning techniques have been employed to predict CVD development. Numerous studies have explored these techniques; Shorewala et al. provided one such example [3], which achieved an accuracy rate of approximately 75.1% using a Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) combined into a stacked model. Maiga et al.[6] attained approximately 70% accuracy using four algorithms such as RFt, Naive Bayes, KNN, and Logistic Regression (LR). Waigi et al. [7] achieved similar success (70%) using RF, while Decision Tree (DT) was employed with a 72.77% accuracy rate. Khan and Mondal [8] utilized various algorithms, such as SVM, Neural Networks (NN), and LR, with success ranging from 71.82%-72.72% accuracy across different datasets. Maini E. et al. also applied algorithms with similar results on several datasets. Maini et al. and Kavitha et al. reported an impressive accuracy of 90.74% using artificial neural networks (ANNs) in the Cleveland Heart Disease dataset [9,10]. Using the same dataset, Kavitha et al. [10] applied a "Hybrid Model," composed of DT and RF models, to achieve approximately 88.70%. Shah, D. et al. used multiple algorithms such as KNN, reaching 90.79% accuracy, while Bharti R. et al.[12] applied Deep Learning (DL) directly onto The University of California, Irvine (UCI) datasets, achieving 94.20% precision.

Building upon machine learning's demonstrated efficacy in heart disease prediction, this study integrates data from diverse sources for comprehensive risk forecasting. Employing feature selection and machine learning strategies enhances model interpretability, providing insights into predictions and alleviating financial strain associated with early detection amidst projected increases in cardiovascular deaths [1]. Utilizing three datasets and advanced imputation methods like XGBoost and Random Forest (RF), this study predicts heart disease for 4,728 individuals. The SpFSR feature selection method condenses variables to four key elements—cholesterol, age, restecg, and thalach—maintaining a remarkable 99% accuracy. Beyond simplifying the model, this reduction enhances interpretability, providing efficient diagnostic tools. The incorporation of

explainable machine learning, including summary plots and SHAP values, fortifies trust in the model's decisions, aligning with ethical considerations.

Contributing significantly, this study introduces SpFSR as a potent tool for variable reduction without compromising predictive accuracy—a noteworthy departure from conventional approaches. It accentuates the applicability and interpretability of predictive models, underscoring the importance of accurate and understandable predictions amidst global cardiovascular health challenges. Addressing a critical gap, the research explicitly presents diverse methodologies employed in cardiovascular disease prediction. By drawing on three distinct datasets and employing advanced imputation methods, it establishes a comprehensive basis for methodological comparison. Notably, the SpFSR method contributes to methodological diversity by reducing variables to a minimal set while maintaining a 99% accuracy rate—an innovative perspective for researchers considering feature selection techniques. This commitment to methodological transparency, augmented by summary plots and SHAP values, fosters a culture of rigorous scrutiny and refinement in cardiovascular disease prediction research.

## 2. METHODS

### *2.1 Datasets and Data Preprocessing*

This study utilized three distinct datasets for forecasting heart disease incidence. The initial dataset originated from the well-established Heart Disease Data Set [13], the second dataset was procured from IEEEDataPort [14], and the third dataset was secured from Kaggle, identified as the Heart Disease Prediction dataset [15]. The combined dataset involved 4,728 individuals and encompassed 14 characteristics, 13 attribute values, and a target variable. Parameters in the Heart Disease dataset include **age** (expressed in years), **sex** (0 for female and 1 for male), chest pain (**cp**) categorized as 0 for typical angina, 1 for atypical angina, 2 for non-anginal pain, and 3 for asymptomatic. Other parameters cover resting blood pressure (**treetops**), serum cholesterol level (**chol**), fasting blood sugar levels (**fbs**), resting electrocardiographic measurements (**restecg**), maximum heart rate achieved during testing (**thalch**), exercise-induced angina (**exang**), ST depression induced by exercise relative to rest (**oldpeak**), slope of the peak exercise ST segment (**slope**), number of major vessels (**ca**), thalassemia (**thal**), and **target**, where 0 indicates no heart disease, and 1 signals the likelihood of developing heart disease.

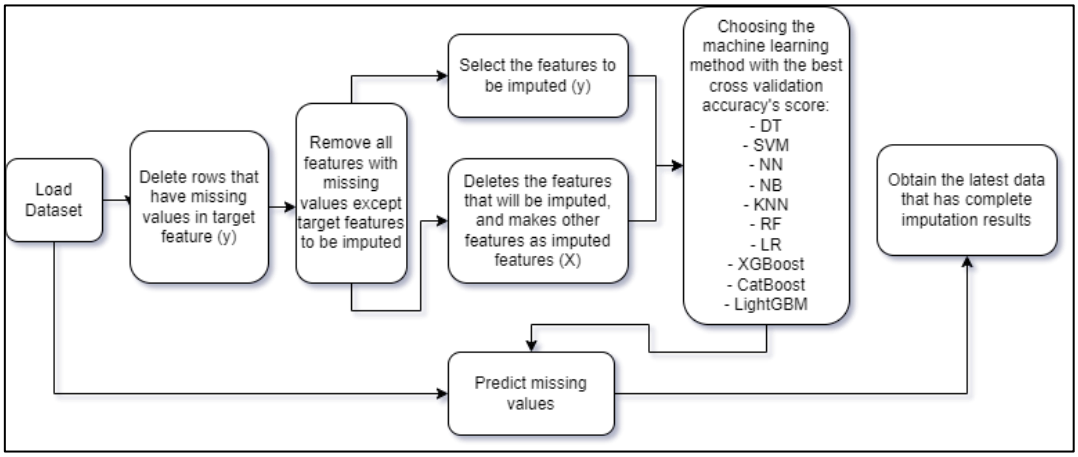


Figure 1 Missing value imputation algorithm using machine learning methods

After our analysis was complete, we employed a range of machine learning algorithms, including decision trees (DT), support vector machines (SVM), random forests (RF), neural

networks (NN), naive bayes (NB), KNN classifier, logistic regression (LR), XGBoost, CatBoost, and LightGBM, to predict and fill in missing values. Following cross-validation, we selected the model with the highest accuracy for imputation. Using accuracy measures, we assessed the success of our machine learning model in replacing missing information with artificial imputation. All these stages can be explained in Figure 1. This comprehensive approach resulted in a complete dataset, enhancing data quality and enabling more precise and relevant analyses for cardiovascular disease examination.

*2.2 Feature Elimination with Simultaneous Perturbation Feature Selection and Ranking (SpFSR)*

Feature selection can help maximize computational resources, reduce data acquisition and storage costs, ensure compliance with feature constraints, and strengthen model robustness. Simultaneous Perturbation Feature Selection and Ranking (SpFSR) is an advanced feature selection and ranking technique in predictive modeling [16]. SpFSR begins with the initial weight $w_0$ and there is a recursion to find the local minimum $\hat{w}$ as shown in equation (1).

$$\hat{w}_k + 1 \coloneqq \hat{w}_k - a_k \hat{G}(\hat{w}_k) \quad (1)$$

where $a_k$ is the order of iteration gain; $a_k \geq 0$ and $\hat{G}(\hat{w}_k)$ are estimates of the gradient at k.

SpFSR research has proven its worth across multiple domains, from cardiovascular disease prediction and Autism Spectrum Disorder diagnosis [17, 18] to spinal cord injuries, where SpFSR feature selection research is essential in predictive modeling [20-22].

*2.3 Classification Models*

Classification is a crucial aspect of applying machine learning techniques for disease diagnosis. We provide an overview of operational methods employed by DT, RF, SVM, NN, LR, NB, KNN, XGBoost, CatBoost, and LightGBM classifiers to classify models that best suit this task.

**DT** algorithm creates tree-like structures to represent decision-making processes through hierarchical rules; this highly interpretable approach is a valuable way of exploring factors impacting heart disease incidence by splitting datasets according to specific attributes and building decision trees [19,20]. Let's consider a classification problem with $\boldsymbol{K}$ classes. We aim to classify an instance with features $\boldsymbol{X} = (\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n})$ into one of the $\boldsymbol{K}$ classes. At each node of the tree, we make a decision based on a feature test $\boldsymbol{f_i}(\boldsymbol{X})$ and its associated threshold as shown in equation (2):

$$Node\ i\text{: } (f_i(X) <= \theta_i) \quad (2)$$

Where $\boldsymbol{Node\ i}$ represents an internal node in the tree, $\boldsymbol{f_i}(\boldsymbol{X})$ is a feature test on one of the features $(\boldsymbol{x_i})$ in the dataset, and $\boldsymbol{\theta_i}$ is the threshold associated with the feature test. DT uses entropy to measure the level of impurity disorder in a set of items. The formula is explained in equation (3):

$$Entropy(p) = -\Sigma\ (p_i * log2(p_i)) \quad (3)$$

where $Entropy(p)$ is the entropy for a set of items $p$, and $p_i$ is the probability of an item belonging to class i out of all possible classes.

**RF** is an ensemble method composed of multiple DTs that collectively reduces overfitting potential and improves classification accuracy. While individual decision trees tend to be straightforward to interpret, random forests present additional challenges when simultaneously dealing with multiple decision trees. Yet, this additional difficulty often improves predictive performance, making RF highly effective tools for various prediction tasks [20].

**SVM** is an algorithm created to find an optimal hyperplane for class separation in any dataset, making it particularly suitable for those featuring defined class boundaries. SVM employs kernel

functions when mapping data into higher dimensional spaces to differentiate classes further and help increase differentiation. The formula is shown in equation (4):

$$H: sign(\boldsymbol{w^T}(\boldsymbol{x}) + b) \quad (4)$$

where $b$ is the bias term and intercept if the hyerplane equation. Furthermore, employing structural risk minimization principles rather than empirical risk minimization principles makes SVM suitable for fitting models onto small datasets [21].

**NN** is a model that mimics the function and structure of human neurons, similar to how our brain operates. The formula is explained in equation (5):

$$z_j = \sum (w_{ij} * a_i) + b_j ; \quad a_i = f(z_i) \quad (5)$$

where $z_j$ is the input to neuron $j$ in the hidden layer, $w_{ij}$ is the weight connecting neuron $i$ in the previous layer to neuron $j$ in the hidden layer, $a_i$ is the output from neuron $i$ in the previous layer, $b_i$ is the bias of neuron $j$ in the hidden layer, and $f$ is the activation function used such as sigmoid, ReLu, and so on.

**NB** is a probabilistic algorithm that excels at handling datasets with categorical attributes based on Bayes' Theorem. Equation (6) explains how NB works based on Bayes' Theorem. NB stands out not only as an efficient and effective classification algorithm but also for its elegant simplicity and apparent effectiveness, even if its independence assumptions aren't fully fulfilled [22].

$$P(c_i|X) = \frac{P(X|c_i) * P(c_i)}{P(X)} \quad (6)$$

where $P(c_i \mid X)$ is the posterior probability of class $c_i$ given features $X$, $P(X \mid c_i)$ is the likelihood of observing features $X$ given class $c_i$, $P(c_i)$ is the prior probability of class $c_i$, representing the likelihood of a random instance belonging to class $c_i$ without considering the features. $P(X)$ is the evidence probability, a normalization constant ensuring that the probabilities sum to 1 over all classes.

**LR** is a statistical technique that models relationships between independent and binary target variables using log odds probability. A logistic function transforms linear possibilities to logit probabilities, making this technique particularly suitable for classification challenges rather than regression scenarios [23]. The LR formula is explained in equation (7):

$$P(Class\ 1|X) = \frac{1}{(1 + e^{-z})}; z = b_0 + (b_1 * x_1) + (b_2 * x_2) + \cdots + (b_n * x_n) \quad (7)$$

where $P(Class\ 1 \mid X)$ is the probability that the instance belongs to Class 1 given the features $X$, $e$ is the base of the natural logarithm (approximately 2.7183), and z is the linear combination of the features and their associated weights: $b_0$ is the bias term (intercept); $b_1, b_2, \ldots, b_n$ are the coefficients (weights) associated with each input features $x_1, x_2, \ldots, x_n$.

**KNN**: Classification decisions are determined based on which class has majority membership among its nearest neighbors. If one data point has more neighbors from one class than another, it will be classified accordingly. Every new input data point is assigned its class with the highest number of nearest neighbors - representing its closest match from that dataset [24].

**XGBoost** is an efficient gradient-boosting algorithm rooted in decision trees that performs well on large or complex datasets, using multiple decision trees to increase classification accuracy. Gradient Boosting Machine (GBM), however, remains one of the leading artificial intelligence techniques thanks to its seamless parallel processing capability and superior predictive accuracy [25]. **CatBoost** is an efficient gradient-boosting algorithm tailored to categorical datasets that solve preprocessing issues associated with them, making their analysis much simpler [26].

**LightGBM** is an efficient gradient-boosting algorithm based on decision trees that deliver fast performance for large datasets [27].

*2.4 Evaluation*

The optimal performance of the machine learning technique with all features was assessed by comparing the accuracy of each method, using a formula that includes True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). To minimize variability and optimize computational time, a stratified 10-fold cross-validation method for performance evaluation was employed and the process was repeated three times. The performance matrix scores are calculated using accuracy, precision, recall, and F1-score that can be explained in equation (8), equation (9), equation (10), and equation (11):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

$$Precision = \frac{TP}{TP + FP} \quad (9)$$

$$Recall = \frac{TP}{TP + FP} \quad (10)$$

$$F1\ Score = 2 * \frac{Recall \text{ x } Precision}{Recall + Precision} \quad (11)$$

*2.5 Explainable Machine Learning*

Explainable machine learning aims to make machine learning models more transparent and understandable. It addresses the complexity of models like deep neural networks, providing insights into decision-making processes [28]. This transparency is crucial for trust-building between machine learning systems and users and has found applications in healthcare and other fields [31-34]. Key techniques in explainable machine learning include feature importance scores and SHapley Additive ExPlanations (SHAP). Feature importance scores help assess variable importance in models [36], while SHAP offers explanations for predictions by quantifying each feature's impact [37]. SHAP's versatility extends to various models and applications [38, 39]. Explainable machine learning enhances model interpretability and trust, benefiting multiple domains.

## 3. RESULTS AND DISCUSSION

*3.1 Missing Value Imputation*

We evaluate the percentages of missing values in the dataset's features. For instance, the "cp" feature has 143 missing data (3,02% of the total), "exang" has 48 missing data (1,02%), and "slope" has 22 missing data (0,46%). However, "sex", "age", "chol", "trestbps", "oldpeak", "restecg", "fbs", "thalach", and "target" features have no missing data (0%). The process of handling missing values is explained in Figure 1.

The most effective model for addressing missing values in the 'cp' variable is XGBoost, achieving a score of 0.949, closely followed by RF with a score of 0.946. In handling missing values in the 'exang' variables, both RF and XGBoost demonstrated superior performance, scoring 0.994, with DT (0.986) and LightGBM (0.982) following closely. Regarding imputing missing values for the 'slope' variable, RF emerged as the top-performing model, scoring 0.922, while XGBoost followed closely with a score of 0.919. Hence, XGBoost and RF were robust models for missing values in the 'cp' variable. At the same time, RF excelled in addressing missing values in the 'slope' and 'exang' variables. The Results Selection of the best model for feature imputation is shown in Table 1.

Table 1 Results Selection of the best model for feature imputation

| Model | Accuracy | | |
|---|---|---|---|
| | CP | EXANG | SLOPE |
| CatBoost | 0.92 | 0.98 | 0.90 |
| Decision Tree | 0.93 | 0.99 | 0.90 |
| KNN Classifier | 0.74 | 0.88 | 0.75 |
| LightGBM | 0.93 | 0.98 | 0.89 |
| Logistic Regression | 0.44 | 0.73 | 0.57 |
| Neural Network | 0.48 | 0.76 | 0.63 |
| Naive Bayes | 0.41 | 0.71 | 0.53 |
| Random Forest | 0.95 | 0.99 | 0.92 |
| Support Vector Machine | 0.48 | 0.75 | 0.58 |
| XGBoost | 0.95 | 0.99 | 0.92 |

*3.2 Feature Elimination and Classification Model Selection4*

Table 2 The accuracy values of each model using all available features

| Model | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| CatBoost | 0.98 | 0.98 | 0.98 | 0.98 |
| Decision Tree | 0.97 | 0.98 | 0.97 | 0.98 |
| KNN Classifier | 0.95 | 0.96 | 0.94 | 0.95 |
| LightGBM | 0.97 | 0.98 | 0.97 | 0.97 |
| Logistic Regression | 0.76 | 0.77 | 0.77 | 0.77 |
| Naive Bayes | 0.71 | 0.79 | 0.61 | 0.69 |
| Neural Network | 0.82 | 0.82 | 0.84 | 0.83 |
| Random Forest | 0.99 | 0.99 | 0.98 | 0.99 |
| Support Vector Machine | 0.81 | 0.82 | 0.82 | 0.82 |
| XGBoost | 0.99 | 0.99 | 0.98 | 0.99 |

In this study, an analysis was conducted on various classification models used to identify the risk of heart disease. Initially, the performance of the models was evaluated using all available features. The results in Table 2 showed that two models, RF and XGBoost, achieved the highest accuracy, almost 0.99.

Table 3 The accuracy values of the Random Forest models using the feature selection results from SpFSR using the Random Forest Wrapper

| Number of features | List Feature | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|---|
| 2 | slope, cp | 0.75 | 0.73 | 0.82 | 0.77 |
| 3 | slope, chol, cp | 0.87 | 0.89 | 0.86 | 0.87 |

| Number of features | List Feature | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|---|
| 4 | restecg, chol, thalach, age | 0.99 | 0.99 | 0.99 | 0.99 |
| 5 | cp, restecg, thalach, chol, trestbps | 0.99 | 0.99 | 0.98 | 0.99 |
| 6 | cp, slope, thalach, chol, trestbps, restecg | 0.98 | 0.99 | 0.97 | 0.98 |
| 7 | cp, chol, thalach, slope, trestbps, age, restecg | 0.98 | 0.99 | 0.98 | 0.98 |

Table 4 The accuracy values of the XGBoost models using the feature selection results from SpFSR using XGBoost Wrapper

| Number of features | List Feature | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|---|
| 2 | slope, cp | 0.75 | 0.73 | 0.82 | 0.77 |
| 3 | cp, slope, thalach | 0.84 | 0.84 | 0.84 | 0.84 |
| 4 | slope, restecg, cp, chol | 0.91 | 0.91 | 0.91 | 0.91 |
| 5 | slope, cp, restecg, thalach, chol | 0.97 | 0.98 | 0.96 | 0.97 |
| 6 | cp, slope, thalach, chol, trestbps, restecg | 0.98 | 0.99 | 0.98 | 0.98 |
| 7 | cp, restecg, chol, thalach, trestbps, slope, oldpeak | 0.98 | 0.99 | 0.98 | 0.98 |

In this study, feature selection using the SpFSR method identified the most critical subset of features to enhance model performance. The results showed that by selecting only four features (chol, age, restecg, and thalach), the RF model achieved an accuracy of 0.99 (Table 3), while XGBoost reached an accuracy of 0.91 (Table 4). This suggests that RF remains the preferred model after feature reduction. Additionally, this RF model with four features can provide highly accurate predictions of heart disease risk with straightforward interpretation of results, providing valuable guidance in developing more efficient diagnostic tools to assess this risk. Furthermore, findings from this research could prove instrumental in creating more effective diagnostic tools.

Reducing the features to these four variables significantly impacts model interpretability. Here's how they can be explained: (chol) Blood cholesterol levels provide essential data for assessing cardiovascular disease risks; (Age) is a significant risk factor; (Restecg) results offer insights into the heart's condition at rest, and; (ThalaCh) measures the maximum heart rate during stress testing, offering a key indicator of overall cardiovascular health. Moreover, this RF model with four features can make highly accurate predictions of heart disease risk with easily interpretable results.

*3.3 Explainable Machine Learning with Permutation Importance and SHAP Summary Plot*

Permutation importance is an invaluable tool in feature analysis, quantifying each feature's importance in accurate predictions by tracking how feature changes impact model performance. Randomly permuting feature values calculate it and note any significant variance in model performance - with greater variance indicating increased significance for accurate predictions. Based on permutation importance results for particular features, **Restecg** stands out with its exceptional average permutation significance of 0.4021 and low standard deviation of 0.0087, signaling its significant effect on model performance and vulnerability to random permutations. **Thalach** also makes an impactful statement with a permutation importance of roughly 0.3245, making him sensitive to changes that could impede model performance. Although

his influence is less profound, **Chol** still manages to have a meaningful permutation impact of around 0.2813, significantly altering prediction accuracy. Finally, **Age** significantly identifies cardiovascular risks with an average permutation importance value of 0.2577, further validating the previously selected features. Restecg and Thalach mainly contribute towards increasing the RF model's ability to detect heart disease risks.

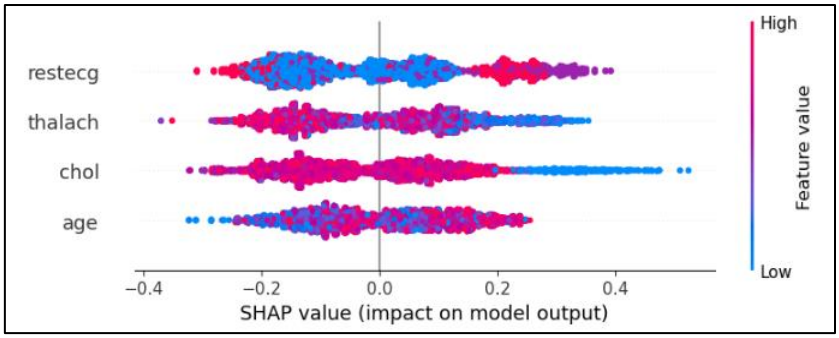


Figure 2 SHAP Summary Plot

In Figure 2, the SHAP Summary plot provides insights into feature impacts on predictions. For 'restecg,' points scatter on both sides of the zero line, indicating mixed positive and negative impacts depending on feature values. Lower values (blue) tend to have a negative impact, while higher values (red) are more positive. 'Thalach' generally has a positive impact, with most points on the positive side and higher feature values (red) having a more substantial positive effect. 'Chol' shows positive and negative impacts, with no clear color pattern, suggesting a complex influence on predictions. 'Age' has a predominantly negative impact, as most points lie on the negative side, and higher feature values (red) intensify this negative effect.

Extensive research has examined the relationship between restecg, thalach, chol, and age in the context of heart disease. Lakatta & Levy [40] emphasized the role of arterial and cardiac aging, especially age's impact on cardiovascular disease. Zaini & Awang [41] listed major heart disease risk factors, including age, cholesterol (chol) levels, and electrocardiographic results (restecg). Kostis et al. [42] provided valuable insights into age's connection to cardiovascular health. Further studies have reinforced these findings. Jacobs et al. [43] found a link between low blood cholesterol levels and mortality, while Hedayatnia et al. [44] explored dyslipidemia's impact on cardiovascular disease events. Electrocardiograms and maximum heart rate have also been identified as crucial predictors of heart disease. Kostis et al. [42] delved into the relationship between age and heart rate in individuals without cardiovascular diseases. This body of research comprehensively underscores the significant influence of age, cholesterol, and electrocardiographic results on heart disease, offering a comprehensive understanding of factors contributing to cardiovascular well-being.

## 4. CONCLUSIONS

This study addresses the challenge of missing data in a heart disease dataset with 3.02%, 1.02% and 0.46% missing entries across three dimensions - clinical profile (cp), exang (0.01%) and slope (0.004%) missing values respectively. Through advanced imputation methods such as XGBoost's accuracy rates of 0.95 for cp and 0.99 for exang while RF recorded rates of 0.95, 0.99, and 0.92 - these results highlight their effectiveness at managing missing values while improving data quality analysis for heart disease risk analysis.

Furthermore, our research comprehensively evaluated various classification models using the SpFSR feature selection method. Initial findings included RF and XGBoost models with high accuracies of up to 0.99. Once feature selection was complete, four main features (chol, age, restecg and thalach) remained central, with RF reaching an accuracy of 0.99 while XGBoost hit

0.91. This simplified approach to four variables enhanced interpretability without jeopardizing accuracy. We employed the RF model to identify heart disease using four features - blood cholesterol ('chol'), age ('age'), resting electrocardiogram results (restecg) and maximum heart rate during stress tests (thalach). Our findings provide more efficient diagnostic tools, providing accurate yet understandable models to predict cardiovascular disease more quickly and reliably.

## REFERENCES

[1] "Global health estimates: Leading causes of death." Accessed: Oct. 19, 2023. [Online]. Available: https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death

[2] C. J. L. Murray, "The Global Burden of Disease Study at 30 years," *Nat Med*, vol. 28, no. 10, pp. 2019–2026, Oct. 2022, doi: 10.1038/s41591-022-01990-1.

[3] V. Shorewala, "Early detection of coronary heart disease using ensemble techniques," *Informatics in Medicine Unlocked*, vol. 26, p. 100655, 2021, doi: 10.1016/j.imu.2021.100655.

[4] J. Li, A. Loerbroks, H. Bosma, and P. Angerer, "Work stress and cardiovascular disease: a life course perspective," *Journal of Occupational Health*, vol. 58, no. 2, pp. 216–219, 2016, doi: 10.1539/joh.15-0326-OP.

[5] Purushottam, K. Saxena, and R. Sharma, "Efficient Heart Disease Prediction System," *Procedia Computer Science*, vol. 85, pp. 962–969, 2016, doi: 10.1016/j.procs.2016.05.288.

[6] J. Maiga, G. G. Hungilo, and Pranowo, "Comparison of Machine Learning Models in Prediction of Cardiovascular Disease Using Health Record Data," in *2019 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, Oct. 2019, pp. 45–48. doi: 10.1109/ICIMCIS48181.2019.8985205.

[7] R. Waigi, S. Choudhary, P. Fulzele, and G. Mishra, "Predicting the risk of heart disease using advanced machine learning approach," *European Journal of Molecular and Clinical Medicine*, vol. 7, pp. 1638–1645, Sep. 2020.

[8] M. Khan and M. R. Mondal, "Data-Driven Diagnosis of Heart Disease," *International Journal of Computer Applications*, vol. 176, pp. 46–54, Jul. 2020, doi: 10.5120/ijca2020920549.

[9] E. Maini, B. Venkateswarlu, and A. Gupta, "Applying Machine Learning Algorithms to Develop a Universal Cardiovascular Disease Prediction System," in *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018*, J. Hemanth, X. Fernando, P. Lafata, and Z. Baig, Eds., in Lecture Notes on Data Engineering and Communications Technologies. Cham: Springer International Publishing, 2019, pp. 627–632. doi: 10.1007/978-3-030-03146-6_69.

[10] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai, and R. S. Suraj, "Heart Disease Prediction using Hybrid machine Learning Model," in *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, Jan. 2021, pp. 1329–1333. doi: 10.1109/ICICT50816.2021.9358597.

[11] D. Shah, S. Patel, and S. K. Bharti, "Heart Disease Prediction using Machine Learning Techniques," *SN COMPUT. SCI.*, vol. 1, no. 6, p. 345, Oct. 2020, doi: 10.1007/s42979-020-00365-y.

[12] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, "Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning," *Computational Intelligence and Neuroscience*, vol. 2021, pp. 1–11, Jul. 2021, doi: 10.1155/2021/8387680.

[13] W. S. Andras Janosi, "Heart Disease." UCI Machine Learning Repository, 1989. doi: 10.24432/C52P4X.

[14] M. Siddhartha, "Heart Disease Dataset (Comprehensive)." IEEE DataPort, Nov. 05, 2020. doi: 10.21227/DZ4T-CM36.

[15] "Heart Disease Predication." Accessed: Oct. 24, 2023. [Online]. Available: https://www.kaggle.com/datasets/durgesh2050/heart-disease-predication
[16] F. H. Alfebi and M. D. Anasanti, "Improving Cardiovascular Disease Prediction by Integrating Imputation, Imbalance Resampling, and Feature Selection Techniques into Machine Learning Model," *Indonesian J. Comput. Cybern. Syst.*, vol. 17, no. 1, p. 55, Feb. 2023, doi: 10.22146/ijccs.80214.
[17] A. Novianto and M. D. Anasanti, "Autism Spectrum Disorder (ASD) Identification Using Feature-Based Machine Learning Classification Model," *Indonesian J. Comput. Cybern. Syst.*, vol. 17, no. 3, p. 259, Jul. 2023, doi: 10.22146/ijccs.83585.
[18] A. Yarahmadi *et al.*, "Curcumin attenuates development of depressive-like behavior in male rats after spinal cord injury: involvement of NLRP3 inflammasome," *J. Contemp. Med. Sci.*, vol. 8, no. 3, Jun. 2022, doi: 10.22317/jcms.v8i3.1230.
[19] P. Geurts, A. Irrthum, and L. Wehenkel, "Supervised learning with decision tree-based methods in computational and systems biology," *Mol. BioSyst.*, vol. 5, no. 12, p. 1593, 2009, doi: 10.1039/b907946g.
[20] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," *The Stata Journal*, vol. 20, no. 1, pp. 3–29, Mar. 2020, doi: 10.1177/1536867X20909688.
[21] Q. Wang, "Support Vector Machine Algorithm in Machine Learning," in *2022 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, Dalian, China: IEEE, Jun. 2022, pp. 750–756. doi: 10.1109/ICAICA54878.2022.9844516.
[22] K. M. Al-Aidaroos, A. A. Bakar, and Z. Othman, "Naïve bayes variants in classification learning," in *2010 International Conference on Information Retrieval & Knowledge Management (CAMP)*, Shah Alam, Selangor: IEEE, Mar. 2010, pp. 276–281. doi: 10.1109/INFRKM.2010.5466902.
[23] K. Siddique, Z. Akhtar, H. Lee, W. Kim, and Y. Kim, "Toward Bulk Synchronous Parallel-Based Machine Learning Techniques for Anomaly Detection in High-Speed Big Data Networks," *Symmetry*, vol. 9, no. 9, p. 197, Sep. 2017, doi: 10.3390/sym9090197.
[24] K. Taunk, S. De, S. Verma, and A. Swetapadma, "A Brief Review of Nearest Neighbor Algorithm for Learning and Classification," in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, Madurai, India: IEEE, May 2019, pp. 1255–1260. doi: 10.1109/ICCS45141.2019.9065747.
[25] 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.
[26] J. T. Hancock and T. M. Khoshgoftaar, "CatBoost for big data: an interdisciplinary review," *J Big Data*, vol. 7, no. 1, p. 94, Dec. 2020, doi: 10.1186/s40537-020-00369-8.
[27] G. Ke *et al.*, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Accessed: Nov. 04, 2023. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html
[28] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nat Mach Intell*, vol. 1, no. 5, pp. 206–215, May 2019, doi: 10.1038/s42256-019-0048-x.
[29] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A Review of Machine Learning Interpretability Methods," *Entropy*, vol. 23, no. 1, p. 18, Dec. 2020, doi: 10.3390/e23010018.
[30] U. Bhatt *et al.*, "Explainable Machine Learning in Deployment," 2019, doi: 10.48550/ARXIV.1909.06342.
[31] A. Ejmalian *et al.*, "Prediction of Acute Kidney Injury After Cardiac Surgery Using Interpretable Machine Learning," *Anesth Pain Med*, vol. 12, no. 4, Sep. 2022, doi: 10.5812/aapm-127140.

[32] K. Kobylińska, T. Orłowski, M. Adamek, and P. Biecek, "Explainable Machine Learning for Lung Cancer Screening Models," *Applied Sciences*, vol. 12, no. 4, p. 1926, Feb. 2022, doi: 10.3390/app12041926.
[33] J. Jiménez-Luna, F. Grisoni, and G. Schneider, "Drug discovery with explainable artificial intelligence," *Nat Mach Intell*, vol. 2, no. 10, pp. 573–584, Oct. 2020, doi: 10.1038/s42256-020-00236-4.
[34] F. Gabbay, S. Bar-Lev, O. Montano, and N. Hadad, "A LIME-Based Explainable Machine Learning Model for Predicting the Severity Level of COVID-19 Diagnosed Patients," *Applied Sciences*, vol. 11, no. 21, p. 10417, Nov. 2021, doi: 10.3390/app112110417.
[35] U. Bhatt *et al.*, "Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, Virtual Event USA: ACM, Jul. 2021, pp. 401–413. doi: 10.1145/3461702.3462571.
[36] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional variable importance for random forests," *BMC Bioinformatics*, vol. 9, no. 1, p. 307, Dec. 2008, doi: 10.1186/1471-2105-9-307.
[37] R. Kitani and S. Iwata, "Verification of Interpretability of Phase-Resolved Partial Discharge Using a CNN With SHAP," *IEEE Access*, vol. 11, pp. 4752–4762, 2023, doi: 10.1109/ACCESS.2023.3236315.
[38] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent Individualized Feature Attribution for Tree Ensembles," 2018, doi: 10.48550/ARXIV.1802.03888.
[39] Y. Arslan *et al.*, "Towards Refined Classifications Driven by SHAP Explanations," in *Machine Learning and Knowledge Extraction*, vol. 13480, A. Holzinger, P. Kieseberg, A. M. Tjoa, and E. Weippl, Eds., in Lecture Notes in Computer Science, vol. 13480. , Cham: Springer International Publishing, 2022, pp. 68–81. doi: 10.1007/978-3-031-14463-9_5.
[40] E. G. Lakatta and D. Levy, "Arterial and Cardiac Aging: Major Shareholders in Cardiovascular Disease Enterprises: Part II: The Aging Heart in Health: Links to Heart Disease," *Circulation*, vol. 107, no. 2, pp. 346–354, Jan. 2003, doi: 10.1161/01.CIR.0000048893.62841.F7.
[41] N. A. M. Zaini and M. K. Awang, "Hybrid Feature Selection Algorithm and Ensemble Stacking for Heart Disease Prediction," *IJACSA*, vol. 14, no. 2, 2023, doi: 10.14569/IJACSA.2023.0140220.
[42] J. B. Kostis, A. E. Moreyra, M. T. Amendo, J. Di Pietro, N. Cosgrove, and P. T. Kuo, "The effect of age on heart rate in subjects free of heart disease. Studies by ambulatory electrocardiography and maximal exercise stress test.," *Circulation*, vol. 65, no. 1, pp. 141–145, Jan. 1982, doi: 10.1161/01.CIR.65.1.141.
[43] D. Jacobs *et al.*, "Report of the Conference on Low Blood Cholesterol: Mortality Associations.," *Circulation*, vol. 86, no. 3, pp. 1046–1060, Sep. 1992, doi: 10.1161/01.CIR.86.3.1046.
[44] M. Hedayatnia *et al.*, "Dyslipidemia and cardiovascular disease risk among the MASHAD study population," *Lipids Health Dis*, vol. 19, no. 1, p. 42, Dec. 2020, doi: 10.1186/s12944-020-01204-y.