Vol. 8, No. 4, August 2026

ISI Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

# Optimizing KNN Classification for Heart Disease Prediction Using Sequential Forward Selection

**Herman[1], Rusydi Umar[2], Deni Kuswandani[3]**

[1,2,3]Informatics Department, Universitas Ahmad Dahlan, Yogyakarta

**Received:**
February 6, 2026
**Revised:**
July 15, 2026
**Accepted:**
July 29 2026
**Published:**
August 22, 2026

Corresponding Author:
**Author Name*:**
Herman
**Email*:**
hermankaha@mti.uad.ac.id

DOI:
10.63158/journalisi.v8i4.1744



**Abstract.** Irrelevant attributes often degrade the effectiveness of distance-based algorithms like K-Nearest Neighbors (KNN) in heart disease prediction. This study enhances a KNN model using Sequential Forward Selection (SFS) on the Cleveland dataset to optimize computational efficiency and precision while maintaining stable recall. To rigorously prevent data leakage, data partitioning was executed prior to mode imputation and normalization, followed by feature selection within a 5-fold stratified cross-validation framework. To further guarantee model robustness and rule out arbitrary selection, a 5-repeated 10-fold cross-validation and a 50-iteration feature stability analysis were executed. Compared to a baseline model (k=7, Euclidean; 80.33% accuracy) utilizing all 13 original attributes, the optimal 7-feature subset (cp, trestbps, chol, thalach, oldpeak, ca, thal) reduced the dimensional space by 46% and achieved 81.97% Accuracy, 77.42% Precision, 85.71% Recall, 78.79% Specificity, 0.9183 ROC-AUC, and an 81.36% F1-Score on an independent hold-out test set comprising 61 samples. Although McNemar's test (p = 1.0000) indicated the absolute accuracy improvement was not statistically significant, the 7-feature model successfully eliminated unstable features and reduced statistical noise. Ultimately, applying SFS provides a highly efficient, computationally lightweight framework for heart disease prediction without compromising predictive reliability.

**Keywords**: Heart Disease Prediction, Sequential Forward Selection (SFS), K-Nearest Neighbors (KNN), Feature Selection, Wrapper Feature Selection.

## 1. INTRODUCTION

To address the rising mortality from cardiovascular disease (CVD), reliable early detection systems are critical [1]. Machine learning algorithms increasingly outperform conventional clinical risk scores due to their flexibility with non-linear data [2], [3], [4]. Among these, K-Nearest Neighbors (KNN) is highly favored in medical applications for its interpretable, non-parametric approach that effectively identifies local data patterns [5], [6]. However, applying KNN to clinical datasets frequently triggers the curse of dimensionality. Irrelevant medical features distort distance metric computations (e.g., Euclidean or Manhattan), thereby degrading diagnostic precision and increasing computational burden [7], [8]. Using inappropriate distance metrics further misrepresents sample proximity in complex feature spaces.

While Sequential Forward Selection (SFS) is a proven wrapper-based solution, existing studies explicitly applying KNN with forward selection on the Cleveland dataset such as the framework proposed by Zeniarja et al. [9], predominantly optimize overall accuracy while overlooking the stability of sensitivity (recall). Although other recent research integrates recall evaluation [10] , [11] the focus remains heavily on ensemble algorithms rather than KNN. Furthermore, broader comparative studies [12] rarely address the crucial interaction between feature dimensionality reduction and KNN's internal distance metric parameters within this specific dataset. In medical diagnostics, minimizing false negatives remains a critical consideration that must be carefully monitored, rather than strictly maximized, during model optimization [13], [14].

This study addresses these gaps by systematically optimizing an SFS-based KNN model to maximize overall accuracy while ensuring recall stability. It evaluates the dynamic interaction between feature reduction (5, 6, and 7 features), distance metrics (Euclidean and Manhattan), and small k values (3, 5, and 7). These specific k values were chosen to capture local decision boundaries without introducing excessive computational complexity or overfitting to noise, thereby optimizing the balance between efficiency and predictive reliability. Although the k = 3 Manhattan configuration yielded the highest raw accuracy on the hold-out test set, this study deliberately adopted the k = 7 Euclidean configurations as the baseline, since smaller k values combined with the Manhattan

metric are inherently more sensitive to local noise, whereas a larger k with the Euclidean metric was expected to yield a smoother and more generalizable decision boundary.

Formally, this study hypothesizes that optimizing KNN through SFS will significantly mitigate dimensional noise, improving classification precision and specificity without sacrificing disease detection sensitivity (recall). Accordingly, this research addresses the core question: How does the dynamic interaction between SFS-driven dimensionality reduction and internal KNN parameter tuning impact the balance between computational efficiency and predictive reliability?

Positioning this work strictly as a machine-learning experiment on a public dataset, it must be explicitly clarified that this study focuses on computational optimization and does not establish immediate clinical diagnostic readiness. Within this experimental scope, the contributions include: (1) determining an optimal combination of heart disease features on the Cleveland dataset; (2) presenting empirical evidence of improved accuracy and precision while maintaining recall stability; and (3) providing technical insights into optimal k values and distance metrics. Consequently, the model demonstrates how targeted feature selection enhances algorithmic efficiency and accuracy, serving as a robust computational foundation for future clinical validation.

## 2. METHODS

The methodology employed in this study involves the development and evaluation of a heart disease classification model utilizing the K-Nearest Neighbors (KNN) algorithm integrated with the Sequential Forward Selection (SFS) feature selection method. The research workflow encompasses data collection, data preprocessing, determination of the optimal k value, feature selection, and model performance evaluation, as illustrated in Figure 1.

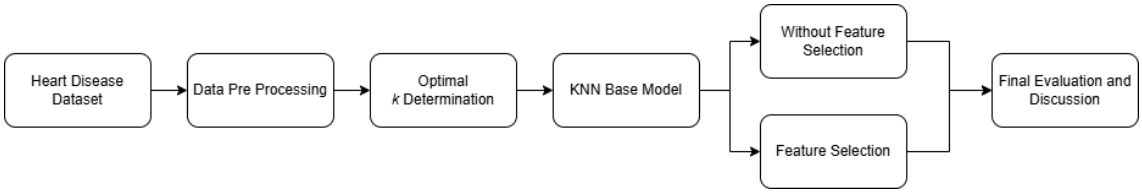


**Figure 1**. Research Methodology

### 2.1. Research Dataset

The heart disease dataset utilized in this study was sourced from the UCI Machine Learning Repository, specifically employing the Cleveland Heart Disease Dataset [15]. This dataset consists of 303 patient samples with 13 predictor medical variables encompassing demographic information and clinical observations. Although the original dataset features a target variable scaled from 0 (no indication) to 4 (severe indication), this study applies binarization to simplify the classification space. The original target values ranging from 1 to 4 were converted into label 1, representing patients diagnosed with heart disease (positive class), whereas the value of 0 was retained as label 0, representing a healthy condition (negative class). Based on this transformation, the class distribution reveals 164 healthy samples and 139 diseased samples. This proportion indicates a relatively balanced data profile, rendering it suitable for classification modeling. The representation of the structure and format of the raw data used can be observed in the snippet of the first five rows presented in Table 1.

**Table 1**. Excerpt of the First Five Rows of the Dataset

| age | sex | cp | trestbps | ... | slope | ca | thal | target |
|---|---|---|---|---|---|---|---|---|
| 63 | 1 | 1 | 145 | ... | 3 | 0 | 6 | 0 |
| 67 | 1 | 4 | 160 | ... | 2 | 3 | 3 | 2 |
| 67 | 1 | 4 | 120 | ... | 2 | 2 | 7 | 1 |
| 37 | 1 | 3 | 130 | ... | 3 | 0 | 3 | 0 |
| 41 | 0 | 2 | 130 | ... | 1 | 0 | 3 | 0 |

### 2.2. Data Preprocessing and Partitioning

To enhance data quality and consistency, a preprocessing stage was conducted as an initial step prior to the modeling process [16]. To rigorously prevent data leakage during this stage, the complete dataset (N=303) was first partitioned into 80% training (n=242) and 20% independent hold-out testing (n=61) sets using stratified random sampling (random_state=42). This ensured the original class distribution 164 healthy (54.1%) and 139 diseased (45.9%) samples was perfectly preserved across both subsets, preventing model bias [17]. Crucially, implementing a strict pipelined approach to prevent data leakage, all preprocessing parameters were fitted exclusively on the training data and subsequently applied to transform the independent test set. Missing values in the clinical attributes ca and thal were handled via mode imputation [18]. Categorical features (cp, restecg, slope,

thal) were retained in their ordinal numeric format. This deliberate choice avoids the dimensionality expansion of one-hot encoding while maintaining graded severity levels compatible with KNN distance computations.

Subsequently, Min-Max Normalization was applied to equalize feature scales, ensuring that variables with large ranges (e.g., cholesterol or maximum heart rate) do not disproportionately dominate the distance algorithms [19]. The testing set was strictly reserved for the final evaluation to provide an objective estimate of unseen data performance [20]. All computational experiments, including data handling and statistical validations, were executed using Python 3.10, integrating pandas (v2.3.3) and NumPy (v2.3.5) for manipulation, Scikit-learn (v1.2.2) for modeling, Mlxtend (v0.22.0) for feature selection, and Statsmodels (v0.14.0) for statistical testing.

### 2.3. KNN Classification Model

All modeling and evaluation procedures were implemented using the Python programming language, utilizing the Scikit-Learn library for the KNN algorithm and the Mlxtend library for the SFS implementation. The K-Nearest Neighbors (KNN) algorithm a non-parametric, distance-based classification method originally formalized by Cover and Hart [20], was utilized to develop a classification model that determines the class of an instance based on its proximity to a predefined number of nearest neighbors within the feature space. In this study, a uniform weighting scheme was applied, where all points in each neighborhood were weighted equally.

To ensure model stability, testing was conducted across three k values representing the number of nearest neighbors, specifically k = 3, 5, and 7. These specific small k values were selected to provide a focused tuning space that effectively captures local decision boundaries on a relatively small dataset (303 samples) without introducing excessive computational complexity or over-smoothing the classifications. This study evaluates the model's performance by employing two distinct distance metrics to identify the most effective method for measuring inter-patient similarity. The first metric is the Manhattan distance, which calculates the sum of the absolute differences between features. The second metric is the Euclidean distance, which measures the distance based on the square root of the sum of the squared differences between features. The application of these two metrics facilitates a comparative analysis regarding the impact of different

distance measurement methods on enhancing the performance of the KNN algorithm during the classification process.

Model performance evaluation was conducted using two approaches. First, a baseline model was constructed utilizing all 13 features to determine the combination of the k value and distance metric that yielded optimal performance. This baseline configuration was not selected solely on the basis of the single highest raw accuracy on the hold-out test set; rather, k = 7 with the Euclidean metric was adopted because smaller k values combined with the Manhattan metric are theoretically more susceptible to the influence of local noise, whereas the Euclidean metric is generally considered more appropriate for normalized, continuous feature spaces such as the one used in this study. Second, the model was optimized using the Sequential Forward Selection (SFS) method, based on the established optimal parameters.

### 2.4. Sequential Forward Selection

The optimization of the KNN model was executed using the Sequential Forward Selection (SFS) method, a well-established wrapper-based heuristic search algorithm [21]. In contrast to previous approaches that relied heavily on a recall-driven narrative, this study establishes a balanced SFS-KNN framework with a primary emphasis on maximizing overall accuracy while ensuring the stability of recall performance. Within this framework, SFS serves not only as a feature reduction technique but also as a heuristic strategy to bridge the translational gap by generating a simpler and more interpretable model. By prioritizing this balanced evaluation, the framework shifts the focus toward a level of classification precision that bolsters predictive reliability, as illustrated in Figure 2 through systematic feature selection.

The integration of Sequential Forward Selection (SFS) into the K-Nearest Neighbors (KNN) model aims to enhance the model's effectiveness by strengthening three primary aspects. First, it simplifies the feature space to overcome dimensional complexity. Because the KNN algorithm operates based on distance measurements (Euclidean/Manhattan) within a high-dimensional space, the effectiveness of these measurements in medical datasets comprising 13 features frequently diminishes due to data sparsity [22]. SFS addresses this by simplifying the feature space, thereby rendering distance calculations more relevant and precise. Second, it eliminates noise and

redundancy. Given that KNN is highly susceptible to irrelevant features, SFS acts as a dynamic filter that removes interfering variables. For instance, it eliminates statistical variance anomalies in specific metrics (e.g., cholesterol or fasting blood sugar) that, while medically relevant in real-world clinical scenarios, may lack discriminative statistical utility within the specific boundaries of this dataset Third, it optimizes the decision boundary. By selecting the most discriminative feature subsets, SFS assists KNN in establishing a more robust decision boundary that is resilient to outliers, ultimately improving the model's generalization capability [23].

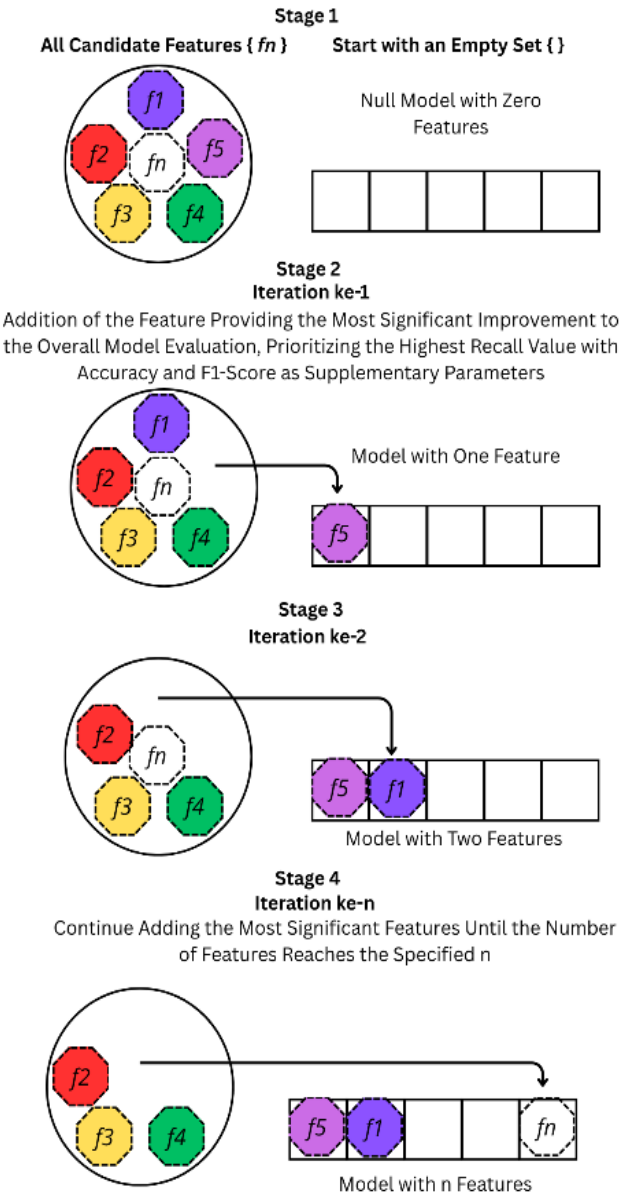


**Figure 2**. SFS-KNN Framework

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

In this study, the feature selection process was conducted using a wrapper-based SFS method, with the KNN algorithm serving as the internal evaluation engine. To strictly prevent data leakage, the SFS algorithm was configured to use a 5-fold stratified cross-validation loop as its scoring mechanism, operating solely on the training dataset. KNN was selected due to its flexibility in utilizing distance metrics, which effectively influences classification performance on medical data, as investigated by Prasath et al. [13]. Mathematically, the search process for new features to be added to achieve optimal performance is based on evaluation criteria according to Equation (1):

$$X_t = \arg max_{x \in Y \setminus St-1} \; J(S_{t-1} \cup \{x\}) \quad (1)$$

Here, $x_t$ represents the single selected feature, Y is the complete set of available candidate features, $S_{t-1}$ is the subset of features already selected in the previous iteration, and J represents the objective scoring function (in this case, the 5-fold cross-validation Accuracy). The *arg max* operator is utilized to determine and select a single feature x that possesses the optimal value from the pool of available candidate features, which technically yields the maximum value for the objective function J compared to other candidates. Based on this rationale, the process is executed iteratively according to the steps illustrated in Figure 2, as follows:

Stage 1: Initialization (Empty Subset): The algorithm initiates with an empty feature set, $S_0 = \emptyset$. At the top of Figure 2, this stage is depicted as the "null model," containing no predictor features. Stage 2: Addition of the First Feature: Each of the total 13 features is evaluated independently using KNN. The feature yielding the highest Accuracy value is selected as the first feature ($x_1$). The visualization at "Iteration 1" demonstrates that feature $f_5$ is selected to enter the "One-feature model" as it provides the most significant improvement to the overall model evaluation. In the event of a tie in Accuracy values, Recall and F1-Score are employed as supplementary tie-breaker parameters. Stage 3: Continuous Expansion: The remaining features are individually tested for integration with the previously selected features ($S_{t-1}$). The transition process occurs at "Iteration 2," where, in Figure 2, feature $f_1$ is drawn from the candidate pool to be combined with $f_5$, creating a "Two-feature model". Stage 4: Iterative Progression: This process continues until it reaches "Iteration n" to generate the final subset. The objective function in this study

prioritizes the Accuracy metric to improve overall prediction correctness, while rigorously monitoring Recall to minimize false negatives (diseased patients incorrectly classified as healthy).

To ensure objective optimization, the SFS algorithm was configured to halt automatically when marginal improvement in 5-fold cross-validation accuracy reached ≤0.0% or if recall declined. Consequently, adding features beyond the 7th attribute degraded performance, isolating the 5-, 6-, and 7-feature subsets as the optimal candidates. This restriction effectively mitigates the impact of irrelevant features that obscure KNN distance metric calculations. As illustrated in Figure 2, selectively reducing the feature space enhances predictive precision compared to utilizing all available variables. This aligns with Noor and Wardiah [10]. who demonstrated that a streamlined subset (9 of 13 features) increased KNN accuracy to 92%, further confirming that indiscriminate feature inclusion distorts predictive performance.

### 2.5. Model Evaluation

Model performance was comprehensively evaluated using a confusion matrix (to track True/False Positives and Negatives) and multiple metrics: Accuracy, Precision, Recall, Specificity, Balanced Accuracy, ROC-AUC, and F1-Score. This balanced approach prioritizes overall accuracy and precision while maintaining stable recall to minimize clinically critical false negatives, ensuring that performance improvements genuinely reflect better disease detection rather than class bias.

To guarantee objective measurement and prevent data leakage, the final evaluation was strictly conducted on an independent hold-out test set. Furthermore, two advanced validations were implemented to assure internal robustness: a 5-repeated 10-fold stratified cross-validation evaluated metric consistency across 50 iterations, and a Feature Stability Analysis. Crucially, during this repeated cross-validation, the SFS algorithm was strictly re-run inside each training fold across all 50 iterations, rather than just testing the pre-selected 7-feature model. This rigorous procedure prevented data leakage and accurately quantified the consistency of the SFS algorithm in iteratively selecting core clinical features. Finally, McNemar's exact test was applied to the hold-out predictions to determine the statistical significance of the model's improvements.

## 3. RESULTS AND DISCUSSION

### 3.1. Baseline KNN Model Performance

A baseline model was initially developed as a preliminary step in the testing process, utilizing all 13 features present in the dataset. The objective of this evaluation was to establish an initial performance standard prior to proceeding with the dimensionality reduction process using Sequential Forward Selection (SFS). The evaluation was conducted by setting the values for the nearest neighbors (k) within an odd range (3, 5, 7) to prevent ambiguity in class determination, and by comparing the Manhattan and Euclidean distance metrics to assess the performance of the KNN algorithm on the heart disease dataset. The performance results of the baseline model are presented in Table 2.

**Table 2**. KNN Evaluation Results for Various k Values

| k Values | k=3 | | k=5 | | k=7 | |
|---|---|---|---|---|---|---|
| **Metric** | **Manhattan** | **Euclidean** | **Manhattan** | **Euclidean** | **Manhattan** | **Euclidean** |
| Accuracy | 81.97% | 80.33% | 80.33% | 80.33% | 80.33% | 80.33% |
| Precision | 77.42% | 75.00% | 75.00% | 75.00% | 75.00% | 75.00% |
| Recall | 85.71% | 85.71% | 85.71% | 85.71% | 85.71% | 85.71% |
| F1-Score | 81.36% | 80.00% | 80.00% | 80.00% | 80.00% | 80.00% |

Although the k = 3 Manhattan configuration achieved the highest accuracy on the independent hold-out test set (81.97%), this configuration was not adopted as the baseline. As all six k–metric combinations were evaluated directly on the same hold-out test set, selecting the single highest-accuracy configuration as the baseline would risk tailoring the baseline choice to this particular test partition. Instead, the k = 7 Euclidean configuration which achieved a consistent 80.33% accuracy shared across most of the tested k–metric combinations was selected as a more conservative baseline, given its comparatively lower sensitivity to local noise relative to the Manhattan metric at small k.

**Table 3**. Confusion matrix Baseline

| **Actual** | **Model Prediction** | |
|---|---|---|
| | **Diseased** | **Healthy** |
| **Diseased** | 24 (TP) | 4 (FN) |
| **Healthy** | 8 (FP) | 25 (TN) |

Based on the findings obtained from the analysis at k = 7 utilizing the Euclidean distance, as illustrated in Table 3, the KNN classification model demonstrated satisfactory results with an accuracy rate reaching 80.33%. Out of the total 61 test data instances, the model successfully and accurately identified 24 patients in the 'Diseased' category (True Positives). Given that there were actually 28 diseased patients in the test data, this yielded a recall value of 85.71%.

### 3.2. KNN Performance with Feature Subsets (SFS)

The subsequent step focused on evaluating the performance of specialized feature subsets obtained through the Sequential Forward Selection (SFS) method and comparing them against the baseline model utilizing all 13 original features. To guarantee objectivity and prevent data leakage, the search process for feature combinations was strictly conducted only within the training set using a 5-fold cross-validation approach, with accuracy as the target evaluation metric. The final performance evaluation of the selected subsets was then conducted separately on an independent hold-out test set. The comparative results of the model's performance for the baseline and the specific subsets of 5 features (cp, trestbps, thalach, ca, thal), 6 features (cp, trestbps, chol, thalach, ca, thal), and 7 features (cp, trestbps, chol, thalach, oldpeak, ca, thal) are presented in Table 4.

**Table 4**. Comparative Performance of SFS Feature Subsets on the Test Data

| Evaluation Metric | Baseline (13 Features) | SFS (5 Features) | SFS (6 Features) | SFS (7 Features) |
|---|---|---|---|---|
| True Positive | 24 | 23 | 23 | 24 |
| True Negative | 25 | 28 | 26 | 26 |
| False Positive | 8 | 5 | 7 | 7 |
| False Negative | 4 | 5 | 5 | 4 |
| Accuracy | 80.33% | 83.61% | 80.33% | 81.97% |
| Precision | 75.00% | 82.14% | 76.67% | 77.42% |
| Recall | 85.71% | 82.14% | 82.14% | 85.71% |
| Specificity | 75.76% | 84.85% | 78.79% | 78.79% |
| BACC | 80.74% | 83.50% | 80.47% | 82.25% |
| F1-Score | 80.00% | 82.14% | 79.31% | 81.36% |
| ROC-AUC | 0.8858 | 0.9194 | 0.9085 | 0.9183 |

The utilization of feature subsets obtained through SFS demonstrated varying impacts on performance. Although the 5-feature subset achieved the highest absolute accuracy (83.61%), it reduced the recall metric to 82.14%, indicating an increase in false-negative predictions. Since clinical safety requires preserving disease detection capability, the 7-feature subset (cp, trestbps, chol, thalach, oldpeak, ca, and thal) was selected as the optimal final model. This 7-feature configuration was prioritized because it successfully maintained the baseline recall (85.71%) to minimize false negatives, while simultaneously improving overall accuracy, precision, and specificity. Based on the Confusion Matrix evaluation of this optimal 7-feature model, as presented in Table 5, it demonstrated robust performance in distinguishing between the two classes without sacrificing sensitivity.

**Table 5**. Confusion matrix of the Optimal SFS Model (7 Features)

| **Actual** | **Model Prediction** | |
|---|---|---|
| | **Diseased** | **Healthy** |
| **Diseased** | 24 (TP) | 4 (FN) |
| **Healthy** | 7 (FP) | 26 (TN) |

Derived from these confusion matrix values, the optimal model successfully maintained the baseline recall (85.71%) while concurrently improving Accuracy (to 81.97%), Precision (to 77.42%), and ROC-AUC (to 0.9183). Furthermore, the model recorded a Specificity of 78.79% and a Balanced Accuracy of 82.25%, confirming an unbiased capability in class distinction without skewing toward the majority class.

To ensure that these performance metrics were not merely the result of a favorable data split, the models were further subjected to a rigorous 5-repeated 10-fold stratified cross-validation on the training set. For the baseline model, this comprehensive stability evaluation across 50 iterations yielded a mean accuracy of 81.40% ± 5.69%. The presence of a 5.69% standard deviation validates the inherent variance expected from the relatively small Cleveland dataset, yet confirms that the model maintains a consistent predictive baseline. To provide a complete comparative view, the distribution of performance metrics for both the baseline model and the proposed SFS subsets across all 50 evaluation runs is comprehensively detailed in Table 6.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

**Table 6**. Performance Distribution Across Repeated Stratified Cross-Validation (5-Repeated 10-Fold CV)

| Model Architecture | Mean Accuracy (%) | Mean Precision (%) | Mean Recall (%) | Mean F1-Score (%) |
|---|---|---|---|---|
| Baseline (13 Features) | 81.40 ± 5.69 | 83.17 ± 9.56 | 76.53 ± 11.09 | 78.85 ± 6.79 |
| SFS (5 Features) | 83.05 ± 7.19 | 84.87 ± 10.60 | 78.55 ± 11.74 | 80.79 ± 8.42 |
| SFS (6 Features) | 82.90 ± 7.25 | 83.62 ± 10.42 | 79.83 ± 11.48 | 80.95 ± 8.11 |
| SFS (7 Features) | 82.72 ± 6.96 | 83.73 ± 10.24 | 79.27 ± 11.78 | 80.64 ± 7.95 |

To rigorously evaluate the final performance change on the 61 hold-out test samples, McNemar's exact test was conducted. The baseline model accurately predicted 49 out of 61 cases (80.33%), whereas the 7-feature SFS model accurately predicted 50 cases (81.97%). McNemar's test comparing these paired predictions yielded a p-value of 1.0000 ($p > 0.05$). This indicates that the addition of one correct prediction is not statistically significant. However, this finding is crucial as it demonstrates that the 7-feature subset successfully maintained the model's maximum predictive reliability and improved the ROC-AUC while drastically reducing the computational dimensionality by 46%.

### 3.3. Iteration Dynamics and Feature Inclusion Chronology

The iterative behavior of Sequential Forward Selection (SFS) evaluated within the 5-fold cross-validation training framework is illustrated in Figure 3. The 7-feature configuration yielded the optimal internal cross-validation performance, achieving an independent hold-out test accuracy of 81.97%.

As shown in Figure 3, the inclusion of features progressed through sequential steps. Step 1 incorporated Thalassemia (thal, 77.25% internal accuracy), followed by the Number of major vessels (ca, 77.30%) in Step 2. The most substantial internal performance increase occurred at Step 3 with the addition of Chest pain type (cp), raising the CV accuracy to 84.29%. Subsequent additions from Step 4 through Step 7 (thalach, trestbps, chol, and oldpeak) resulted in a stable internal accuracy plateau around 83.87%, culminating in the final 7-feature subset that balanced overall predictive performance on the test set.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

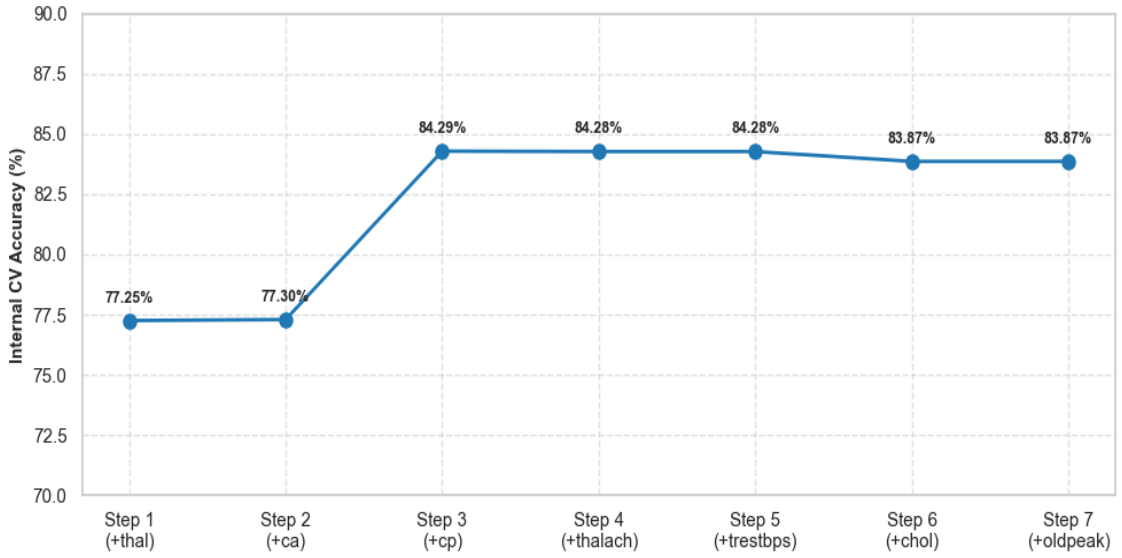


**Figure 3**. Feature Inclusion Chronology and Performance Synergy during 5-Fold Cross-Validation (Training Phase)

Furthermore, a Feature Stability Analysis across 50 iterations of 5-repeated 10-fold cross-validation evaluated the consistency of the selected attributes. As summarized in Table 7, primary clinical features such as thal and ca exhibited exceptionally high stability (98.00% and 94.00% selection frequency, respectively), whereas secondary features demonstrated lower selection consistency.

**Table 7**. Feature Stability Analysis Results

| Medical Feature Name | Original Feature Code | Selection Frequency | Stability Percentage |
|---|---|---|---|
| Thalassemia (thal) | Feature 12 | 49 | 0.98 |
| Number of major vessels (ca) | Feature 11 | 47 | 0.94 |
| Chest pain type (cp) | Feature 2 | 35 | 0.70 |
| Age (age) | Feature 0 | 34 | 0.68 |
| Serum cholestoral (chol) | Feature 4 | 33 | 0.66 |
| ST depression / oldpeak | Feature 9 | 28 | 0.56 |
| Sex (sex) | Feature 1 | 27 | 0.54 |
| Resting blood pressure (trestbps) | Feature 3 | 25 | 0.50 |
| Fasting blood sugar (fbs) | Feature 5 | 23 | 0.46 |
| Maximum heart rate (thalach) | Feature 7 | 19 | 0.38 |
| Exercise induced angina (exang) | Feature 8 | 18 | 0.36 |

| Medical Feature Name | Original Feature Code | Selection Frequency | Stability Percentage |
|---|---|---|---|
| Slope of peak ST segment (slope) | Feature 10 | 7 | 0.14 |
| Resting ECG results (restecg) | Feature 6 | 5 | 0.10 |

Several features in the final 7-feature subset—such as chol (66%), oldpeak (56%), resting blood pressure or trestbps (50%), and particularly maximum heart rate or thalach (38%)—demonstrated moderate to low selection consistency. This occurs because SFS acts as a greedy wrapper, selecting features based on local synergistic interactions within specific training splits rather than their independent predictive strength. Consequently, this variance limits confidence in the final subset. It should not be presented as a fully stable set of universal clinical biomarkers. Instead, it must be cautiously interpreted as a mathematically optimal configuration tailored specifically to reduce dimensionality and refine KNN decision boundaries for this limited Cleveland dataset partition.

### 3.4. Comparative Analysis

To evaluate the SFS-KNN framework, its performance was compared against previous studies using the Cleveland dataset. Zeniarja et al. achieved 87.00% accuracy using KNN with Forward Selection (8 features), while Takcı reported 90.11% using tuned KNN on all 13 features. Furthermore, standard Logistic Regression benchmarks generally range from 80.00% to 85.00%, whereas Nasution et al. reached 88.00%–91.00% using complex algorithms like Support Vector Machine (SVM) and Random Forest on 13 features. In comparison, our proposed model achieved a strictly leakage-free accuracy of 81.97% on the independent test set using a reduced 7-feature subset.

Direct performance comparisons are inherently limited because previous studies often utilized varying train-test splits, different validation protocols, or lacked strict pre-split data isolation. Although the absolute accuracy of the proposed model appears lower than some reported figures, the SFS-KNN approach proves highly effective. Reducing the feature space by 46% creates a simpler, interpretable, and computationally efficient model while maintaining predictive accuracy and avoiding overfitting. Importantly, all reported metrics derive solely from the internal hold-out test set; no external validation on independent cohorts was performed.

### 3.5. Discussion

#### 3.5.1. Performance Parameter Evaluation and Baseline Model

In medical dataset classification, minimizing false negatives is highly desirable. However, optimizing a model solely based on recall often compromises overall accuracy, leading to a high rate of false positives. Therefore, this study establishes overall accuracy as the principal objective function during the subsequent Sequential Forward Selection (SFS) process, while rigorously monitoring the recall metric to ensure balanced predictive performance. By focusing on this balanced optimization, it is anticipated that the selected features will emerge as the most discriminative variables in detecting indications of heart disease. Consequently, the probability of false negatives can be minimized through the elimination of interfering features (noise) that could distort the model's decision boundary. This approach aligns with the research conducted by Nasution et al., which highlights that appropriate parameter optimization effectively influences the competitiveness of KNN when compared against other classification models on heart disease datasets [24].

Table 2 summarizes the baseline performance of the KNN classifier using different combinations of neighborhood size (k = 3, 5, and 7) and distance metrics (Manhattan and Euclidean). Based on the evaluation results, the configuration using k = 7 with the Euclidean distance metric was selected as the baseline model for the subsequent Sequential Forward Selection (SFS) optimization. Although smaller k values may produce competitive performance, they are generally more susceptible to overfitting because the classification decision relies on a limited number of neighboring samples, making the model more sensitive to noise and outliers. This observation is consistent with the theoretical analysis by Prasath et al., which explains that excessively small k values tend to generate unstable decision boundaries that are highly susceptible to noise distortion. [13].

Conversely, the configuration of k = 7 combined with the Euclidean metric was established as the baseline model. Although the k = 3 (Manhattan) configuration achieved a slightly higher absolute accuracy on the single hold-out test set (81.97%), relying on such a small neighborhood size in small medical datasets frequently leads to overfitting due to high sensitivity to local noise. The k = 7 Euclidean configuration was deliberately

selected because it creates a smoother, more generalized decision boundary that is theoretically less susceptible to local noise, reflecting a more favorable bias-variance tradeoff [13]. This robust generalization is also supported by empirical research in the context of heart disease classification by Mukherjee and Ali [25]. Furthermore, this configuration successfully maintains a stable and high recall value (85.71%), which remains essential in medical datasets to mitigate the risk of false negatives. In evaluating the distance metrics, both the Manhattan and Euclidean metrics demonstrated identical predictive performance at larger neighborhood sizes (k = 5 and k = 7). However, the Euclidean metric was prioritized because it computes the squared differences between features, rendering it more effective in capturing the numerical data distribution within the attributes of this dataset following transformation via Min-Max normalization. As elucidated by Prasath et al., within a normalized feature space, the Euclidean metric tends to provide mathematically more reliable performance due to its ability to account for diagonal disparities in multidimensional space more accurately than the Manhattan metric [13].

The results obtained from the baseline model exhibit relatively strong performance; however, several critical aspects require further in-depth evaluation. First, the simultaneous use of all 13 medical features increases both the computational burden and the complexity of the model. Second, the curse of dimensionality phenomenon can render the model sensitive to irrelevant features, thereby making the Euclidean distance calculations between patient samples less precise. Therefore, the configuration of k = 7 with the Euclidean metric was established as the optimal baseline model to be utilized in the subsequent Sequential Forward Selection (SFS) optimization stage. The establishment of this robust baseline constitutes a crucial methodological step to ensure that the feature selection phase possesses a valid and reliable reference point [9], [26].

### 3.5.2. The Systematic Impact of Dimensionality Reduction (SFS)

The success of SFS in reducing the number of features demonstrates that this method is highly effective in eliminating uninformative and noisy features, which could previously disrupt the formation of the feature space in distance-based classification algorithms such as KNN. The research findings indicate that each feature added during the selection process, up to a total of 7 features, contributes to establishing a more balanced predictive model.

As indicated by McNemar's test (p = 1.0000), the absolute accuracy enhancement from 80.33% to 81.97% does not reach statistical significance, primarily due to the limited sample size of the hold-out test set (n=61). However, evaluating a medical diagnostic model based solely on absolute accuracy obscures the true systematic impact of SFS. The reduction of features from 13 to 7 successfully mitigated the curse of dimensionality, which is evident from the systematic improvements in ROC-AUC (from 0.8858 to 0.9183) and Specificity (from 75.76% to 78.79%). By discarding noisy features, the model established a more precise decision boundary that slightly reduced false-positive predictions (from 8 down to 7) without sacrificing a single true positive. This configuration was selected because it successfully enhances overall predictive reliability and AUC while maximizing computational efficiency (a 46.15% reduction in feature space). The successful implementation of SFS in this study emphasizes that the quality of selected features exerts a far more substantial influence on model specificity and reliability than the mere quantity of input features utilized.

### 3.5.3. Feature Synergy and Non-Linear Interactions

Referring to the iterative behavior presented in Figure 3, there is a strong indication that these features interact synergistically to support the model's performance. In Step 1, Thalassemia (thal) emerges as the strongest initial predictor, which is subsequently reinforced by the condition of major blood vessels (ca) in Step 2. The most critical synergy occurs in Step 3; the addition of chest pain type (cp) dramatically increases the internal cross-validation accuracy. Interestingly, the subsequent additions of maximum heart rate (thalach), resting blood pressure (trestbps), serum cholesterol (chol), and ST depression (oldpeak) in Steps 4 through 7 function collectively to refine and stabilize the final decision boundary against unseen test data. This demonstrates that SFS successfully identifies non-linear interactions; for instance, features like oldpeak and chol become highly discriminative when combined with primary data regarding chest pain type (cp) and major blood vessels (ca). This synergy aligns with the findings of Sang et al., which indicate that these features serve as primary predictive elements in the diagnosis of heart disease [5].

The reduction in the number of features from 13 to 7 (a reduction of 46.15%) plays a pivotal role in addressing the curse of dimensionality problem inherent in KNN. In a high-dimensional space, Euclidean distance calculations tend to lose their significance because

the distances between data points become nearly uniform. By eliminating irrelevant features through SFS, the distance calculations between patient samples become more precise and focus strictly on features with high biological relevance [19]. As a result, the model is able to form a more defined decision boundary, as evidenced by the increase in absolute accuracy from the 80.33% baseline to 81.97% on the hold-out test set.

Furthermore, the deliberate inclusion of traditional features such as serum cholesterol (chol) in the optimal subset highlights its statistical utility in establishing the distance-based decision boundary for this specific dataset. Conversely, the exclusion of features that demonstrated extreme instability during cross-validation, such as resting ECG results (restecg) and fasting blood sugar (fbs), reinforces the analysis conducted by Firdaus et al. [26], which demonstrates that features with high variance or low correlation to the target frequently act as noise in distance-based algorithms [22]. In addition to technical efficiency, the model's success in achieving optimal accuracy with a perfectly consistent sensitivity (recall) at Step 7 (85.71%) demonstrates promising predictive performance within this experimental machine-learning setting [7]. This deliberate emphasis on maintaining stable recall ensures that the developed framework not only excels computationally by eliminating interfering features but also provides a reliable statistical foundation to help mitigate the risk of false negatives, which is a critical consideration for future clinical applications [27].

While the final 7-feature subset successfully mitigates the curse of dimensionality and improves the model's predictive performance, the results from the Feature Stability Analysis necessitate a highly cautious interpretation. As observed, the final optimal subset is anchored by highly stable primary predictors such as thal (98%) and ca (94%). However, it concurrently incorporates features with moderate to low selection stability, most prominently maximum heart rate or thalach (38% stability). The inclusion of such a low-stability feature indicates that SFS, acting as a greedy wrapper method, optimized the subset based on specific synergistic interactions that maximized accuracy within the evaluated training splits, rather than selecting universally robust individual predictors. Consequently, the proposed 7-feature configuration must be strictly interpreted as the optimal mathematical subset for maximizing KNN performance on this specific Cleveland dataset partition. It highlights data-specific non-linear interactions rather than serving

as a definitive, universal set of independent clinical biomarkers for broader medical diagnosis.

### 3.5.4. Comparative Advantages and Practical Implications

The comparative analysis results indicate that the proposed model achieves a highly robust and strictly leakage-free accuracy rate (81.97%) on the Cleveland dataset. While some previous studies report higher absolute accuracies, many of these evaluations lack strict pre-split imputation and rigorous cross-validation protocols, making them susceptible to data leakage [17]. This advantage in methodological reliability is particularly evident when comparing the framework to the baseline performance of methods proposed by Takcı, which utilizes all 13 original features [28]. Methodologically, this reinforces the premise that in the absence of a stringent feature selection process, distance-based algorithms such as KNN will experience performance degradation due to distortion caused by irrelevant features (noise). The implementation of SFS in this study shows potential in extracting the most contributory information, thereby establishing a more robust feature subset compared to the simpler feature selection method employed in the study by Zeniarja et al. [9].

A compelling aspect of these findings emerges when compared to the results reported by Nasution et al., who applied more complex algorithms such as Support Vector Machine (SVM) and Random Forest [24]. Although these models are frequently considered more effective in handling high-dimensional data, the results of this study demonstrate that a computationally lightweight algorithm like KNN can remain highly competitive, particularly when preceded by an appropriate feature reduction process. However, it is reiterated that definitive conclusions regarding algorithmic superiority require benchmarking under identical data splits and validation protocols. By reducing the feature set to 7 primary features, the curse of dimensionality problem that typically hinders KNN was successfully mitigated, rendering the Euclidean distance calculations more effective in differentiating between positive and negative classes.

Viewed beyond the aspect of accuracy, the resulting efficiency (a feature reduction of 46.15%) provides practical advantages in computational efficiency and model interpretability. By relying exclusively on the 7 selected medical features, this model highlights the potential to streamline data processing workflows without compromising

predictive reliability. Another principal advantage is the model's balanced evaluation approach. Through the integration of the balanced SFS-KNN framework, the proposed model achieved higher overall absolute Accuracy (81.97%), Precision (77.42%), and ROC-AUC (0.9183) while perfectly maintaining the sensitivity (recall) value (85.71%) comparable to that of the baseline model. This finding indicates that the observed performance improvement was successfully associated with a reduction in false-positive predictions without degrading the disease detection rate. This balanced performance effectively mitigates the risk of false negatives, a crucial requirement when evaluating machine learning applications on medical datasets [7].

Despite these promising results, several limitations should be acknowledged. First, the proposed framework was evaluated exclusively using the Cleveland Heart Disease Dataset, which consists of only 303 patient records from a single publicly available source, and no external validation was performed. Although this dataset is widely adopted as a benchmark for heart disease prediction, its relatively small sample size, its historical nature as an older dataset originally collected in 1988, its limited population diversity, and the absence of external validation on independent contemporary clinical cohorts may significantly restrict the generalizability of the findings to modern clinical settings. Therefore, the reported performance should be interpreted strictly as evidence of computational effectiveness within this experimental machine-learning setting rather than as confirmation of clinical applicability. Additionally, given the limited size of the independent hold-out test set (n = 61), confidence intervals for the classification metrics were not computed, as they would be statistically unreliable at this sample size; future work with larger test sets should report these intervals to better quantify estimation uncertainty.

Second, this study transformed the original target labels (0–4) into a binary classification problem by grouping labels 1–4 into a single positive class. While this preprocessing strategy is commonly adopted in previous studies to facilitate binary classification, it also removes information regarding disease severity. Consequently, the proposed model is intended to distinguish between the presence and absence of heart disease rather than to predict different stages of disease progression. While repeated cross-validation demonstrated consistent internal predictive performance, future research should prioritize validating the proposed framework using larger external multicenter datasets

and evaluating its performance under multiclass classification settings to further assess its true generalizability across diverse clinical environments.

## 4. CONCLUSION

Integrating K-Nearest Neighbors (KNN) with Sequential Forward Selection (SFS) demonstrates potential for optimizing heart disease prediction. Streamlining the feature space from 13 to an optimal 7-feature subset (cp, trestbps, chol, thalach, oldpeak, ca, thal) improved absolute classification accuracy from 80.33% to 81.97% on the independent test set. Although this increase was not statistically significant (McNemar's p = 1.0000), the 7-feature model provided critical systematic benefits. By eliminating unstable attributes, it refined Euclidean distance calculations, achieving 77.42% Precision, 78.79% Specificity, and 0.9183 ROC-AUC. Importantly, recall remained stable at 85.71%, confirming that specificity gains did not sacrifice disease detection rates. Relying on a small, single public dataset, these findings must be interpreted strictly within an experimental machine-learning context and should not be directly generalized to clinical practice. Because feature subset stability remains partially unresolved, future research must verify the consistency of these selected features on larger, external multicenter datasets. Subsequent developments should also incorporate nested cross-validation, model probability calibration, and comprehensive medical expert validation to ensure these optimized frameworks can safely translate into reliable clinical tools.

## REFERENCES

[1] F. J. Pinto, E. Harding, and O. Fenton, "A manifesto from Global Heart Hub for early detection and diagnosis of cardiovascular disease," *Eur. Heart J.*, vol. 46, no. 4, pp. 339–340, Jan. 2025, doi: 10.1093/eurheartj/ehae455.

[2] T. Liu, A. Krentz, L. Lu, and V. Curcin, "Machine learning based prediction models for cardiovascular disease risk using electronic health records data: Systematic review and meta-analysis," *Eur. Heart J. Digit. Health*, vol. 6, no. 1, pp. 7–22, Jan. 2025, doi: 10.1093/ehjdh/ztae080.

[3] M. B. Matheson, Y. Kato, S. Baba, C. Cox, J. A. C. Lima, and B. Ambale-Venkatesh, "Cardiovascular risk prediction using machine learning in a large Japanese cohort," *Circ. Rep.*, vol. 4, no. 12, pp. 595–603, Dec. 2022, doi: 10.1253/circrep.cr-22-0101.

[4] H. J. Kim *et al.*, "Machine learning-based analysis of lifestyle risk factors for atherosclerotic cardiovascular disease: Retrospective case-control study," *JMIR Med. Inform.*, vol. 13, Art. no. e74415, 2025, doi: 10.2196/74415.

[5] H. Sang *et al.*, "Prediction model for cardiovascular disease in patients with diabetes using machine learning derived and validated in two independent Korean cohorts," *Sci. Rep.*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-63798-y.

[6] B. Pfeifer and M. Kreuzthaler, "Calibrated kNN classification via second-layer neighborhood analysis," *Adv. Data Anal. Classif.*, vol. 20, no. 1, pp. 145–165, 2026, doi: 10.1007/s11634-025-00654-5.

[7] D. I. Kasartzian and T. Tsiampalis, "Transforming cardiovascular risk prediction: A review of machine learning and artificial intelligence innovations," *Life*, vol. 15, no. 1, Art. no. 94, Jan. 2025, doi: 10.3390/life15010094.

[8] F. Shishehbori and Z. Awan, "Enhancing cardiovascular disease risk prediction with machine learning models," *arXiv Preprint* arXiv:2401.17328, Feb. 2024, doi: 10.48550/arXiv.2401.17328.

[9] J. Zeniarja, A. Ukhifahdhina, and A. Salam, "Diagnosis of heart disease using K-nearest neighbor method based on forward selection," *J. Appl. Intell. Syst.*, vol. 4, no. 2, pp. 39–47, Dec. 2019, doi: 10.33633/jais.v4i2.2749.

[10] Dhiyaussalam, M. H. Noor, Herlinawati, and I. Wardiah, "Impact of feature selection on the performance of KNN and SVM in heart disease prediction," *Tech: J. Eng. Sci.*, vol. 1, no. 1, pp. 14–25, 2025, doi: 10.69836/tech.v1i1.353.

[11] S. T. Veena, R. Jeevetha, and N. Abirami, "Prediction of heart disease using hybrid feature selection," *J. Posit. Sch. Psychol.*, vol. 6, no. 4, pp. 10454–10469, 2022.

[12] M. A. Bouqentar *et al.*, "Early heart disease prediction using feature engineering and machine learning algorithms," *Heliyon*, vol. 10, no. 19, Oct. 2024, doi: 10.1016/j.heliyon.2024.e38731.

[13] V. B. S. Prasath *et al.*, "Distance and similarity measures effect on the performance of K-nearest neighbor classifier: A review," *arXiv Preprint* arXiv:1708.04321, 2017.

[14] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," *BMC Med.*, vol. 17, no. 1, Art. no. 195, Oct. 2019, doi: 10.1186/s12916-019-1426-2.

[15] R. Detrano *et al.*, "Heart Disease," *UCI Mach. Learn. Repository*, 1988, doi: 10.24432/C52P4X.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[16] E. N. Wanyonyi and N. W. Masinde, "The impact of data preprocessing on machine learning model performance: A comprehensive examination," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 11, no. 2, pp. 3814–3827, Apr. 2025, doi: 10.32628/CSEIT25112854.

[17] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, Sep. 2023, doi: 10.1016/j.patter.2023.100804.

[18] A. Desiani *et al.*, "Handling missing data using combination of deletion technique, mean, mode and artificial neural network imputation for heart disease dataset," *Sci. Technol. Indones.*, vol. 6, no. 4, pp. 333–342, 2021, doi: 10.26554/sti.2021.6.4.303-312.

[19] A. Alsarhan, F. Hussein, S. Moh, and F. S. El-Salhi, "The effect of preprocessing techniques, applied to numeric features, on classification algorithms' performance," *Data*, vol. 6, no. 7, Art. no. 74, 2021, doi: 10.3390/data6020011.

[20] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967, doi: 10.1109/TIT.1967.1053964.

[21] P. Pudil, J. Novovičová, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognit. Lett.*, vol. 15, no. 11, pp. 1119–1125, Nov. 1994, doi: 10.1016/0167-8655(94)90127-9.

[22] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O'Sullivan, "A review of feature selection methods for machine learning-based disease risk prediction," *Front. Bioinform.*, vol. 2, Jun. 2022, doi: 10.3389/fbinf.2022.927312.

[23] U. M. Khaire and R. Dhanalakshmi, "Stability of feature selection algorithm: A review," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, no. 4, pp. 1060–1073, Apr. 2022, doi: 10.1016/j.jksuci.2019.06.012.

[24] N. Nasution, M. A. Hasan, and F. B. Nasution, "Predicting heart disease using machine learning: An evaluation of logistic regression, random forest, SVM, and KNN models on the UCI heart disease dataset," *IT J. Res. Dev.*, vol. 9, no. 2, pp. 140–150, Apr. 2025, doi: 10.25299/itjrd.2025.17941.

[25] R. Mukherjee, S. Sadhu, and A. Kundu, "Heart disease detection using feature selection based KNN classifier," in *Proc. Data Analytics and Management*, Singapore: Springer, 2022, doi: 10.1007/978-981-16-6289-8_48.

[26] F. F. Firdaus, H. A. Nugroho, and I. Soesanti, "A review of feature selection and classification approaches for heart disease prediction," *Int. J. Inf. Technol. Electr. Eng.*, vol. 4, no. 3, Sep. 2020, doi: 10.22146/ijitee.59193.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[27] A. A. Lamir, S. Razzagzadeh, and Z. Rezaei, "A comprehensive machine learning framework for heart disease prediction: Performance evaluation and future perspectives," *arXiv Preprint* arXiv:2505.09969, May 2025, doi: 10.48550/arXiv.2505.09969.

[28] H. Takcı, "Performance-enhanced KNN algorithm-based heart disease prediction with the help of optimum parameters," *J. Fac. Eng. Archit. Gazi Univ.*, vol. 38, no. 1, pp. 451–460, 2023, doi: 10.17341/gazimmfd.977127.