International Journal of Informatics and Computation (IJICOM)
Vol. 7, No. 2, 2025
ISSN: 2685-8711, E-ISSN: 2714-5263 DOI: 10.35842/ijicom

Arrhythmia Disease Detection using SVM with Recursive Feature Elimination

Setia Anfyasa Hadi[1], Tikaridha Hardiani[2]

Abstract
Arrhythmia is a critical cardiovascular disorder affecting approximately 1.5% to 5% of the global population. The issue of early detection remains challenging due to asymptomatic presentation and complex electrocardiogram (ECG) signal interpretation. Traditional diagnostic methods and existing machine learning approaches often struggle with high-dimensional medical data containing irrelevant features, leading to suboptimal classification performance. This study proposes an integrated approach combining Support Vector Machine (SVM) with Recursive Feature Elimination (RFE) for automated arrhythmia detection from the UCI Machine Learning Repository dataset containing 452 patient records with 278 features. The methodology incorporates comprehensive preprocessing, including normalization, Synthetic Minority Oversampling Technique (SMOTE) for class balancing, and RFE-based feature selection. Both Linear and Radial Basis Function (RBF) kernels were evaluated across four train-test split scenarios (90:10, 80:20, 70:30, 60:40). The proposed method achieved superior performance with 91.30% accuracy, 88.00% precision, 95.65% recall, and 91.67% F1-score using the RBF kernel in the 90:10 scenario. RFE successfully reduced dimensionality by 96.4%, selecting 10 optimal features from 278 original parameters while maintaining high classification accuracy. These findings demonstrate that the integration of SVM with RFE significantly enhances arrhythmia detection capability.

Keywords:
SVM, RFE, Arrhythmia, Detection

1. Introduction

The heart functions as a vital organ responsible for circulating blood throughout the human body, ensuring the delivery of oxygen and nutrients to all tissues. Cardiovascular diseases represent one of the leading causes of mortality globally. According to the World Health Organization (WHO), heart disease claims more than 17 million lives annually worldwide (WHO, 2024). In Indonesia specifically, the Ministry of Health reported that 651,481 deaths occurred due to heart disease in 2023 alone (Ministry of Health, 2023). This alarming mortality rate has positioned heart disease as a major public health concern for the Indonesian government, necessitating focused efforts on the prevention and treatment of non-communicable diseases [1]. Arrhythmia, characterized by irregular heart rhythm patterns, represents one of the most prevalent cardiac conditions encountered in cardiology practice. This electrophysiological disorder can manifest as either tachyarrhythmia (abnormally fast heartbeat) or bradyarrhythmia (abnormally slow heartbeat), and poses significant health risks if left undetected [2]. Global epidemiological data from 2021 indicate that atrial fibrillation/flutter (AF/AFL) affected approximately 52.55 million individuals worldwide, with an incidence rate of 4.48 million new cases and a prevalence of approximately 620.5 per 100,000 population [3]. The Indonesian Heart Rhythm Society (InaHRS) estimates that arrhythmia prevalence ranges from 1.5% to 5%

Corresponding Author: Tikaridha Hardiani (tikaridha@unisayogya.ac.id)
1 Setia Anfyasa Hadi, Universitas 'Aisyiyah Yogyakarta (setiaanfyasa@gmail.com)
2 Tikaridha Hardiani, Universitas 'Aisyiyah Yogyakarta (tikaridha@unisayogya.ac.id)

of the global population based on 2023 data. Atrial Fibrillation (AF) emerges as the most common arrhythmia type, with current global prevalence at 46.3 million cases. Projections suggest this number will escalate dramatically by 2050, reaching 6-16 million cases in the United States, 14 million in Europe, and 72 million in Asia, including an estimated 3 million cases in Indonesia.

The clinical challenge with arrhythmia lies in its often-asymptomatic presentation. Many patients remain unaware of their condition until it is discovered incidentally during routine cardiac examination [4]. Arrhythmia constitutes an electrophysiological disorder caused by disruptions in either the formation or conduction of electrical impulses within the cardiac tissue [5]. Consequently, early detection becomes crucial for preventing serious complications such as stroke, heart failure, and sudden cardiac death. The primary diagnostic method for arrhythmia detection involves analyzing electrocardiogram (ECG) signals, which capture the electrical activity of the heart.

With the advancement of artificial intelligence and machine learning technologies, various computational methods have been increasingly employed to assist in disease classification and detection. Support Vector Machines (SVMs) have emerged as particularly effective tools in medical signal classification due to their capability to handle high-dimensional and complex datasets [6]. SVMs operate by identifying the optimal hyperplane that maximally separates different data classes, making them highly suitable for binary and multi-class classification problems. However, the effectiveness of SVM models heavily depends on the selection of appropriate features from the input data.

High-dimensional medical datasets often contain numerous irrelevant or redundant features that can negatively impact model performance by increasing computational complexity and reducing generalization capability. Recursive Feature Elimination (RFE) addresses this challenge by systematically removing less important features through an iterative process [7]. RFE works by recursively training the model, ranking features based on their contribution to classification accuracy, and eliminating the least significant ones. This approach not only improves model accuracy but also enhances interpretability and reduces computational overhead [8].

This research presents several novel contributions to the field of arrhythmia detection, Integrated Approach: We propose a comprehensive framework that combines SVM with RFE specifically optimized for arrhythmia detection, addressing the challenge of high-dimensional ECG data while maintaining high classification performance, Comparative Kernel Analysis: We conduct a systematic evaluation of both Linear and RBF kernels across multiple train-test split scenarios, providing insights into which kernel functions best capture the non-linear patterns inherent in arrhythmia data, Feature Importance Analysis: We identify and validate the most critical ECG parameters for arrhythmia detection through RFE, contributing a better understanding of which medical attributes are most predictive of cardiac arrhythmias, Clinical Applicability: We demonstrate practical implementation strategies, including data balancing using SMOTE and normalization techniques that can be readily adopted in clinical decision support systems.

Rationale for technique selection, the choice of SVM as the primary classification algorithm is justified by its proven track record in medical diagnosis applications, particularly its ability to handle high-dimensional data and non-linear decision boundaries. SVM's kernel trick enables the transformation of non-separable data into higher-dimensional spaces where linear separation becomes possible. RFE was selected as the feature selection technique because it considers feature dependencies and evaluates features in the context of the entire model, unlike univariate methods that assess features independently. This integrated approach addresses the specific challenges of arrhythmia detection: high dimensionality of ECG data (278 features), class imbalance, and the need for interpretable results in medical applications.

The primary objective of this study is to develop and validate an effective arrhythmia

detection system by leveraging SVM's classification capabilities enhanced through RFE-based feature selection. By optimizing feature selection, we aim to improve diagnostic accuracy while reducing data complexity, ultimately contributing to better patient outcomes through earlier and more accurate arrhythmia identification.

2. Related Works

To support the development of this research, it is necessary to outline the findings of previous research published in relevant journals. These findings will serve as a reference in formulating the research concept. Research conducted by [7] on breast cancer classification using SVM and RFE resulted in an accuracy of 0.98, precision of 1.00, recall of 0.94, and F-1 score of 0.97. Compared to employing all features, RFE has allowed for a 50% feature reduction without compromising model performance. The evaluation value of the SVM study for heart disease prediction by [1] was demonstrated by an accuracy value of 0.85, a precision of 0.93, a recall of 0.76, and an F-1 score of 0.83. In research [2] on arrhythmia detection using the development of the Deep Neural Network (DNN) algorithm that supports increasing the accuracy of arrhythmia classification by classifying ECG signals, resulting in the best accuracy validation obtained of 71.91% and a loss of 0.6647. Research [9] using the K-Nearest Neighbor (KNN) algorithm for early detection of arrhythmia, resulting in 90% accuracy from the extraction results with the PQRST feature, then classified using KNN. Research [10] shows the results of heart disease classification using the SVM method; the SVM model has superior performance in classifying heart disease patients who are potentially affected by heart disease, with an accuracy level of more than 90%.

In a study using a combination of SVM and FS to optimize heart failure prediction, the Radial Kernel showed the best performance after FS application with the highest AUC of 0.881, accuracy of 84.64%, precision of 86.51%, and recall of 92.55%. The Dot Kernel provided a significant increase in accuracy (84.30%) and recall (95.12%) after FS application, although the AUC decreased slightly to 0.837. The Polynomial Kernel experienced a moderate increase with an accuracy of 79.93% and a recall of 95.57%, but had a lower AUC (0.801) than the other kernels. A study [12] using the Artificial Neural Network method for heart disease classification produced an accuracy value of 73.77% and a precision value of 80.43%, a recall of 84.09% and an F-measure of 82.22%. A study [13] showed that the random forest approach obtained a high degree of accuracy with a 94% classification rate for heart illness [11].

Research on heart disease classification using the Naïve Bayes algorithm, the implementation of Naïve Bayes with K-Fold Cross Validation showed optimal performance at K=4 (accuracy 85.1%, precision 81.1%, recall 86.1%, AUC-ROC 0.914). The model was consistent across all scenarios, with stable AUC (0.91–0.92) and recall always higher than precision. The K=4 confusion matrix showed high accuracy in classification with minimal error. In research [15] on the classification of heart attack patient conditions that lead to ventricular arrhythmia based on QT dispersion using logistic regression, the performance of the logistic regression method with stratified k-fold cross-validation was good, with an average accuracy value of 0.805 and an average AUC value of 0.891 [14].

Another study conducted a comparative study of CNN, SVM, and other machine learning algorithms for intrusion detection. The study also emphasizes the importance of proper feature selection and algorithm optimization in improving classification accuracy. In the context of arrhythmia detection, the integration of SVM with Recursive Feature Elimination (RFE) aims to achieve similar performance improvement by selecting the most relevant ECG features [16]. An article applied CNN and several machine learning algorithms, including SVM, for cardiovascular disease classification. Their findings highlighted that feature selection and model optimization play a crucial role in improving diagnostic accuracy. Building on this insight, the present study integrates Recursive Feature

Elimination (RFE) with SVM to enhance the detection performance of arrhythmia based on ECG data [17]. Another work applied a Random Forest model combined with semi-supervised learning and SMOTE to classify public sentiment about the Indonesian animated film JUMBO. Their study demonstrated that integrating resampling and feature optimization significantly improved model accuracy on imbalanced datasets. Inspired by this, the present research adopts Recursive Feature Elimination (RFE) with SVM to enhance feature relevance and improve classification performance in arrhythmia detection.

3. Proposed Method

In this study, we adopted the SVM algorithm as a supervised learning technique that constructs an optimal hyperplane to separate data points from multiple classes with maximum margin. The fundamental principle of SVM is to find the decision boundary that maximizes the distance between different classes, thereby improving generalization capability on unseen data. For a training dataset $D = \{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in R^d$ and $y_i \in$ {-1,+1}, the optimization problem can be formulated as:

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + c\sum_{i=1}^{n} \xi i \quad (1)$$

Equation 2 formulates the SVM constraints as:

$$y_i(w^T x_i + b) \geq 1 - \xi_i, \ \xi_i \geq 0 \quad (2)$$

where w is the weight vector, b is the bias, ξ_i are slack variables, and C is a regularization parameter that controls the trade-off between maximizing the margin and minimizing classification error.
Where:

- w: weight vector defining the orientation of the separating hyperplane
- b: bias term determining the position of the hyperplane
- ξ_i: slack variables allowing for soft margin classification by permitting some misclassification
- C: regularization parameter controlling the trade-off between maximizing the margin and minimizing classification error

The parameter C plays a crucial role in model performance. A large C value leads to a smaller margin but fewer misclassifications in the training data, potentially causing overfitting. Conversely, a small C value produces a larger margin but allows more training errors, which may result in underfitting.

In this study, we gathered medical data with exhibits non-linear patterns that cannot be separated by linear decision boundaries. Kernel functions address this limitation by implicitly mapping the original feature space into higher-dimensional spaces where linear separation becomes feasible.

In this paper, we utilize the Linear kernel represents the simplest form, computing the inner product of two vectors directly:

$$K(x_i, x_j) = x_i^T x_j \quad (3)$$

The Linear kernel is computationally efficient and works well when the data is linearly separable or nearly linearly separable. It is particularly suitable when the number of features is very large relative to the number of samples. The RBF kernel, also known as the Gaussian kernel, is one of the most popular kernel functions for non-linear classification:

$$K(x_i, x_j) = exp\left(-\gamma \left\| x_i - x_j \right\|^2\right) \quad (4)$$

Where:

- γ (gamma): parameter controlling the width of the Gaussian function.
- $\left\| x_i - x_j \right\|^2$: squared Euclidean distance between feature vectors.

The gamma parameter determines the influence of individual training samples. A small gamma value means a large similarity radius, leading to smoother decision boundaries, while a large gamma creates more complex, tightly fitted boundaries around individual points. The RBF kernel can handle cases where the relationship between class labels and attributes is non-linear, making it particularly suitable for complex medical datasets. After training, the SVM model makes predictions using the following decision function:

$$y = sign\left(\sum_{i=1}^{n} \alpha_i \, y_i K(x_i, x) + b\right) \quad (5)$$

Where:

- α_i: Lagrange multipliers are obtained during training through quadratic programming optimization.
- $K(x_i, x)$: kernel function computing similarity between training samples x_i and test sample x
- $sign$: function returning +1 or -1 based on the sign of its argument.

The Lagrange multipliers α_i determine which training samples become support vectors (samples with $\alpha_i > 0$). These support vectors are the critical data points that define the decision boundary. Most training samples have $\alpha_i = 0$ and do not influence the final decision function, making SVM memory-efficient during prediction.

RFE is a wrapper-based feature selection technique that systematically removes less important features to determine the optimal subset for model construction. Unlike filter-based methods that evaluate each feature independently, RFE accounts for feature interdependencies by assessing their relevance within the context of the learning algorithm. The process begins with initial model training, where an SVM model is trained using all available features. Feature ranking is performed by computing importance scores derived from the model's weight coefficients. This iterative cycle of ranking, elimination, and retraining continues until the desired number of features is reached, resulting in a refined and high-performing model that balances predictive accuracy with computational efficiency.

For linear SVM models, the importance of each feature is determined by the squared weight:

$$Rangking(f_j) = |w_j|^2 \quad (6)$$

Where:

f_j: the j-th feature

w_j: weight coefficient associated with the feature f_j in the SVM model

Features with larger $|w_j|^2$ values contribute more to the decision function and are therefore considered more important. This criterion provides a direct measure of each feature's contribution to the classification decision. RFE offers several key advantages that make it a robust technique for feature selection in machine learning applications. It

performs contextual evaluation by assessing each feature in relation to others, effectively capturing dependencies and interactions that might otherwise be overlooked. Since it is model-specific, RFE tailors the feature selection process to the learning algorithm in use, optimizing overall model performance. Through iterative refinement, RFE systematically removes less relevant features, enabling the model to adjust and stabilize with each iteration. This process not only simplifies the model but also improves generalization by reducing overfitting caused by redundant or noisy features.

The integrated SVM-RFE framework combines the strengths of SVM classification and RFE-based feature selection in a unified methodology. The process begins with data preprocessing, where the dataset is cleaned, normalized, and balanced to ensure consistency. Next, RFE identifies the optimal subset of features, which are then used to train an SVM classifier employing both Linear and Radial Basis Function (RBF) kernels. Hyperparameter tuning is performed using cross-validation to optimize parameters such as (C) and (\gamma). Finally, the model undergoes a comprehensive evaluation on independent test data to assess accuracy, sensitivity, and specificity. This integrated approach effectively tackles the core challenges of arrhythmia detection, including managing high-dimensional ECG data, addressing class imbalance, and ensuring clinical interpretability.

4. Experimental Setup

4.1 Dataset

This study utilized the Arrhythmia dataset from the UCI Machine Learning Repository, a widely recognized benchmark dataset in cardiovascular research. The dataset comprises 452 patient records, each characterized by 278 medical parameters derived from electrocardiogram (ECG) measurements and patient demographic information. These parameters include various ECG features such as heart rate, QRS duration, P-R interval, Q-T interval, T wave characteristics, and other morphological measurements from different ECG leads.

The target variable categorizes patients into 16 different classes representing various types of arrhythmias and one normal class. However, the dataset exhibits severe class imbalance, with the majority of samples belonging to the normal class while other arrhythmia types have relatively few instances. This imbalance poses significant challenges for machine learning algorithms and necessitates careful handling through appropriate preprocessing techniques. The high dimensionality (278 features) combined with relatively modest sample size (452 records) creates the risk of overfitting, making feature selection crucial for developing robust classification models.

4.2 Data Preprocessing

Comprehensive preprocessing is essential to improve data quality and ensure optimal model performance. The following preprocessing steps were systematically applied:

- **Data Cleaning**: Missing values in the dataset were denoted by the "?" symbol. These missing values were first replaced with NaN (Not a Number) to enable proper handling. Subsequently, records with missing values were analyzed. Depending on the proportion of missing data, we employed either imputation strategies (for features with few missing values) or removal (for records with excessive missing values). This cleaning process ensures data integrity and prevents errors during model training.
- **Normalization**: Feature scaling is critical when using SVM algorithms because they are sensitive to the scale of input features. We applied Min-Max normalization

to scale all features into the [0,1] range using the following formula:

$$Normalisasi = \frac{data - \min(data)}{\max(data) - \min(data)} \quad (7)$$

Where data is the original feature value, min(data) is the minimum value in the feature column and max(data) is the maximum value in the feature column. This normalization ensures that features with larger numerical ranges do not dominate the distance calculations in SVM, allowing all features to contribute proportionally to the classification decision. The transformation preserves the relationships between data points while standardizing the scale.

The severe class imbalance in the arrhythmia dataset could lead to models biased toward the majority class, resulting in poor detection of minority arrhythmia types. To address this issue, we employed the Synthetic Minority Oversampling Technique (SMOTE). SMOTE generates synthetic samples for minority classes by interpolating between existing minority class instances and their nearest neighbors. This approach creates realistic synthetic samples that lie along the line segments connecting minority class samples in feature space, effectively balancing the class distribution without simply duplicating existing samples. SMOTE helps prevent overfitting while enabling the model to learn meaningful patterns from underrepresented arrhythmia types.

4.3 Experimental Scenarios

To comprehensively evaluate model performance across different data availability conditions, we designed four experimental scenarios based on varying train-test split ratios. These scenarios simulate different practical situations where varying amounts of training data might be available:

- Scenario 1: 90% training, 10% testing
- Scenario 2: 80% training, 20% testing
- Scenario 3: 70% training, 30% testing
- Scenario 4: 60% training, 40% testing

Both SVM with Linear kernel and SVM with RBF kernel were trained under these scenarios, with and without RFE-based feature selection.

4.4 Evaluation Metrics

The models' performance was assessed using metrics included in the Confusion Matrix, including Accuracy, Precision, Recall, and F1-score. These are explained as:

1. Accuracy
   Accuracy is the frequency with which a model makes accurate predictions for both positive and negative classifications. It is the most commonly used evaluation statistic and provides a broad picture of a model's overall performance.
   $$Accuracy = \frac{TP + TN}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \; x\; 100\% \quad (8)$$
2. Precision
   The precision of a model's positive predictions is determined by its accuracy. This shows what percentage of data are actually positive as opposed to those that were predicted to be such.
   $$Precision = \frac{TP}{\mathrm{TP} + \mathrm{FP}} \; x\; 100\% \quad (9)$$
3. Recall
   The model's recall gauges its capacity to identify all positive data. How many of all the genuinely positive facts were accurately predicted

$$Recall = \frac{TP}{\text{TP} + \text{FN}} \; x \; 100\% \tag{10}$$

4. F1-Score
   The harmonic mean of recall and precision is used to compute the F1-score. This metric is used when precision and recall need to be balanced, especially in uneven data.

$$F1 = 2 \text{ x } \frac{Precision * Recall}{\text{Precision} + \text{Recall}} \; x \; 100\% \tag{11}$$

   where the numbers of correctly predicted positive cases are denoted by TP (True Positives), correctly predicted negative cases by TN (True Negatives), incorrectly predicted positive cases by FP (False Positives), and incorrectly predicted negative cases by FN (False Negatives). The suggested strategy may be thoroughly assessed in terms of classification accuracy, sensitivity, and robustness to data imbalance thanks to this experimental design.

5. Result and Analysis

The dataset was split into four scenarios: 90:10, 80:20, 70:30, and 60:40 for training and testing. Table 5.1 shows the number of training and testing samples for each split.

Table 5.1 Dataset Splitting Scenarios

Scenario	Training Dataset	Testing Dataset
90:10	407	45
80:20	362	90
70:30	316	136
60:40	271	181

Both linear and RBF kernels were used to test the SVM models. Evaluation measures include F1-score, recall, accuracy, and precision. Table 5.2 presents the performance results for each scenario using the RBF kernel and Linear kernel with optimal hyperparameters.

Table 5.1. Performance comparison of SVM with Linear and RBF kernels

Scenario	Kernel	Accuracy	Precision	Recall	F1-Score
90:10	Linear	85.21%	82.00%	87.00%	84.43%
90:10	RBF	91.30%	88.00%	95.65%	91.67%
80:20	Linear	82.60%	80.00%	84.00%	81.96%
80:20	RBF	88.40%	85.00%	90.00%	87.40%
70:30	Linear	80.10%	78.00%	81.00%	79.47%
70:30	RBF	85.60%	83.00%	87.00%	84.97%
60:40	Linear	78.20%	75.00%	79.00%	77.00%
60:40	RBF	82.50%	80.00%	84.00%	81.96%

The experimental results reveal several important patterns. The RBF kernel consistently outperformed the Linear kernel across all scenarios, demonstrating its superior capability in capturing non-linear relationships inherent in arrhythmia ECG data. The performance advantage of RBF over the Linear kernel ranged from 3.0% to 6.1% in accuracy, with the largest gap observed in Scenario 1 (90:10 split). Scenario 1 (90:10 split) yielded the highest

performance metrics with the RBF kernel achieving 91.30% accuracy, 88.00% precision, 95.65% recall, and 91.67% F1-score. The exceptionally high recall value of 95.65% is particularly significant for medical diagnosis, indicating that the model successfully identified 95.65% of actual arrhythmia cases, thereby minimizing dangerous false negatives.

Progressive performance degradation was observed in the training data decreased from Scenario 1 to Scenario 4. The RBF kernel's accuracy declined from 91.30% to 82.50%, while the Linear kernel dropped from 85.21% to 78.20%. This trend confirms the expected relationship between training data quantity and model performance, though the RBF kernel demonstrated greater resilience to reduced training data, maintaining a more gradual performance decline. The recall values for the RBF kernel remained consistently high across all scenarios (ranging from 84.00% to 95.65%), indicating robust sensitivity to arrhythmia detection even with varying amounts of training data. This characteristic is crucial for clinical applications where missing true arrhythmia cases could have severe health consequences for patients.

Recursive Feature Elimination successfully reduced the original 278 features to an optimal subset of 10 features, achieving 96.4% dimensionality reduction while maintaining or improving classification performance. The selected features were: ['GE', 'HY', 'IT', 'IV', 'JH', 'JL', 'JO', 'KB', 'KK', 'KL']. These feature codes represent specific ECG parameters and patient characteristics identified as most predictive for arrhythmia classification. Table 5.3 presents the feature importance scores computed by RFE for both Linear and RBF kernels, revealing how different kernel functions emphasize different aspects of the feature space.

Table 5.3. Feature Importance Scores

No	Feature	Importance (Linear)	Importance (RBF)
1	IV	0.195	0.475
2	JH	0.148	0.025
3	IT	0.143	0.000
4	KK	0.112	0.000
5	HY	0.111	0.000
6	GE	0.111	0.200
7	JL	0.079	0.000
8	JO	0.071	0.300
9	KL	0.026	0.000
10	KB	0.003	0.000

In the Linear kernel, IV, JH, and IT were the most significant contributors, while in the RBF kernel, IV, JO, and GE were more dominant. This highlights how different kernels capture different aspects of the feature space. The feature importance analysis reveals distinct patterns between kernel functions. In the Linear kernel model, importance scores are distributed relatively evenly among the top features, with IV (0.195), JH (0.148), and IT (0.143) emerging as the most significant contributors. This distribution suggests that Linear SVM relies on multiple features working together to establish the decision boundary.

In contrast, the RBF kernel shows highly concentrated importance scores, with IV (0.475), JO (0.300), and GE (0.200) dominating while many other features have zero or near-zero importance. This concentration indicates that the RBF kernel identifies specific critical features that capture the non-linear patterns in arrhythmia data. The stark difference in feature importance distributions between kernels highlights how kernel choice fundamentally affects which aspects of the data the model emphasizes.

Feature IV consistently demonstrates the highest importance across both kernels, suggesting it represents a universally critical ECG parameter for arrhythmia detection regardless of the modeling approach. The fact that JH and IT show high importance in Linear but minimal importance in RBF, while JO and GE show the opposite pattern,

indicates these features contribute differently to linear versus non-linear decision boundaries.

6. Conclusion

This study successfully demonstrated that integrating SVM with Recursive Feature Elimination provides an effective approach for automated arrhythmia detection from ECG data. Through systematic experimentation across four train-test split scenarios, the RBF kernel consistently outperformed the Linear kernel in capturing complex non-linear patterns in arrhythmia data. The optimal performance was achieved in the 90:10 scenario with 91.30% accuracy, 88.00% precision, 95.65% recall, and 91.67% F1-score. The exceptionally high recall value is particularly significant for medical diagnosis, minimizing the critical risk of missing actual arrhythmia cases. According to the experimental results, the RFE successfully reduced dataset dimensionality by 96.4%, selecting only 10 optimal features (GE, HY, IT, IV, JH, JL, JO, KB, KK, KL) from 278 original parameters while maintaining high classification accuracy. This reduction enhances computational efficiency, improves model interpretability for clinical use, and decreases overfitting risk. The comprehensive preprocessing pipeline, including data cleaning, Min-Max normalization, and SMOTE-based class balancing, proved essential for achieving optimal performance. The proposed SVM-RFE framework represents a viable approach for arrhythmia detection that could be integrated into clinical decision support systems.

Acknowledgment

The authors would like to thank Allah SWT from the bottom of their hearts for giving them the strength and blessings to finish this research. We would also want to thank the information technology department's professors and supervisors for their important advice and unwavering support during the study. The writers also thank their coworkers and family for their support.

References

[1] R. Hidayat, Y. S. Sy, T. Sujana, M. Husnah, and H. T. Saputra, “Implementation of Machine Learning for Heart Disease Prediction Using Support Vector Machine Algorithm,” vol. 5, no. 2, pp. 161–168, 2024.

[2] M. Fajar and Z. Nugraha, “Arrhythmia Detection Using Deep Neural Network (DNN) Algorithm on Electrocardiogram Signals,” *eProceedings …*, vol. 10, no. 5, pp. 4155–4158, 2023, [Online]. Available: https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/download/21141/20704

[3] X. Li, Z. Li, H. He, S. Wang, H. Su, and G. Kang, “Global burden and health inequality of atrial fibrillation/atrial flutter from 1990 to 2021,” *Front. Cardiovasc. Med.*, vol. 12, no. May, pp. 1–19, 2025, doi: 10.3389/fcvm.2025.1585980.

[4] A. Hanifa, “Expert System for Diagnosing Arrhythmia Disease Using Certainty Factor,” *J. SANTI - Sist. Inf. dan Tek. Inf.*, vol. 2, no. 1, pp. 41–48, 2022, doi: 10.58794/santi.v2i1.63.

[5] A. Musfirah Putri Lukman11), Armin Lawi22), Desi Widyaningsih33), “Arrhythmia Disease Detection System Based on Heartbeat Count Based on Internet of Things and Cloud Storage,” *Pros. Semin. Nas. Tek. Elektro dan Inform.*, vol. 1, pp. 1–6, 2022.

[6] V. Piccialli and M. Sciandrone, “Nonlinear optimization and support vector machines,” *Ann. Oper. Res.*, vol. 314, no. 1, pp. 15–47, 2022, doi: 10.1007/s10479-022-04655-x.

[7] H. Sundari, M. A. Amrustian, A. Dwi, and P. Wicaksono, “Application of Recursive Feature Elimination on Support Vector Machine for Breast Cancer Classification,” vol. 8798, pp. 60–65,

2024.
[8] N. A. Maori *et al.*, “Application of Recursive Feature Elimination (RFE) on Tree-Based Classifier for Diabetes Risk Identification,” vol. 6, no. 2, pp. 465–471, 2024.
[9] J. Given Hamonangan, A. S. Rahmi, S. J. Lase, N. Suhendi Syafei, and A. Turnip, “JIIF (Jurnal Ilmu dan Inovasi Fisika) EARLY DETECTION OF ARRHYTHMIA USING K-NEAREST NEIGHBOR,” vol. 08, no. 01, pp. 86–95, 2024, [Online]. Available: https://doi.org/10.24198/jiif.v8i1
[10] B. G. Sihombing and N. R. Al-fiqri, “Literature Review: Classification of Heart Disease Using Support Vector Machine (SVM) Method,” vol. 2, no. 3, pp. 442–448, 2024.
[11] V. Riyanto, H. Destiana, T. Prihatin, and G. Wijaya, “OPTIMIZING HEART FAILURE PREDICTION WITH A COMBINATION,” vol. 8, no. 1, pp. 103–111, 2025.
[12] D. Pradana, M. Luthfi Alghifari, M. Farhan Juna, and D. Palaguna, “Classification of Heart Disease Using Artificial Neural Network Method,” *Indones. J. Data Sci.*, vol. 3, no. 2, pp. 55–60, 2022, doi: 10.56705/ijodas.v3i2.35.
[13] H. Hidayat, A. Sunyoto, and H. Al Fatta, “Heart Disease Classification Using Random Forest Classifier,” *J. SISKOM-KB (Sistem Komput. dan Kecerdasan Buatan)*, vol. 7, no. 1, pp. 31–40, 2023, doi: 10.47970/siskom-kb.v7i1.464.
[14] D. P. P. Salman, W. Yunanto, and M. Akbar, “Coronary Heart Disease Classification Using the Naive Bayes Algorithm,” *J. Aksara Komput. …*, pp. 115–127, 2017, doi: 10.33364/algoritma/v.22-1.2178.
[15] Y. W. Patmonobo, “Classification of Heart Attack Patient Conditions Leading to Ventricular Arrhythmia Based on QT Dispersion Using Logistic Regression,” *J. UIN Syarif Hidayatullah*, 2022.
[16] M. H. Wathan and M. Aziz, “Establishing CNN for Network Intrusion Detection : A Comparative Approach,” vol. 6, no. 1, 2024, doi: 10.35842/ijicom.
[17] H. Romana and J. S. R, “Enhancing Cardiovascular Diseases Classification using CNN Algorithm,” vol. 5, no. 2, 2023, doi: 10.35842/ijicom.
[18] E. Rahmat, S. Hidayat, and A. C. Frobenius, “Sentiment Analysis of Animated Film ‘ JUMBO ’ on Twitter Using Random Forest and Semi-Supervised Learning,” vol. 7, no. 2, 2025, doi: 10.35842/ijicom.