Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 9, 54-64
2026
Publisher: Learning Gate
DOI: 10.55214/2576-8484.v10i9.13549

A multi-stage data preprocessing framework for ECG arrhythmia classification

Sunil Kumar Bansal[1], Ashwani Kumar Aggarwal[2*]
[1,2]Department of Electrical and Instrumentation Engineering, Sant Longowal Institute of Engineering and Technology, Longowal-148106, Punjab, India; skbansal@sliet.ac.in (S.K.B.) ashwani.ist@sliet.ac.in (A.K.A.).

Abstract: Arrhythmia classification from ECG signals has been widely attempted by researchers using conventional feature-based methods and deep learning approaches. The UCI arrhythmia dataset is widely used for evaluating such classifiers. however, its missing values, outliers, and class imbalance present significant preprocessing challenges that can adversely affect classification performance and interpretability. The focus of the majority of the works is on achieving a fast, accurate, and reliable classifier; however, little effort has been made toward preprocessing the data. This study proposes a multi-stage, class-aware preprocessing framework for robust arrhythmia classification using the UCI dataset. The proposed framework systematically addresses data quality issues and evaluates the contribution of individual preprocessing modules across multiple classification algorithms, including low-complexity support vector machines. Experimental results demonstrate that each preprocessing stage contributes significantly to overall classifier performance. The proposed framework, yielding the best performance with an accuracy of 88.51%, with 0.8905 precision, 0.8851 recall, and 0.8834 F1-score. Furthermore, PCA achieved an accuracy of 85.14% with the added benefits of dimensionality reduction and noise mitigation. These results show the effectiveness of the proposed framework for arrhythmia classification, particularly when dealing with missing, noisy, and imbalanced ECG data.

Keywords: *Arrhythmia classification, Class imbalance, ECG, Feature extraction, Missing data imputation, Outlier detection, UCI dataset.*

1. Introduction

Arrhythmia is a state in which the heart beats too slowly, too rapidly, or irregularly [1]. Some arrhythmia conditions are normal without any harm to a person's health, whereas others are related to an abnormal heart condition [2]. The symptoms of arrhythmia include palpitations, trembling, and fainting, among many others [3]. The normal heart rate ranges from 60–100 beats per minute. If the heart beats too fast, the condition is called tachycardia; if it beats too slowly, the condition is called bradycardia [4]. If extra beats appear along with normal beats, the condition is called premature beats [5]. Sometimes, the frequency at which the heart beats varies continuously or intermittently; the condition is known as irregular heartbeats [6]. Thus, there are four categories of arrhythmia: tachycardia, bradycardia, premature beats, and irregular beats [7].

Arrhythmia detection is very important in deciding the treatment plan for heart patients and as an indicator of the overall health of a person. Arrhythmias can be detected using various invasive and non-invasive methods [8]. Non-invasive methods include standard electrocardiogram (ECG), Holter monitors, vectorcardiography, photoplethysmography, impedance cardiography, echocardiography, and wearable devices such as smartwatches [9]. Invasive methods include electrophysiology, invasive hemodynamic monitoring, and catheter ablation.

Electrocardiogram (ECG) signals represent the electrical activity of the heart [10]. The signals are captured using ECG electrodes placed on the skin. ECG signals are used for the detection as well as

History: Received: 3 June 2026; Revised: 5 August 2026; Accepted: 10 August 2026; Published: 4 September 2026
* Correspondence: ashwani.ist@sliet.ac.in

diagnosis of arrhythmia at an early stage [11]. The choice of ECG signal analysis plays a significant role in the diagnosis. An automated ECG signal classification technique assists cardiologists in making treatment decisions and recommending further tests such as Holter and echo [12]. Several automatic techniques have been proposed by many researchers for arrhythmia classification, which are based on machine learning approaches. These techniques focus either on improving classification accuracy or reducing computational time [13]. The methods use raw ECG data, extract features, and then use these features for classification. The performance of these classification methods heavily depends on the robustness of the extracted features, which in turn depends on the quality of the input data used [14]. Many ECG datasets are available for classification, including both raw ECG datasets and ECG feature datasets. The UCI Arrhythmia dataset suffers from missing values, noise, and outliers. These factors degrade classification accuracy, leading to false arrhythmia detection and diagnosis [15]. The UCI Arrhythmia dataset contains ECG features with many missing values and class imbalance. This paper primarily focuses on preprocessing the UCI dataset using a multi-stage, class-aware methodology for robust arrhythmia classification. The entire framework includes modules on dataset preprocessing, classification using support vector machines, and performance assessment based on various evaluation metrics. The main contribution of the paper is not to develop a novel classification method but to propose a preprocessing framework for the dataset that helps improve the performance of classification methods in terms of classification accuracy, sensitivity, AUC-ROC, and F1-score, etc.

The rest of the paper is organized as follows. Related work is given in Section 2. Details of the dataset and its preprocessing techniques are provided in Section 3. Section 4 presents the methodology adopted in the work. Section 5 gives a discussion of the results obtained using the proposed framework. Section 6 concludes the work and its prospects.

2. Related Work

A review of deep learning-based arrhythmia detection and classification is given in [16]. The paper focuses on arrhythmia classification methods in which arrhythmia detection is performed using ECG data. Apart from using ECG data for arrhythmia detection, echo, although less common, is another modality sometimes used for this purpose. A semi-supervised machine-learning framework is proposed for arrhythmia detection using echo signals in Hu et al. [17]. The method detects atrial fibrillation by training on a dataset of 260 cines and obtains an accuracy of 79%. However, the other relevant performance metrics have not been evaluated in the paper. A deep learning-based arrhythmia detection and classification method is presented in Hammad et al. [18]. The paper uses continuous ECG signals for arrhythmia detection without the need for segmentation of heartbeat signals. The work can be extended by incorporating efficient and fast feature extraction and processing units so that it can be embedded into wearable devices. To reduce complexity, single-lead ECG data are used for arrhythmia detection in IoT applications in Al-Shammary et al. [19]. The method tackles the problem of model overfitting using batch normalization. The Chi-square distance is used as a fitness function in particle swarm optimization for arrhythmia detection in Dhyani et al. [20]. Unlike conventional machine learning methods, the classifier optimizes feature selection and hence enhances classification accuracy. Wavelet-based ECG data preprocessing, followed by feature extraction and detection, is performed in Liu et al. [21]. The dataset used in the work is the China Physiological Signal Challenge (CPSC) 2018 arrhythmia dataset [22] and He et al. [23]. It is an open-access dataset for evaluating various machine learning models for the detection and classification of arrhythmia. The dataset has been used in Güvenir et al. [24], in which a hybrid deep learning-based classification framework is proposed for arrhythmia classification. The method is evaluated using the F1 score as a performance metric under various conditions.

3. Dataset

This section discusses the description of the dataset used and the preprocessing techniques. The preprocessing techniques include missing data handling, outlier removal, and data normalization. This

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 9: 54-64, 2026
DOI: 10.55214/2576-8484.v10i9.13549

paper uses the arrhythmia classification dataset taken from the UCI Machine Learning Repository [24], which is taken as a benchmark by the research community for arrhythmia classification. The dataset consists of 452 instances of patient records and 279 features derived for each instance. The instances are categorically labelled into 16 arrhythmia classes, including the normal ECG class. The dataset is represented by $X^{(o)} \in \mathbb{R}^{N \times D}$, where $N = 452$ is number of instances and $D = 279$ is the number of features per instance. The class label vector is represented by $y \in \{1,2,\cdots,C\}^N$, where $C = 16$ where is the number of arrhythmia classes. The dataset is summarized in Table 1.

Table 1.
Comprehensive description of the UCI Arrhythmia dataset.

Component	Attribute	Details
Dataset Profile	Dataset Name	UCI Arrhythmia [24]
	Data Domain	ECG signal + patient information
	Instances	452
	Features	279
	Total Classes	16
Data Characteristics	Feature Type	Mixed (numeric + categorical ECG features)
	Majority Class	Class 1 (Normal): 245 samples (~54%)
	Minority Classes	Several classes < 1%
	Imbalance Ratio	> 100:1 (severely skewed)
	Missing Values	Present in multiple features
	Noise	High (ECG signal variability)
	Dimensionality	High (risk of overfitting)
Clinical Relevance	Data Source	ECG recordings + patient metadata
	Diagnostic Scope	Multiple arrhythmia types
Learning Challenges	Key Issues	Class imbalance, missing data, high dimensionality
	Preprocessing	Imputation, normalization, feature selection
	Evaluation Concern	Bias toward majority class

Source: Güvenir et al. [24].

The features are extracted from multivariate data such as raw ECG, heart rate, and anthropometric data. These features are considered the gold standard for various machine learning tools for arrhythmia classification.

4. Methodology

This study presents a robust framework for early heart disease prediction using automated electrocardiogram (ECG) arrhythmia classification. The proposed approach enables accurate analysis of various cardiac patterns, even with missing and imbalanced data. The experiments are performed on the UCI Arrhythmia dataset, which has multiple attributes related to ECG signals and clinical features.

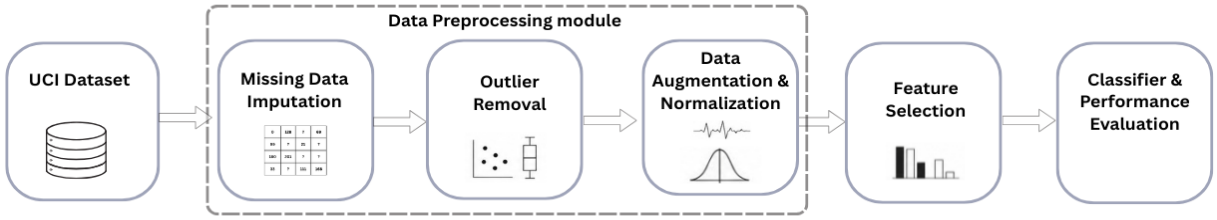


Figure 1.
Multi-stage UCI Dataset Classification pipeline.

A multi-stage classification pipeline, as shown in Figure 1, is proposed to systematically address the challenges associated with noisy, incomplete, and imbalanced data. The shortcomings in the dataset are handled in the preprocessing module using appropriate imputation techniques and by balancing the

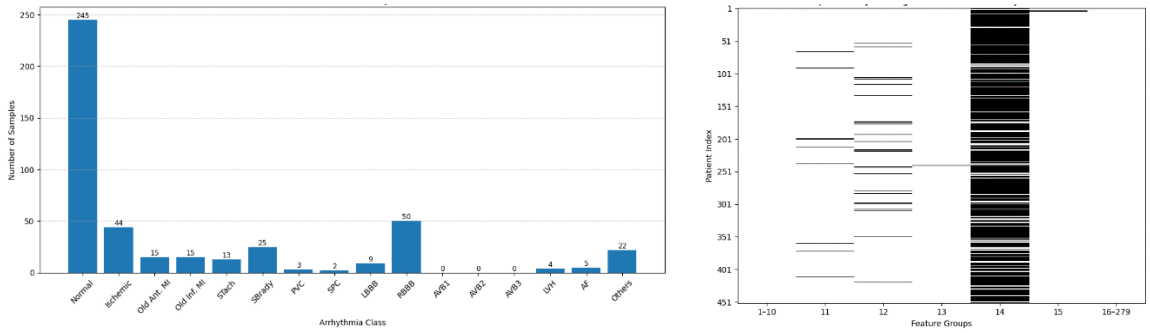
dataset through data augmentation. The relevant features are then selected and used for classification to enhance prediction performance and model robustness.

4.1. Data Preprocessing

The open-source datasets, including the UCI arrhythmia dataset [24], available for research purposes often suffer from various shortcomings, such as missing feature values, non-uniform scaling, data imbalance, large dimensionality of the feature vector, sensor noise, and the presence of outliers. These shortcomings may negatively affect model training and generalization. Therefore, these shortcomings in the dataset should be addressed appropriately before using it for classification.

4.1.1. Missing Data Handling

The UCI arrhythmia dataset contains irregular patterns of missing data across features. Figure 2 (a) represents the general class distribution of the UCI arrhythmia dataset, and a binary indicator representing the missingness pattern in the dataset is shown in Figure 2 (b).



(a) (b)

Figure 2.
(a) Class distribution of UCI arrhythmia dataset; (b) Binary visualization of feature-wise missing data.

The ECG-derived features are sparse and have a non-uniform distribution. These characteristics of the dataset can bias learning, reduce statistical power, and degrade machine learning model generalization. Thus, handling missing data requires a structural imputation approach to preserve feature relationships rather than simply removing the feature. In this section, each feature was analyzed for missing values. The missing ratio r_j of the j^{th} feature is computed as given in Equation (1).

$$r_j = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left(x_{ij}\ is\ missing\right) \qquad (1)$$

where $\mathbb{I}(\cdot)$ is the indicator function. The feature $x_{i,14}$ exhibited substantial missingness of 83%, whereas other features contained less than 5% missing data. One of the naïve methods is to remove a feature from all instances if the feature is missing in any instance. This method skips features that may be significant for model training. Moreover, it will also reduce the dataset significantly. This work employs a class-wise median imputation technique to address the above issue. Unlike the global imputation method, this approach will maintain the inherent characteristics of the dataset and preserve class-specific data distribution. Since, $x_{i,14}$ feature has more than 80% missing data, so $x_{i,14}$ is removed from further investigation, and hence the dataset is reduced to $X^{(A)} \in \mathbb{R}^{452\times278}$. Since the remaining features have missingness below 5%, an appropriate imputation technique is applied to fill the missing elements. For each class $c \in \{1,2,\cdots,C\}$, let $I_c = \{i \mid y_i = c\}$, let be the index set of samples belonging to class c. The imputed value $x_{ij}^{(I)}$ is defined in equation (2).

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 9: 54-64, 2026
DOI: 10.55214/2576-8484.v10i9.13549

$$x_{ij}^{(I)} = \begin{cases} x_{ij}^{(A)}, & \text{if } x_{ij}^{(A)} \text{ is available} \\ m_{cj}, & \text{if } x_{ij}^{(A)} \text{ is missing and } y_i = c \end{cases} \quad (2)$$

where the class-wise median for feature j is computed as $m_{c,j} = median_{i \in I_c}\left(x_{ij}^{(A)}\right)$. Median imputation was selected over other imputation techniques, viz., mean, mode, etc., due to its robustness to outliers. After imputation, the dataset $X^{(I)} \in \mathbb{R}^{452\times278}$ does not have any missing values.

4.1.2. Outlier Removal

The imputed dataset $X^{(I)}$ contains several outliers due to various factors such as sensor noise, electromagnetic interference, and improper electrode placement, etc. The presence of extreme or inconsistent samples can adversely affect classifier performance. Several outlier methods are available in the literature to mitigate the influence of extreme values, such as interquartile range (IQR), RANSAC, kNN, and Isolation Forest. Since the UCI dataset is multivariate with high dimensionality, univariate or distance-based outlier detection methods are less reliable. Therefore, the Isolation Forest Outlier approach is used, which isolates anomalies by recursively partitioning the feature space. For each class c, an Isolation Forest model is applied in a class-wise manner to the scaled feature subset $X^{(c)} \subset X^{(I)}$ to avoid removing minority class samples disproportionately. For each class, samples were classified with a contamination rate of 0.025, as given in equation (3).

$$\hat{y}_i = \begin{cases} 1, & inlier \\ -1, & outlier \end{cases} \quad (3)$$

Classes with fewer than 10 instances are exempted from outlier removal to preserve rare arrhythmia patterns. After class-wise filtering, the cleaned dataset size was reduced to $X^{(c)} \in \mathbb{R}^{440\times278}$.

4.1.3. Data Augmentation & Normalization

The classification accuracy suffers due to class imbalance and the small size of the UCI arrhythmia dataset. The class imbalance is partly handled by using missing data. The issue is further tackled by augmentation of data in each class so that the number of instances corresponding to each class becomes the same. The cleaned dataset $X^{(c)}$ exhibits severe data imbalance, where certain arrhythmia categories contain very few samples compared to the dominant normal class, as depicted in Figure 3. Let n_c denote the number of samples in class as given in equation (4).

$$n_c = \sum_{i=1}^{N} \mathbb{I}(y_i = c) \quad (4)$$

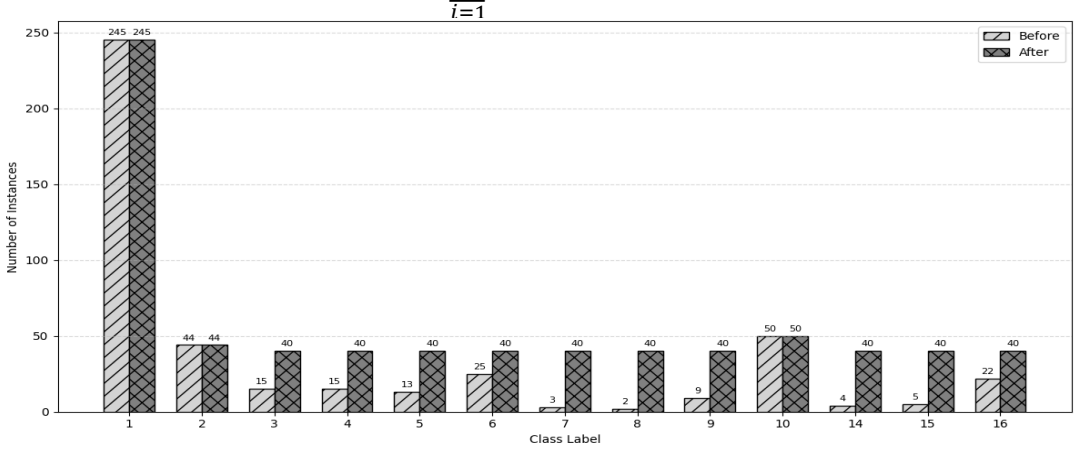


Figure 3.
Effect of data Augmentation on Class-wise Distribution.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 9: 54-64, 2026
DOI: 10.55214/2576-8484.v10i9.13549

This imbalanced dataset might create false model training due to insufficient features in some classes. To overcome this shortcoming, a class-wise, median-driven augmentation strategy is proposed. For each class c, if $n_c < 40$, synthetic samples are generated with bounded uniform noise around the class median, as given in equation (5). This approach preserves the central tendency of each class while introducing controlled variability.

$$x_j^{(C)} = m_{c,j} + \alpha . \epsilon_j . s_{c,j} \quad (5)$$

where the scaling factor α controls the overall magnitude of perturbation. Its value is taken as 0.05 empirically. $\epsilon_j \sim \mathcal{U}(-1,1)$ is uniformly distributed noise to incorporate variability across synthetic samples. The factor $s_{c,j} = \max(|m_{c,j}|, \epsilon)$ is a magnitude-based scaling factor which adapts to the noise commonly present in ECG signals. It helps in preventing near-zero features from collapsing as well as preserving variability across features.

The classes with no instances are excluded from this study. The class distribution for the final dataset $X^{(F)} \in \mathbb{R}^{739\times278}$, $y^{(F)} \in \{1,2,\cdots,C\}^{739}$, became substantially more balanced, as shown in Fig. 3. The UCI arrhythmia dataset contains features based on ECG amplitude in small decimal values, time intervals in larger values, as well as categorical indicators. Therefore, normalization is required, as large-value features may bias the classifier. Many options for data normalization, such as min-max normalization, Z-score normalization, and robust scaling involving the interquartile range, are available in the literature. In this paper, based on the characteristics of the data mentioned above, Z-score normalization is applied, as it brings all features to values with a mean value of 0 and a standard deviation of 1, which makes each feature contribute equally to the classification. This data preprocessing pipeline handles missing data, reduces noise, and overcomes the data imbalance problem. It improves the reliability of downstream arrhythmia classification models.

4.2. Feature Selection

Feature selection is a crucial step in a classification task because it eliminates irrelevant and redundant features. If the number of features is greater, the classifier more robustly classifies the data; however, this leads to higher computational cost. On the other hand, if the number of features is fewer, the robustness of the classifier decreases, while its computational cost is lower. Therefore, a criterion that selects the appropriate features from a set of features is very important in classification. Several feature selection methods, such as information-theoretic methods, Independent Component Analysis (ICA), Linear Discriminant Analysis (LDA), and Principal Component Analysis (PCA), etc., are used for reducing dimensionality and enhancing machine learning performance. Since the UCI arrhythmia dataset is a multiclass dataset with intra-/inter-class variance, the robustness of the classifier highly depends on the feature selection strategy based on variance. It is desired that the intra-class variance should be statistically small in comparison with the inter-class variance. Thus, the SelectKBest method using ANOVA is a suitable candidate method for selecting the relevant and distinct features. The method is simple, computationally efficient, and preserves the original meaningful feature space. The features are selected based on statistical relevance using ANOVA. Although the method appears to be a good choice for feature selection, a few points make the SelectKBest method not a first choice for feature selection. Among the many limitations of SelectKBest are that this method does not involve feature correlation, is sensitive to noise, and depends on the statistical test used, etc.

In this work, Principal Component Analysis (PCA) is used to transform the original feature space into a lower-dimensional subspace while retaining the most informative structure of ECG features and preserving maximum variance. It helps in reducing noise and improving classifier efficiency. The final normalized dataset is represented as $X^{(F)} \in \mathbb{R}^{739\times278}$. Covariance Matrix $\Sigma \in \mathbb{R}^{278\times278}$ is computed, and PCA performs eigen decomposition on the covariance matrix. The projection matrix W_K is formed using the top K eigenvectors corresponding to the largest eigenvalues. The reduced feature representation is obtained as $X^{(PCA)} = X^{(F)}W_K$, where $X^{(PCA)} \in \mathbb{R}^{739\times K}$. The optimal number of

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 9: 54-64, 2026
DOI: 10.55214/2576-8484.v10i9.13549

components (K=39) is selected empirically to achieve the best trade-off between classification accuracy and computational cost. The transformed feature set $X^{(PCA)}$ is then used as input to the model for arrhythmia classification. The advantages of using PCA-based feature selection over alternative methods are multifold. First, it reduces computational cost, as the number of features is significantly reduced using PCA. This happens because many of the features in the UCI arrhythmia dataset are correlated. Second, the effect of noise, which is commonly present in ECG-based features, is mitigated. Third, as PCA maps the original feature space into a new feature space that keeps only the most informative variance components, the stability of the model is enhanced. The PCA-mapped feature set is thereafter fed to the SVM classifier for arrhythmia classification.

4.3. Classifier

The transformed feature space obtained after PCA is used for arrhythmia classification using a Support Vector Machine (SVM). The choice of using SVM is made due to many factors, which include the robustness of SVM toward high-dimensional data, support for small datasets, and tolerance to imbalanced datasets. As the UCI arrhythmia dataset is multiclass, with the number of classes being 16, a multiclass SVM classifier is used for the task. Let $y^{(F)}$ be the labels corresponding to all the instances in $X^{(PCA)}$. The objective is to estimate an optimal hyperplane that separates each class in the dataset. The optimization problem may be formulated as given in Equation (6).

$$min_{w,b,\zeta}\ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\zeta_i \tag{6}$$

$$\text{subject to } y_i = (w^T x_i + b) > 1 - \zeta_i.$$

Where ζ_i are slack variables for handling misclassification cases due to difficult samples in the dataset. C is the regularization parameter, which controls the trade-off between maximizing the margin of the hyperplane from each class and minimizing the error in classification. The classification performance is evaluated using various metrics, such as accuracy, sensitivity, precision, specificity, F1-score, Matthews Correlation Coefficient (MCC), Cohen's Kappa Coefficient, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC), etc., as discussed in Sec. 5.

5. Results and Discussion

The UCI dataset has several challenges, such as high missingness, redundant data, and non-informative features, etc. Figure 4 illustrates the Pearson correlation heatmap of the first 10 raw features. The correlation analysis is done after missing values were imputed using median substitution, and constant features were removed using variance thresholding. There are several localized regions that show stronger positive and negative correlations. This indicates the presence of redundant feature groups obtained from related ECG measurements. The correlation analysis helps to choose preprocessing steps in reducing the effects of irrelevant or redundant attributes.

Statistical analysis is used to conclude a hypothesis. Several statistical tests, such as t-tests and p-values, are used. Statistical analysis helps in evaluating the robustness, reliability, and use of different methodologies for challenges under class-imbalance conditions. Statistical analysis can be done using descriptive statistics or inferential statistics. Descriptive statistics are used to summarize characteristics of the dataset, which include class distribution, the percentage of missing values, and variability in the data. Inferential statistics are used to evaluate the performance of the classifier under various preprocessing stages.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 9: 54-64, 2026
DOI: 10.55214/2576-8484.v10i9.13549

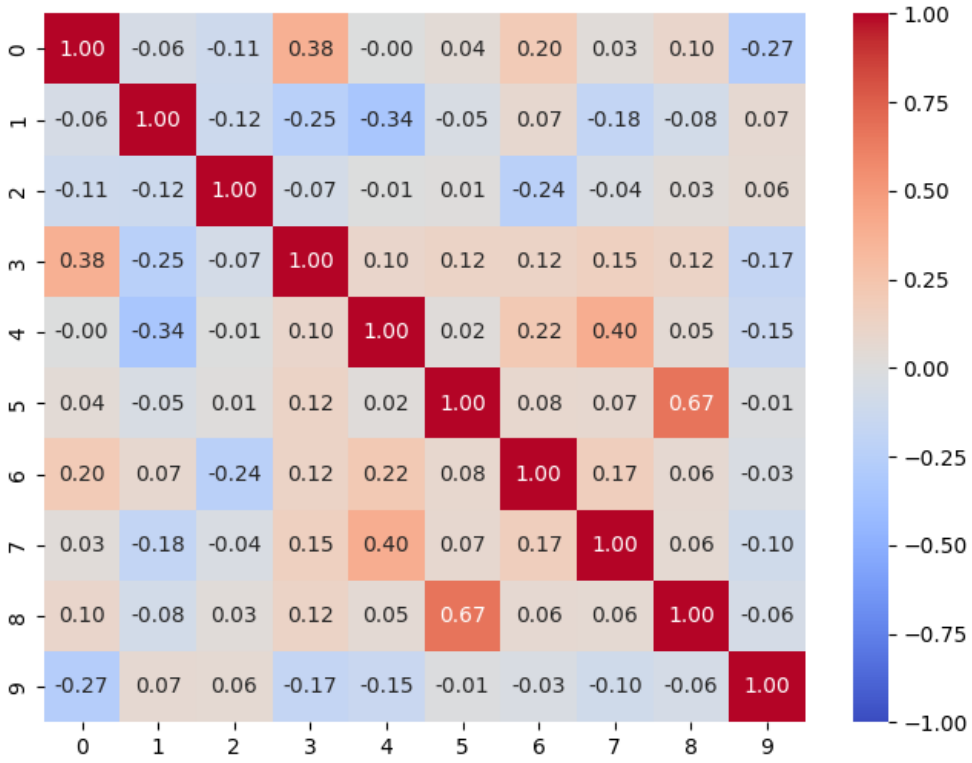


Figure 4.
Feature Correlation Heatmap (first 10 features).

To check the effectiveness of different feature selection methods, dimensionality reduction and feature ranking methods were evaluated using an SVM classifier. The performance of various feature selection methods is summarized in Table 1. It has been observed that SelectKBest based on ANOVA F-statistics achieved the highest classification accuracy of 88.51%, followed by PCA, which achieved an accuracy of 85.14%. PCA reduces the features to 43 principal components. Recursive Feature Elimination (RFE), Random Forest feature importance, and Mutual Information methods gave accuracies from 81.76% to 82.43%. ICA yielded an accuracy of 80.41%, whereas LDA achieved the lowest accuracy of 71.62%. In the proposed framework, PCA, although having lower accuracy than SelectKBest, is used because of its added benefits, such as dimensionality reduction, feature correlation mitigation, and noise insensitivity.

Table 2.
Performance comparison of different feature selection methods.

Method	Features	Accuracy	Precision	Recall	F1-Score
SelectKBest	43	88.51	0.8905	0.8851	0.8834
PCA	43	85.14	0.8661	0.8514	0.8529
Mutual Information	43	82.43	0.8382	0.8243	0.8222
RFE	43	82.43	0.8464	0.8243	0.8208
RF Importance	43	81.76	0.8337	0.8176	0.8151
ICA	43	80.41	0.8281	0.8041	0.8029
LDA	12	71.62	0.7135	0.7162	0.7057

Table 2 compares the performance of various feature selection methods in classification using SVM. The performance metrics used are accuracy, precision, recall, and F1 score. It is observed that the SelectKBest feature selection method gave the best performance in terms of accuracy of 88.51%,

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 9: 54-64, 2026
DOI: 10.55214/2576-8484.v10i9.13549

precision of 0.8905, recall of 0.8851, and F1 score of 88.34. PCA is the second-best feature selection method, with an accuracy of 85.14%, precision of 0.8661, recall of 0.8514, and F1 score of 0.8529.

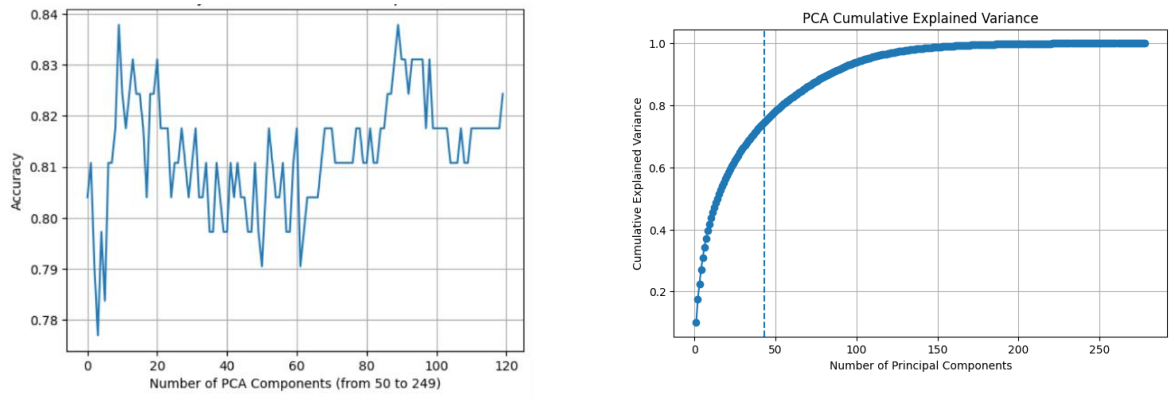


Figure 5.
(a) Impact of PCA Components on Classification Accuracy; (b) PCA Cumulative Explained Variance Analysis.

Figure 5 (a) shows the variation in classification accuracy as the number of PCA components increases. The accuracy varies in the range of 0.78–0.84. It shows that different numbers of principal components have a nonlinear relationship with the performance of the model. The highest accuracy is achieved when the number of PCA components is 90–95, which is an optimum trade-off between dimensionality reduction and information retention. Figure 5 (b) illustrates the cumulative explained variance by principal components. The cumulative variance increases as the number of components increases and reaches 75% when 40 principal components are used. After this, the variance reduces, and this indicates that the information is retained using a smaller number of components.

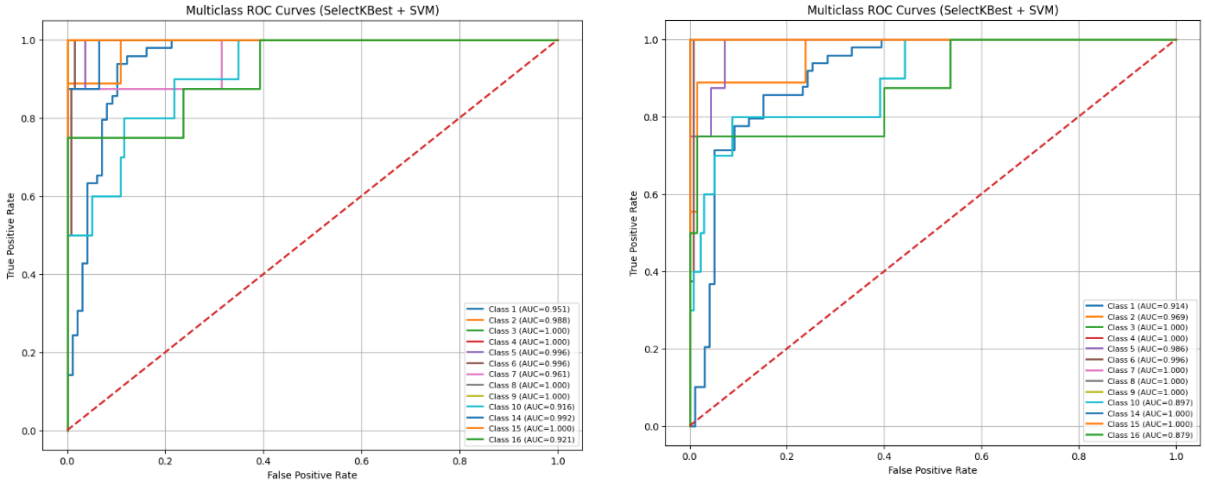


Figure 6.
ROC Curves for (a) SelectKBest + SVM and (b) PCA + SVM.

Figure 6(a) depicts ROC curves obtained by using SelectKBest as the feature selection method and SVM as a classifier. For each of the 13 classes, ROC curves represent the relationship between the TPR and FPR. The dotted diagonal line represents the baseline. It is observed that most of the ROC curves are packed near the upper-left corner of the plot, which denotes that the classification performance is good. Classes 3, 4, 8, and 15 achieved an AUC of 1.0, which shows that the method has good

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 9: 54-64, 2026
DOI: 10.55214/2576-8484.v10i9.13549

discrimination capability. The remaining classes show AUC values in the range of 0.916 to 0.996, which represents good predictive accuracy. Class 10 has an AUC value of 0.916, which is the lowest among all the other classes, but it still indicates good classification accuracy. The trend of high AUC values for all the classes shows that the features selected using feature selection methods can distinguish the class labels because of their highly informative nature. The ROC analysis helps decide that the SelectKBest feature selection method, along with SVM, is a robust multiclass classification framework. Figure 6(b) shows the performance of PCA as the feature selection method and SVM as the classifier. In this case, Classes 10 and 16 have AUC values of 0.897 and 0.879, respectively. This is due to overlap between these two class distributions.

6. Conclusion and Future Prospective

This paper presents a multi-module framework for addressing challenges faced in the classification of datasets having large numbers of missing values, outliers, non-informative and correlated features. The issue of missing values is handled using a correlation analysis method. The outliers are removed using the isolation forest method. The issue of class imbalance is handled using a data augmentation technique based on a median-based approach. The refined dataset is classified using an SVM algorithm. The results obtained are evaluated using various performance metrics, such as accuracy, precision, F1-score, and ROC-AUC, which confirm the effectiveness of the proposed framework. The proposed work can be extended by using deep learning models that can learn complex features. Also, Explainable Artificial Intelligence (XAI) techniques can be used for interpretability of classification decisions. Future work may also explore a hybrid method combining deep learning methods and automated feature selection strategies to further improve classification performance.

Transparency:

The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.

Copyright:

References

[1] P. Mahajan and A. Kaul, "Optimized multi-stage sifting approach for ECG arrhythmia classification with shallow machine learning models," *International Journal of Information Technology*, vol. 16, pp. 53-68, 2024. https://doi.org/10.1007/s41870-023-01641-9

[2] M. A. Talukder, A. S. Talaat, N. J. Muna, A. Alazab, M. Kazi, and U. K. Das, "An explainable deep learning framework for trustworthy arrhythmia detection from ECG signals," *Scientific Reports*, vol. 15, p. 39496, 2025. https://doi.org/10.1038/s41598-025-22986-0

[3] V. Mohammadi and B. S. Bigham, "A trust-aware hybrid unsupervised framework for robust ECG anomaly detection in IoMT monitoring networks," *Research Square*, 2026. https://doi.org/10.21203/rs.3.rs-8917980/v1

[4] M. D. Majid *et al.*, "A hybrid learning framework for automated multiclass electrocardiogram classification with SimCardioNet," *Scientific Reports*, vol. 16, p. 7621, 2026. https://doi.org/10.1038/s41598-026-36932-1

[5] A. S. A. Muhammed, P. R. S. Akaash, L. Lavanya, A. Amaranayani, G. V. Kumar, and T. H. Lakshmi, "A cross-validated LBX ensemble framework for reliable and clinically interpretable heart disease prediction," in *2025 4th International Conference on Innovative Mechanisms for Industry Applications (ICIMIA) (pp. 1558-1566). IEEE*, 2025.

[6] J. K. Francis and M. J. Darr, "Interpretable AI for time-series: Multi-model heatmap fusion with global attention and NLP-generated explanations," *arXiv preprint arXiv:2507.00234*, 2025. https://doi.org/10.48550/arXiv.2507.00234

[7] I. D. Aabdalla and D. Vasumathi, "A hybrid reinforcement and deep learning framework for ECG and PCG based cardiovascular signal classification," *Journal of Computational Analysis & Applications*, vol. 34, no. 5, pp. 143-153, 2025.

[8] P. Pande *et al.*, "Attention-driven hierarchical federated learning for privacy-preserving edge AI in heterogeneous iot networks," *International Journal of Advanced Computer Science & Applications*, vol. 16, no. 5, p. 462, 2025. https://doi.org/10.14569/IJACSA.2025.0160545

[9] R. Anand, S. V. Lakshmi, D. Pandey, and B. K. Pandey, "An enhanced ResNet-50 deep learning model for arrhythmia detection using electrocardiogram biomedical indicators," *Evolving Systems*, vol. 15, pp. 83-97, 2024. https://doi.org/10.1007/s12530-023-09559-0

[10] S. S. Irin, T. A. Mohanaprakash, T. Sindhu, S. Nalini, B. Dhribitri, and K. Cinthuja, "Data transformer—AI driven multimodal fusion network for cardiovascular risk prediction using multimodal data," in *Proceedings of the 5th International Conference on Evolutionary Computing and Mobile Sustainable Networks (ICECMSN) (pp. 203–212). IEEE*, 2025.

[11] R. Pashikanti, C. Y. Patil, and A. Shinde, "Cardiac Arrhythmia multiclass classification using optimized FLS-based 3D-CNN," *Journal of Intelligent & Fuzzy Systems*, vol. 46, no. 1, pp. 1543-1566, 2024. https://doi.org/10.3233/JIFS-230359

[12] R. Anitha, P. Talari, A. Babisha, and B. Tapas Bapu, "IoT-driven heart disease prediction using triple pseudo-siamese network and pyramid attention with termite alate optimization," *Biomedical Materials & Devices*, vol. 4, pp. 4666–4688, 2026. https://doi.org/10.1007/s44174-025-00498-9

[13] S. Kumar, M. Saranya, G. Dheepak, A. Muniraj, N. Baalakumar, and K. Pradeep, "Contrastive learning for personalized cardiovascular risk prediction from multi-lead ECG signal," in *2025 International Conference on Machine Learning and Autonomous Systems (ICMLAS) (pp. 918-927). IEEE*, 2025.

[14] S. U. Haq *et al.*, "Reseek-arrhythmia: Empirical evaluation of ResNet architecture for detection of arrhythmia," *Diagnostics*, vol. 13, no. 18, p. 2867, 2023. https://doi.org/10.3390/diagnostics13182867

[15] Q. Xiao *et al.*, "Deep learning-based ECG arrhythmia classification: A systematic review," *Applied Sciences*, vol. 13, no. 8, p. 4964, 2023. https://doi.org/10.3390/app13084964

[16] F. T. Dezaki *et al.*, "Echo-rhythm net: Semi-supervised learning for automatic detection of atrial fibrillation in echocardiography," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) (pp. 110-113). IEEE*, 2021.

[17] R. Hu, J. Chen, and L. Zhou, "A transformer-based deep neural network for arrhythmia detection using continuous ECG signals," *Computers in Biology and Medicine*, vol. 144, p. 105325, 2022. https://doi.org/10.1016/j.compbiomed.2022.105325

[18] M. Hammad *et al.*, "Deep learning models for arrhythmia detection in IoT healthcare applications," *Computers and Electrical Engineering*, vol. 100, p. 108011, 2022. https://doi.org/10.1016/j.compeleceng.2022.108011

[19] D. Al-Shammary, M. N. Kadhim, A. M. Mahdi, A. Ibaida, and K. Ahmed, "Efficient ECG classification based on Chi-square distance for arrhythmia detection," *Journal of Electronic Science and Technology*, vol. 22, no. 2, p. 100249, 2024. https://doi.org/10.1016/j.jnlest.2024.100249

[20] S. Dhyani, A. Kumar, and S. Choudhury, "Analysis of ECG-based arrhythmia detection system using machine learning," *MethodsX*, vol. 10, p. 102195, 2023. https://doi.org/10.1016/j.mex.2023.102195

[21] F. Liu *et al.*, "An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection," *Journal of Medical Imaging and Health Informatics*, vol. 8, no. 7, pp. 1368-1373, 2018. https://doi.org/10.1166/jmihi.2018.2442

[22] A. L. Goldberger *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215-e220, 2000. https://doi.org/10.1161/01.CIR.101.23.e215

[23] R. He *et al.*, "Automatic cardiac arrhythmia classification using combination of deep residual network and bidirectional LSTM," *IEEE Access*, vol. 7, pp. 102119-102135, 2019. https://doi.org/10.1109/ACCESS.2019.2931500

[24] H. A. Güvenir, B. Acar, H. Muderrisoglu, and J. R. Quinlan, "Arrhythmia [Dataset]," *UCI Machine Learning Repository*, 1997. https://doi.org/10.24432/C5BS32