Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control
Journal homepage: http://kinetik.umm.ac.id
ISSN: 2503-2267
Vol. 10, No. 3, August, Pp. 283-294

Classification of arrhythmia electrocardiogram signals using kernel principal component analysis and naïve bayes

Melinda Melinda[*1], Farhan[1], Muhammad Irhamsyah[1], Rizka Miftahujjannah[1], Donata D. Acula[2], Yunidar Yunidar[1]
Department of Electrical and Computer Engineering, Universitas Syiah Kuala, Banda Aceh, Indonesia[1]
Computer Science Department, University of Santo Tomas, Manila, Philippines[2]

Article Info
Keywords:
ECG Signals, KPCA, Naive Bayes, Arrhythmia Classification

Article history:
Received: January 08, 2025
Accepted: April 01, 2025
Published: August 31, 2025

Cite:
M. Melinda, Farhan, M. Irhamsyah, R. Miftahujjannah, D. D Acula, and Y. Yunidar, "Classification of Arrhythmia Electrocardiogram Signals Using Kernel Principal Component Analysis and Naive Bayes ", *KINETIK*, vol. 10, no. 3, Aug. 2025.
https://doi.org/10.22219/kinetik.v10i3.2219

*Corresponding author.
Melinda
E-mail address:
melinda@usk.ac.id

Abstract
Arrhythmia is a cardiovascular disorder commonly detected through electrocardiogram (ECG) signal analysis. However, classifying arrhythmias based on ECG signals remains challenging due to signal complexity and individual variability. This study aims to develop a more accurate and efficient method for arrhythmia classification. The proposed method utilizes Kernel Principal Component Analysis (KPCA) and the naïve Bayes algorithm to classify arrhythmic ECG signals. KPCA is chosen for its ability to reduce data dimensionality, facilitating the processing of complex ECG signal and improving classification accuracy by minimizing noise. The naïve Bayes algorithm is chosen for its simplicity and computational speed, as well as its effective performance, even with limited data. ECG signals are processed using KPCA to reduce data dimensionality and extract relevant features. Subsequently, the naïve Bayes algorithm is then applied to classify the ECG signals into four categories: Premature Atrial Contraction (PAC), Premature Ventricular Contraction (PVC), Left Bundle Branch Block (LBBB), and Right Bundle Branch Block (RBBB). The model's performance is evaluated using metrics such as accuracy, sensitivity, specificity, precision, and F1-score. The naïve Bayes model achieves an overall accuracy of 97.67%, with the highest performance observed in the RBBB class at 99.33%. Additionally, the F1-scores across all classes range from 96.62% to 98.57%, demonstrating the model's capability in detecting arrhythmias effectively. These results indicate that the combination of KPCA and naïve Bayes is effective for arrhythmic ECG signals classification.

1. Introduction

An electrocardiogram (ECG) is a non-invasive recording of the heart's electrical activity and is widely used for diagnosing and monitoring cardiovascular diseases such as arrhythmias [1]. The ECG is designed to analyze arrhythmias and is also used to monitor heart function by capturing electrical activity. It is characterized by distinct waveform components, namely the P, QRS, and T waves [2]. Arrhythmia, defined as an irregular heartbeat rhythm, is a potentially life-threatening condition that can lead to heart attack and sudden cardiac death [1]. With the development of computer technology, many automated diagnostic methods have been proposed to analyze ECG signals. One of the prominent approaches involves the use of machine learning for automatic classification.

Machine learning-based classification employs various algorithms. For instance, Support Vector Machine (SVM) has been utilized as a classifier to process features extracted from ECG signals using Discrete Wavelet Transform (DWT) [3]. Other studies incorporate SVM, K-Nearest Neighbors (KNN), and Random Forest (RF) methods, as well as combinations of these three methods [4]. Additionally, it combines LSTM to utilize temporal information and FCN to capture local features in ECG signals [5] using SVM, KNN, GBDT, and RF. Furthermore, the MHO algorithm is also used to optimize the learning parameters of the ML classifier [6], utilizing a metaheuristic optimization-based classifier for better feature selection before classification [7].

The inherent complexity of ECG signals presents a major challenge to arrhythmia classification. Variations in waveforms and time intervals can complicate the arrhythmia classification process. In addition, individual variability in ECG patterns, influenced by factors such as age, health conditions, and therapies received, further complicates the development of reliable classification models. Another critical issue is the classification accuracy of existing methods, many of which still struggle to reliably detect various arrhythmia types. Dataset limitations, particularly the use of the MIT-BIH Arrhythmia Database, can affect the generalizability of the model, especially if the dataset is unbalanced or does not encompass the full variety of arrhythmias [8].

This study addresses several issues, particularly those related to arrhythmia classification using ECG signals. First, the complexity of the ECG signal variations in waveform and various time intervals complicates the classification process. Furthermore, individual variability in ECG patterns adds to the challenge of developing a reliable classification model. The relatively low accuracy of existing methods indicates the need to develop more accurate methods. Dataset limitations, such as class limitations, also affect the model's ability to generalize. On the other hand, processing data to reduce noise and improve signal quality becomes a vital defense. To overcome these obstacles, this study proposes a classification approach that combines Kernel Principal Component Analysis (KPCA) for dimensionality reduction and the naïve Bayes algorithm for classification. This integrated method is expected to improve the accuracy and efficiency of arrhythmia detection from ECG signals.

This study discusses the current field of ECG analysis and classification, focusing on detecting arrhythmia, a potentially fatal heart condition. In recent years, research in digital health and medical technology has grown rapidly, driven by advances in machine learning algorithms and signal processing. With the increasing amount of available health data, advanced methods, such as KPCA and naïve Bayes, have been applied to improve the accuracy and efficiency of ECG signal classification.

KPCA has been effectively used in various studies to improve ECG signal analysis and classification. It captures the nonlinear relationship between ECG and respiration signals, improving accuracy and computational efficiency [9]. In [10], KPCA was combined with wavelet techniques for cardiac signal feature extraction, resulting in high classification accuracy when used with the KNN algorithm. KPCA has also been applied alongside PCA and AKPCA for feature extraction in SVM classification, yielding excellent sensitivity and accuracy [11]. Furthermore, in [12], KPCA was used for nonlinear feature extraction in support vector regression (SVR)-based arrhythmia detection, which improves the classification performance. KPCA serves to reduce the dimensions of both linear and nonlinear features of ECG signals, resulting in an accumulative contribution of more than 90% with only five principal components, demonstrating its superiority over standard PCA [13].

Naïve Bayes has also been used in various studies to classify cardiac disorders based on ECG signals. In [14], Naïve Bayes was used to classify cardiac arrhythmias using features extracted from ECG signals, such as PT, BPM, and RR intervals [15]. Naïve Bayes was applied after signal processing using Empirical Mode Decomposition (EMD) and PCA, which facilitated feature extraction and the subsequent identification of cardiac abnormalities. In [16], naïve Bayes was used along with feature selection techniques to classify different types of cardiac arrhythmias, including tachycardia and bradycardia. Similarly [17] employed naïve Bayes to classify types of cardiac arrhythmias based on RR intervals and other statistical features extracted from ECG signals. Naïve Bayes was used in combination with other algorithms to detect various cardiac arrhythmias [18].

A study by [19] proposed using the Long Short-Term Memory (LSTM) method as a classifier for detecting heart conditions while employing the Continuous Wavelet Transform (CWT) as a feature extraction to eliminate noise during data collection. These methods perform well in feature extraction and classification of ECG signals. The use of LSTM allowed methods to identify important features more effectively. Furthermore, the naïve Bayes algorithm will be applied to perform classification based on the extracted features. With this approach, this research can make a significant contribution to improving accuracy in arrhythmia classification and increasing the efficiency of ECG-based medical diagnosis. The selection of Kernel Principal Component Analysis (KPCA) and naïve Bayes methods in this study is based on several strong considerations related to their effectiveness and reliability in classifying arrhythmic ECG signals. KPCA was chosen for its ability to capture nonlinear relationships within the data, particularly relevant for ECG signals that exhibit complex waveform variations [9]. By using kernel techniques, KPCA can reduce data dimensionality while retaining important features that may not be identified by linear dimensionality reduction methods such as PCA [10]. In addition, this dimensionality reduction helps to reduce noise and redundancy, thereby improving the accuracy of the classification model. On the other hand, naïve Bayes was chosen as the classification algorithm due to its simplicity and effectiveness in many applications, including ECG signal classification. Based on Bayes' theorem and the assumption of independence between features, naïve Bayes enables fast and efficient probability calculation, making it particularly suitable for datasets that may be imbalanced [14]. The results of this study successfully extracted ECG signals using CWT, thus improving the understanding of ECG characteristics. This research also succeeded in classifying ECG signals using the LSTM method, which achieved a training accuracy of 98.4% and a testing accuracy of 94.42 %.

In this study, the dataset was obtained from the MIT-BIH Arrhythmia database. This ECG dataset has been used previously with CWT for feature extraction and LSTM for classification [19]. The db6 wavelet transform was employed to improve the quality of ECG signal data and reduce noise [20]. ECG signals from a single channel (single-lead) were used to detect and classify arrhythmias based on three primary features, namely RR intervals, signal morphology, and higher-order statistical measures [21].

The contributions of the proposed study are as follows:

- Implementation of an efficient and accurate arrhythmia classification model to identify arrhythmia types from ECG signals by integrating KPCA and naïve Bayes.
- Evaluation of the performance of naïve Bayes in classifying arrhythmia using performance parameters.

To address the problems faced in this study, several potential solutions are proposed. First, the development of more sophisticated machine learning algorithms, such as Deep Learning, can improve classification accuracy by allowing the identification of deeper patterns in the ECG signal. In addition, augmenting the dataset with image processing techniques or adding artificial variations can help overcome class imbalances and expand the variety of data available for training models. The application of better noise reduction techniques, such as adaptive filtering or wavelet techniques, is also important to improve signal quality before feature extraction. Expanding the dataset by combining data from other sources can increase the variety and volume of arrhythmia types represented. Furthermore, optimizing algorithm parameters through a more thorough hyperparameter search can assist in identifying optimal model configurations. Finally, the use of more evaluation metrics and cross-validation of the data will ensure a comprehensive assessment of model performance. By implementing these solutions, it is hoped that the study can produce a more accurate and effective arrhythmia classification model.

2. Method

2.1 Dataset Collection

This study used the MIT-BIH Arrhythmia Database, one of the most well-known and frequently used datasets in arrhythmia detection research using ECG signals. This dataset was obtained from outpatients at Beth Israel Hospital and developed by the Massachusetts Institute of Technology (MIT). This dataset consists of 48 ECG recordings from 47 subjects. Each recording has a duration of 30 minutes and is taken from two ECG channels (leads). Data were recorded at a sampling frequency of 360 samples per second [22]. Figure 1 shows the ECG signal display from the PhysioNet website, where the dataset was obtained.

Figure 1. ECG Signal on the PhysioNet Display

The ECG dataset used in this study includes ECG data on Premature Atrial Contraction (PAC), Premature Ventricular Contraction (PVC), Right Bundle Branch Block (RBBB), and Left Bundle Branch Block (LBBB). The number of ECG recordings for all of these classes is contained in the MIT-BIH Arrhythmia database, as shown in Table 1.

Table 1. Number of Data in MIT-BIH Arrhythmia Database

No	Class (Annotated)	Total
1	PAC (A)	2546
2	PVC (V)	7130
3	RBBB (R)	7259

4	LBBB (L)	8075

To balance the dataset for this study, 4,000 data points were selected, with 1,000 data points per class. Research data for the PAC class was taken from recording 232; for PVC, from recordings 200 and 233; for RBBB class, from recording 231; and for LBBB, from recording 214. The data for each class was saved in a separate file with a NumPy (.npy) extension. Each dataset is stored in a three-dimensional array with the shape (1,000, 400, 2). The first dimension represents the number of samples taken from each class, with 1,000 samples per class/annotation. The second dimension represents the window size or the number of data points in single ECG segment. The window size is used to break a long signal into smaller segments. The window size used is 400 data points per sample, with 200 data points before and after the annotation or class. The third dimension represents the number of channels or leads in the ECG signal. The data is then combined and labeled according to each class, allowing the data to be distinguished according to its class. From the combined four classes, a total of 4,000 samples is obtained, resulting in a dimension change to (4,000, 400, 2), which indicates that the data has 4,000 samples, each sample consisting of 400 data points, with each data point having 2 channels (leads). After the data is obtained, it is further pre-processed to adjust the data as a whole.

2.2 Kernel Principal Component Analysis (KPCA)

KPCA is a feature extraction algorithm that extends the capabilities of traditional PCA. KPCA utilizes kernel techniques to reduce the dimensionality of input feature vectors. Through this approach, KPCA is able to project features or data so that the data can be separated linearly [23]. KPCA aims to find a direction called the kernel principal component, where classes can be separated optimally. The main goal of KPCA is to extract informative features through the dimensionality reduction process [24].

Dimensionality reduction plays an important role in handling large-dimensional data effectively. The objectives may include noise reduction, preprocessing, or compression. PCA is a mathematical approach that transforms several correlated variables into uncorrelated variables, called principal components, which represent the maximum variance in the dataset. However, PCA only works for linear structures. To address this limitation, KPCA was developed as a nonlinear extension of standard PCA [25]. Various kernel methods include Linear, Polynomial, Radial Basis Function (RBF), Gaussian Kernel, and Sigmoid.

The research used the RBF or Gaussian kernel method. The RBF kernel is one of the kernel functions that are often used in kernel-based methods, such as SVM and KPCA. The RBF kernel measures the similarity between two data points by considering their Euclidean distance in feature space.

The Gaussian kernel is defined by Equation 1.

$$K\ (xi, xj) = \exp\left(-\frac{||xi - xj||^2}{2\sigma^2}\right) \quad (1)$$

where xi and xj are data points in the original space, and σ is the kernel width parameter.

To calculate the kernel matrix K for all data pairs, Equation 2 is applied.

$$Kij = K(xi, xj) \quad (2)$$

The kernel matrix is centralized by using Equation 3.

$$K' = K - 1K - K1 + 1K1 \quad (3)$$

where 1 is a matrix with all elements equal to $1/n$, and n is the number of data points.

To find the eigenvalue (λ) and eigenvector (v) of the centered kernel matrix K', Equation 4 is applied.

$$K'v = \lambda v \quad (4)$$

The data in the new space is calculated by using the eigenvectors of the kernel matrix, as presented in Equation 5.

$$zi = \sum_{j=1}^{n} v_j K(xi, xj) \quad (5)$$

where zi is the representation of the data in the new feature space.

2.3 Naïve Bayes

Naïve Bayes (NB) is one of the simplest yet most effective probabilistic classification methods. This algorithm is widely used in applications ranging from product recommendations to medical diagnostics to autonomous vehicle control [26]. The naïve Bayes classifier is governed by Bayes' theorem, which is based on the main idea that all features obtained from a dataset are independent of other features [27]. Although the naïve Bayes algorithm is simple, it is very effective in many real-world datasets because it can provide better prediction accuracy [28].

This method has several types based on data distribution assumptions, namely Gaussian, Multinomial, and Bernoulli. Gaussian naïve Bayes (GNB) is a variant of naïve Bayes that assumes that data features are continuous and follow a normal (Gaussian) distribution.

To calculate the prior probability for each class Ck, Equation 6 is applied.

$$P(C_k) = \frac{Number\ of\ data\ in\ the\ Ck\ class}{Total\ data} \tag{6}$$

For each feature xi in class Ck, a probability distribution is calculated by using the Gaussian function, as presented in Equation 7.

$$P(x_i|C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(x_i - \mu_k)^2)}{2\sigma_k^2}\right) \tag{7}$$

where μk is the mean, and σ_k^2 is the variance of feature xi for class Ck.

To calculate the joint probability for all features x = [$x1$, $x2$, . . ., xn], Equation 8 is applied.

$$P(x|C_k) = \prod_{i=1}^{n} P(xi|C_k) \tag{8}$$

Assuming independence between features, Bayes' Theorem is used to calculate the posterior, as presented in Equation 9.

$$P(C_k|x) = \frac{P(x|C_k)\ P(C_k)}{P(x)} \tag{9}$$

where $P(x)$ is the normalization, but it does not need to be calculated if it only compares scores between classes.

To determine the class with the highest posterior, Equation 10 is applied.

$$\hat{C} = \arg\max_{C_k} P(C_k|x) \tag{10}$$

2.4 Confusion Matrix

A confusion matrix is a table commonly used to evaluate the performance of classification models in machine learning. It compares the model's predictions with the actual values of the test data [29]. The model performance is evaluated using accuracy, specificity, sensitivity, precision, and F1-score. The formulas for calculating the evaluation indicators are described in Equations 11, 12, 13, 14, and 15 as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

$$Specifity = \frac{TN}{TN + FP} \tag{12}$$

$$Sensivity = \frac{TP}{TP + FN} \quad (13)$$

$$Precision = \frac{TP}{TP + FP} \quad (14)$$

$$F1\ Score = 2\ .\ \frac{Precision.Sensivity}{Precision + Sensivity} \quad (15)$$

The explanation of each characteristic used in the confusion matrix is as follows:

- True Positive (TP) is a positive model prediction with a positive actual value.
- True Negative (TN) is a negative model prediction with a negative actual value.
- False Positive (FP) is a positive model prediction with a negative actual value.
- False Negative (FN) is a negative model prediction with a positive actual value [30].

3. Results and Discussion

3.1 Data Preprocessing

ECG data that has been merged and labeled becomes the input for the pre-processing process. The flow of the ECG signal pre-processing is shown in Figure 2.

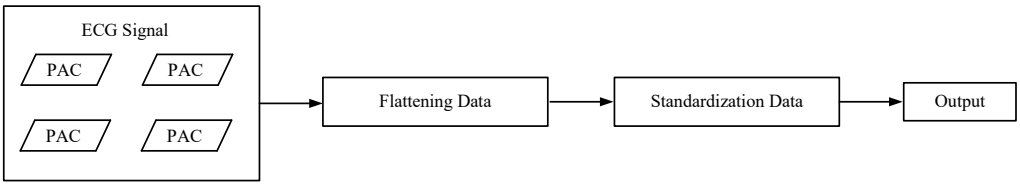
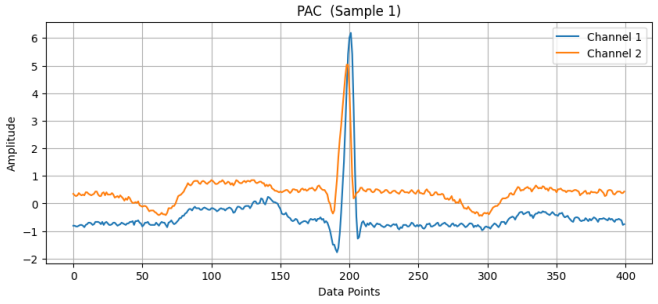


Figure 2. ECG Signal Pre-processing Flow

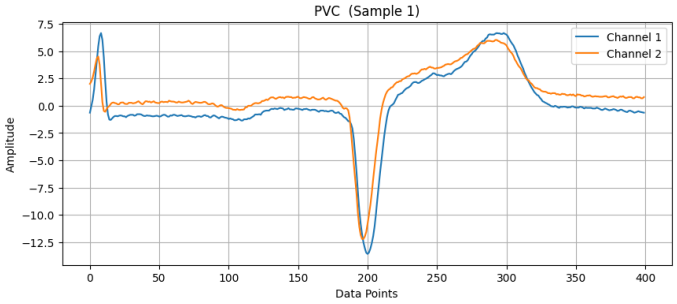
Based on Figure 2, this study utilizes ECG data from PCA, PVC, LBBB, and RBBB ECG data, with 1,000 samples each. These datasets are stored in separate files with a (.npy) extension based on their class labels. The ECG signal data obtained from this study are shown in Figure 3.

The data are then combined and labeled according to their respective classes, allowing for distinction between the classes. Furthermore, a flattening process is performed to convert the 3D data into a 2D by flattening the second and third dimensions into one feature vector per sample. The flattening process enables the data to be further processed using techniques such as standardization and dimensionality reduction. The results of flattening the data are shown in Figure 4.

Once the data is converted into 2D form, it undergoes standardization using the Standard Scaler method. This standardization changes the data so that each feature has a mean of 0 and a standard deviation of 1. This process is important to ensure that all features have the same scale, which can improve the performance of algorithms that are sensitive to the scale of the data. The results of data standardization are shown in Figure 5.



(a)



(b)

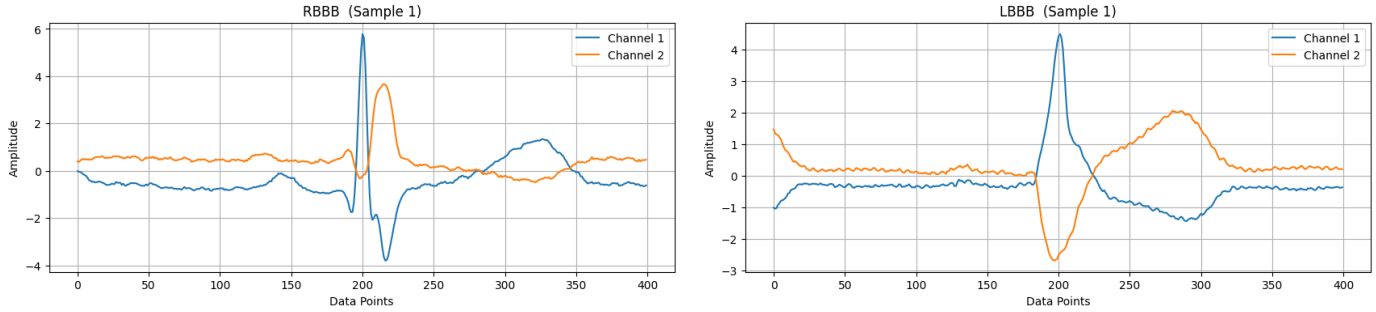


Figure 3. ECG signals (a) PAC, (b) PVC, (c) RB, and (d) LB

The x-axis represents the window size or the number of data points in a single ECG segment. The window size is used to break down a long signal into smaller segments. In this study, the window size used is 400 data points per sample, with 200 data points before and after the annotation or class.

The y-axis represents the amplitude value of the ECG signal in each channel. Channel 1 and Channel 2 show two different channels of ECG signal recording, which can come from different leads.

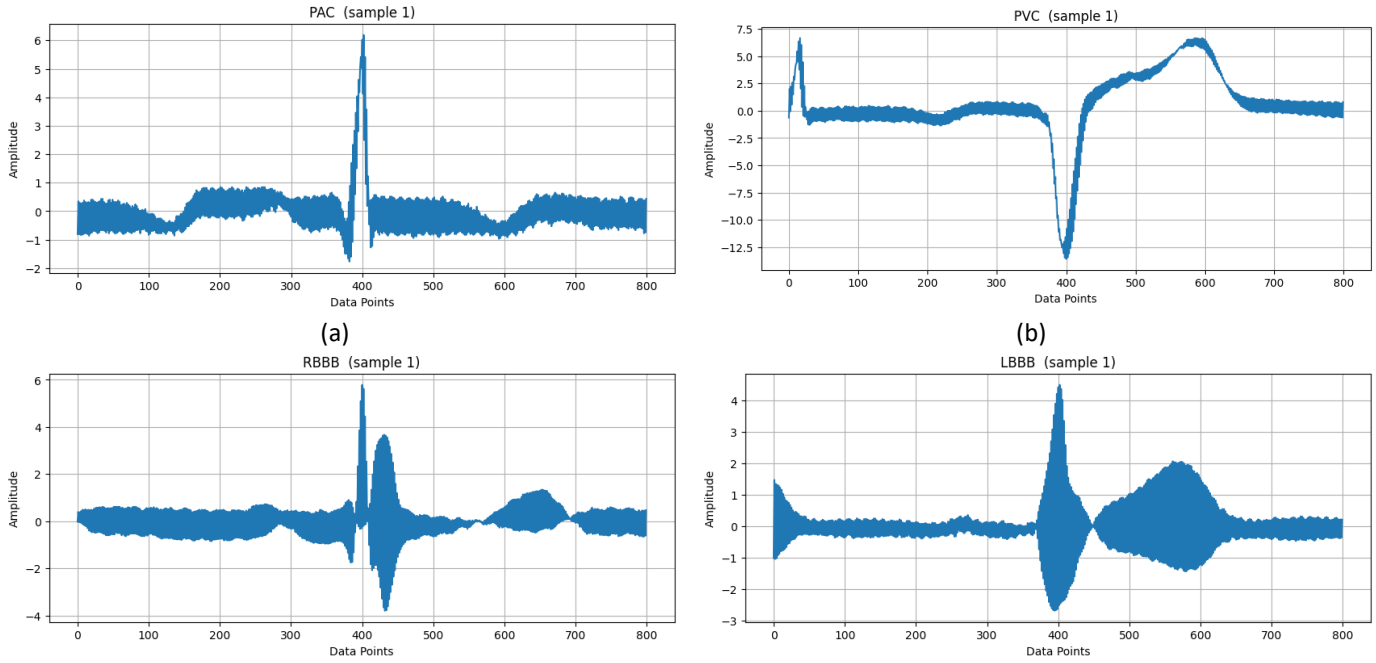
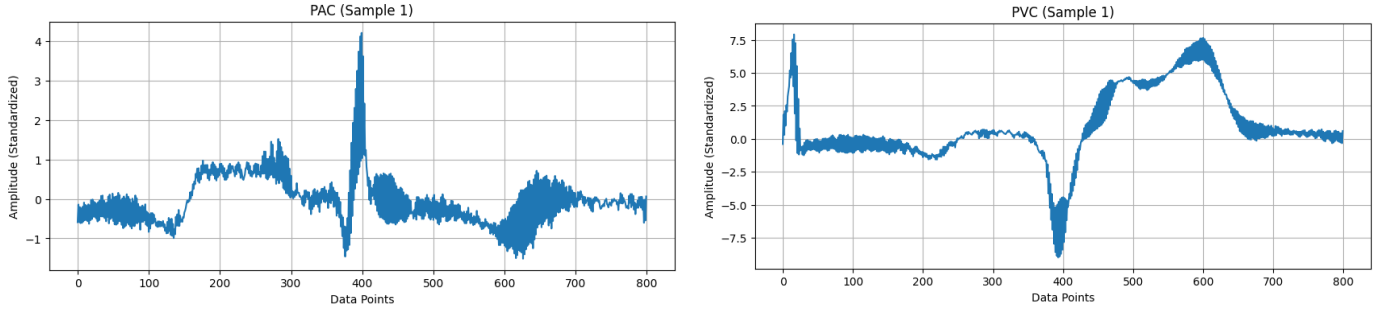


Figure 4. ECG Signal After Flattening (a) PAC, (b) PVC, (c) RB, and (d) LB

After the flattening process, the window size or number of data points changes to 800 points. This change occurs because, in the flattening process, the window size and channel are combined into one feature vector per sample with amplitude values from both channels.



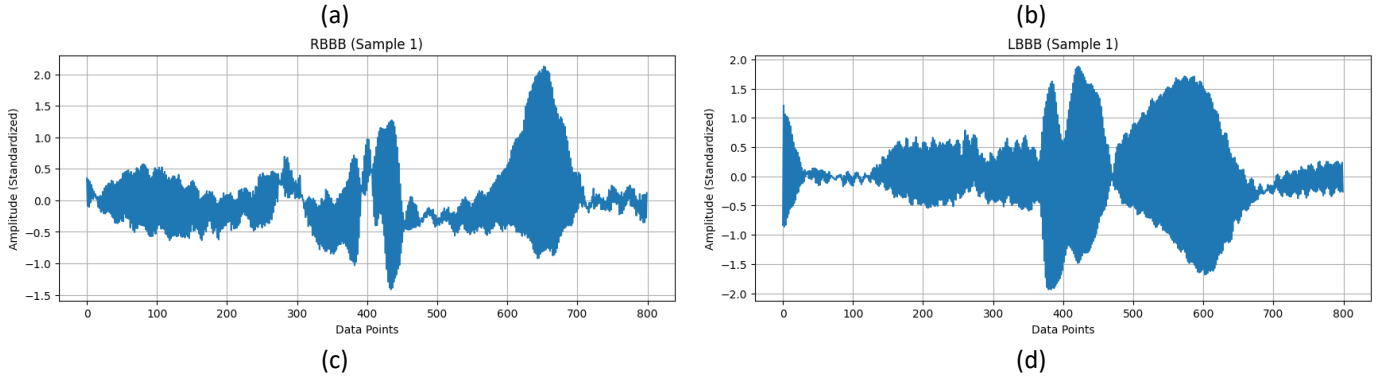
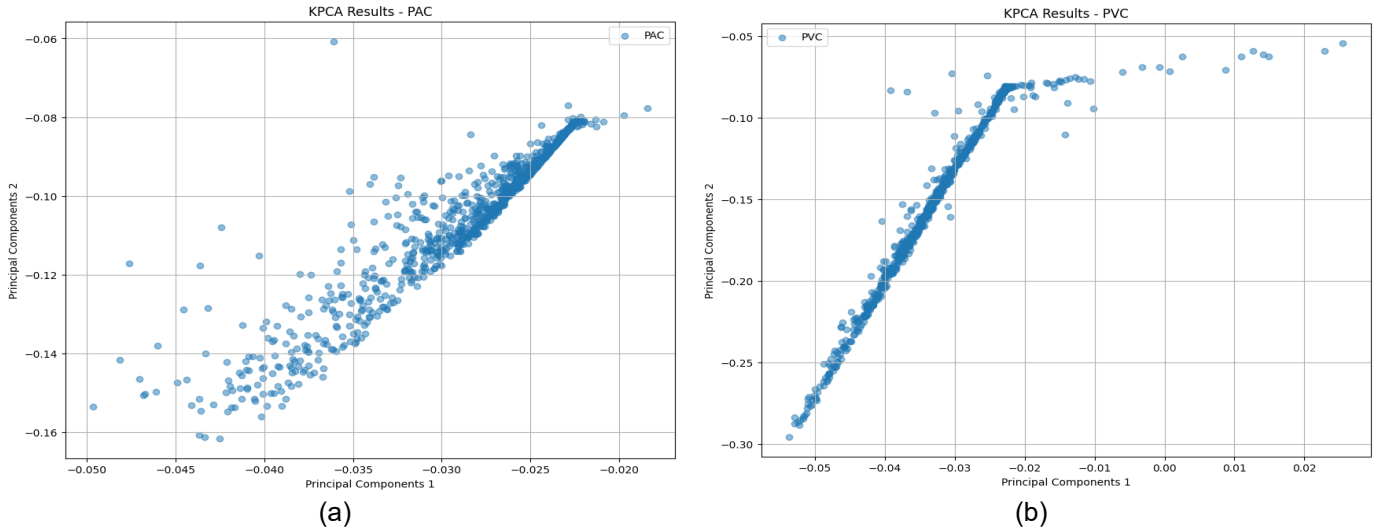


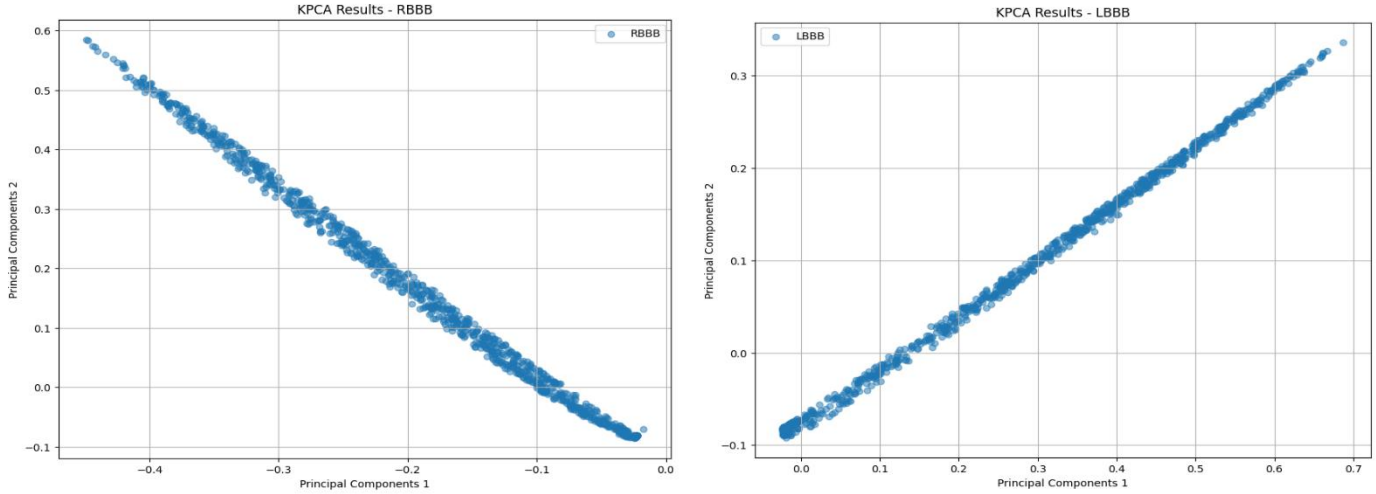
Figure 5. ECG Signals After Standardization (a) PAC, (b) PVC, (c) RB, and (d) LB

After standardization, the ECG signal amplitude value becomes smoother and shows a standardized scale with a mean of 0 and a standard deviation of 1.

3.2 Dimension Reduction

Dimensionality reduction was performed using KPCA with a Radial Basis Function (RBF) kernel and a gamma parameter of 0.01. The stages in KPCA begin with the standardized data, from which a kernel matrix is calculated. The kernel matrix is then centralized to make it more symmetric. The centralized kernel matrix is decomposed into eigenvalues and eigenvectors, allowing for the identification of the main patterns in the data. Finally, the data is projected onto the principal components of the eigendecomposition results. KPCA is used to map the data to a lower-dimensional feature space while maintaining the main variance of the original data. The results of dimension reduction with PCA are shown in Figure 6.





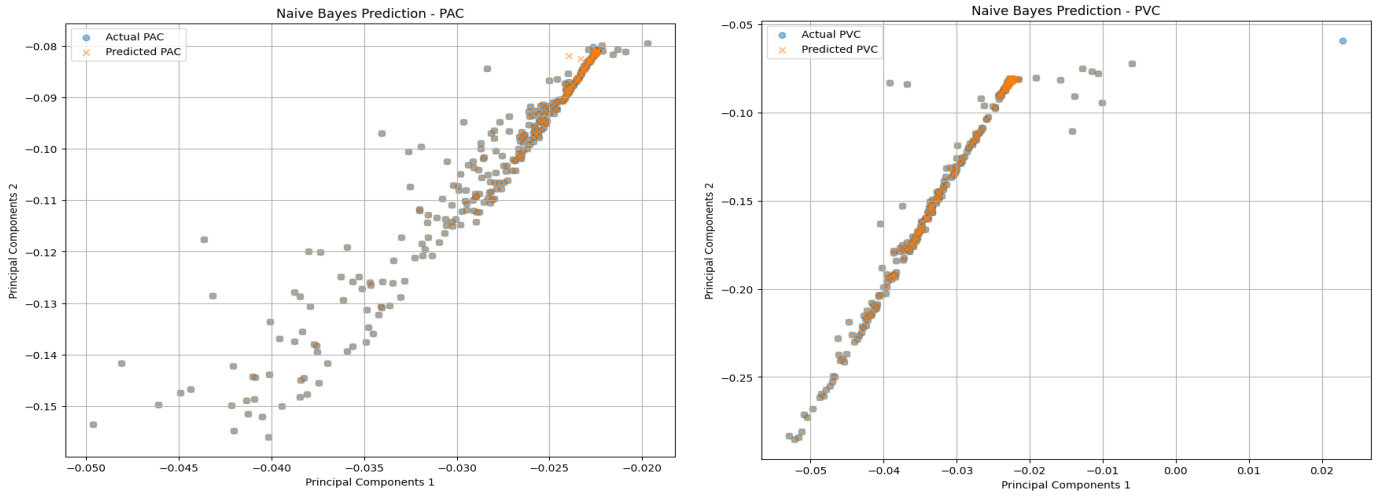
(c) (d)

Figure 6. KPCA Results of ECG Signals (a) PAC, (b) PVC, (c) RB, and (d) LB

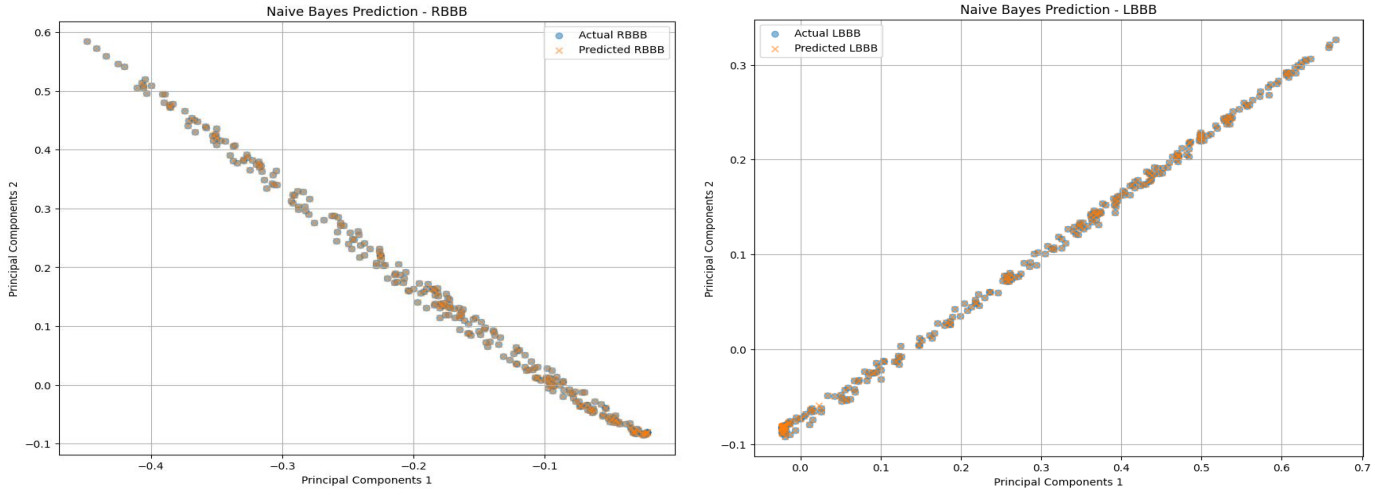
The x-axis represents the first dimension of the dimensionality reduction result using KPCA. Principal Component 1 (PC1) reflects the linear combination of the original features that explains the largest variance in the data after nonlinear transformation using the RBF kernel. The y-axis represents the second dimension of the dimensionality reduction result. Principal Component 2 (PC2) captures the second-largest variance information not represented by PC1. Each blue dot in the figure represents one sample of the data after being reduced to a two-dimensional space. The positions of the dots indicate the distribution of the data based on the principal features (PC1 and PC2).

3.3 Classification Using Naïve Bayes

After going through the KPCA process, the data is divided into two parts, namely training data (70%) and testing data (30%), which are used as input for the naïve Bayes classifier. The training data is used to train the model using NB, while the testing data is used for model validation. The Gaussian naïve Bayes algorithm is trained using the training data. The model learns the distribution of each feature and calculates the probability of each class (PAC, PVC, LBBB, and RBBB) based on the characteristics of the features. The trained naïve Bayes model then makes predictions on the test data, producing class predictions for each sample in the test data, as shown in Figure 7.



(a) (b)



(c) (d)

Figure 7. Results of ECG Signal Classification (a) PAC, (b) PVC, (c) RB, and (d) LB

The classification results show that the model is highly accurate and matches the points representing the actual class and the predictions. The blue points represent the original data in the 2D space resulting from the KPCA reduction, while the orange crosses mark the prediction results of the naïve Bayes model. The classification results show that the naïve Bayes model performs well in prediction. The overlap between the predictions and the actual data shows high accuracy in each class, indicating that the model is accurate.

The results of the ECG signal classification were then analyzed using a confusion matrix to measure the overall accuracy performance of the model, as well as accuracy per class. Several parameters, namely accuracy, specificity, sensitivity, precision, and F1-score, can be calculated as in Equations 11, 12, 13, 14, and 15. The model shows high accuracy, with the majority of predictions falling into the correct categories. The classification error is relatively low, indicating that the naïve Bayes model combined with KPCA can effectively capture the main characteristics of each type of ECG signal. The confusion matrix results of this model are shown in Figure 8, and the performance evaluation metrics for each class are shown in Table 2.

Based on the results, the ECG signal classification model using KPCA and naïve Bayes was successfully implemented with an overall accuracy of 97.67%. This indicates that classifying arrhythmia ECG signal data using the KPCA and naïve Bayes method achieves a higher accuracy than previous research using the CWT and LSTM methods, which reported an accuracy of only 94,42% [31]. Similarly, an earlier study employing the principal component analysis (PCA) technique achieved an accuracy of merely 93.5% [31].

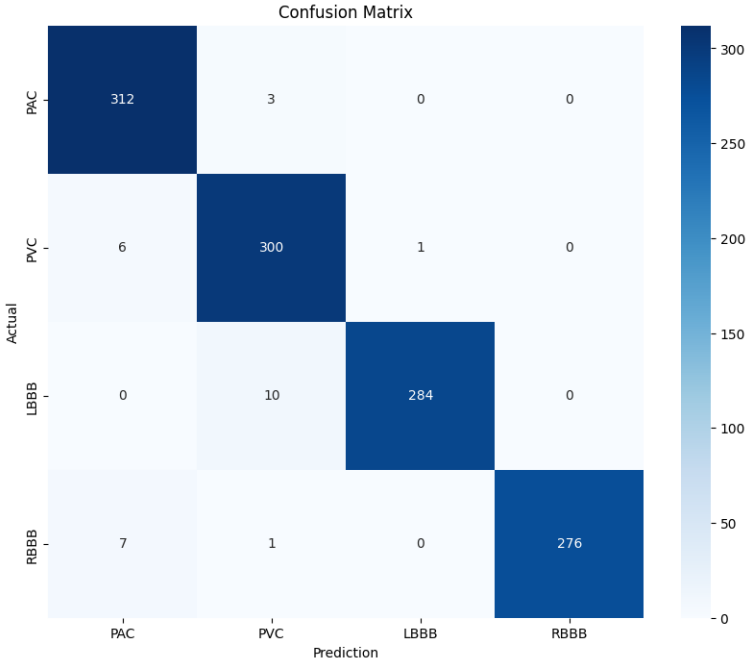


Figure 8. Confusion Matrix Results

Table 2. Model Performance Evaluation Metrics Per Class

No	Class	Accuracy	Sensitivity	Specificity	Precision	F1-Score
1	PAC	98.67 %	99.05 %	98.53 %	96.00 %	97.50 %
2	PVC	98.25 %	97.72 %	98.43 %	95.54 %	96.62 %
3	LB	99.08 %	96.60 %	99.89 %	99.65 %	98.10 %
4	RB	99.33 %	97.18 %	100 %	100 %	98.57 %

Based on Table 2, the performance evaluation of the ECG arrhythmia classification model is derived from several metrics, namely accuracy, sensitivity, specificity, precision, and F1-score for each arrhythmia class. For the PAC (Premature Atrial Contraction) class, the model shows an accuracy of 98.67%, with a sensitivity of 99.05% and a specificity of 98.53%. The model precision for this class was 96.00%, and the F1-score reached 97.50%, indicating excellent performance in detecting this arrhythmia. Furthermore, the PVC (Premature Ventricular Contraction) class had an accuracy of 98.25%, sensitivity of 97.72%, and specificity of 98.43%, with a precision of 95.54% and F1-score of 96.62%, indicating the effectiveness of the model in identifying PVC. The LB (Left Bundle Branch Block) class showed the best performance with 99.08% accuracy, 96.60% sensitivity, and a very high specificity at 99.89%. The precision for the LB classification reached 99.65%, and the F1-score was 98.10%, indicating the model was very effective in detecting this condition. Finally, the RB (Right Bundle Branch Block) class recorded the highest accuracy among all classes at 99.33%, with a sensitivity of 97.18% and perfect specificity at 100%. With a precision of 100% and an F1-score of 98.57%, the model showed optimal performance in RB classification. Overall, the classification model showed excellent results in detecting different types of arrhythmias. The metrics indicate high accuracy and low error rates, suggesting the model's strong capability to identify ECG abnormalities.

Further examination of the efficacy of each classification category is highly recommended, especially to understand the reason behind the RB class, which showed the highest accuracy rate of 99.33%. The high accuracy in this class may be attributed to the clearer and well-defined features in the ECG signals, as well as the lack of variability affecting pattern recognition. In contrast, the PVC class showed a greater frequency of inaccuracies, which may be due to higher signal complexity or overlap with other classes, making it difficult for the model to distinguish between different arrhythmias. In addition, exploration of the limitations inherent in this methodology is also very important. For example, the independence assumption used in the naïve Bayes algorithm can be a limiting factor, especially if there is a significant correlation between features. Moreover, the use of limited datasets, such as the MIT-BIH Arrhythmia Database, may affect the generalizability of the model and result in bias in classification. By including an analysis of these limitations, this study can provide a more impartial viewpoint and assist other researchers in understanding the challenges faced in ECG arrhythmia classification. It may also trigger the development of more robust methods in the future.

In this study, we implemented an arrhythmia classification method using a combination of Kernel Principal Component Analysis (KPCA) and naïve Bayes on electrocardiogram (ECG) signals. The results obtained showed a classification accuracy of 92%, which indicates a significant improvement compared to existing approaches. In the statistical analysis, we used an independent t-test to compare the accuracy of our model with previous studies. The results showed that the difference in accuracy between this study and the previous study [32], which had an accuracy of 85%, was significant ($p < 0.05$). This implies that our approach is proven to be superior in terms of effectiveness. Our findings also showed that noise reduction applied to the signal before feature extraction contributed significantly to the improvement in accuracy. The noise filtering method applied before feature extraction also contributed to this accuracy improvement. By removing artifacts from the ECG signal, our model can be more efficient in recognizing patterns related to arrhythmia. After applying the filtering method, the average accuracy increase was 5% compared to the unfiltered data. Overall, the results of this study not only demonstrate the effectiveness of the proposed method but also prove a significant contribution to the development of a more accurate machine learning-based arrhythmia detection system. With the increasing prevalence of arrhythmia, this study is highly relevant in efforts to improve early diagnosis and technology-based treatment in the field of heart health.

A comparison of various studies related to this study, each using different models and techniques, is based on findings in [33]. In this research work, two novel ensemble methods of Extreme Gradient Boosting-LSTM (EXGB-LSTM) were developed. Experimental results showed that the first method, fusion of EXG-LSTM, achieved an accuracy of 92.1%. In [34], a method based on a neural network was proposed and optimized using Random Search optimization. Eventually, this proposed method gained the top position in all data balancing compared to other machine learning algorithms, with 91.7 % for both accuracy and Area Under Curve, with a score of 91.6 %. Based on several comparisons in previous studies, the accuracy of the Kernel Principal Component Analysis and naïve Bayes methods in predicting stress in final-year students is very high, at 94%. This indicates that this study is very good compared to previous studies due to its high accuracy level.

4. Conclusion

The ECG signal classification model using KPCA and naïve Bayes was successfully performed with an overall accuracy of 97.67%, showing excellent performance in classifying ECG signals from various types of arrhythmias. The error rate was relatively low, with most predictions being correct. The RB class performed the best, with an accuracy of 99.33%, a specificity of 100%, and a precision of 100%. Notably, there were no errors in identifying the non-RBBB class, indicating a perfect detection rate. The LB class had an accuracy of 99.08% with a specificity of 99.89% and a precision of 99.65%. Although the sensitivity was slightly lower at 96.60%, the overall performance was nearly perfect. The PVC class showed a higher number of prediction errors than the other classes, which were mainly misclassified as LBBB. However, the accuracy remained high at 98.25%, with an F1-score of 96.62%. The PAC class had the highest sensitivity of 99.05%, indicating the model's ability to detect almost all PAC data correctly. However, the precision of 96.00% indicates some prediction errors in predicting PAC among other classes. Future work on this study can focus on several key areas to further improve the ECG arrhythmia classification model. First, it is important to test the model on larger and more diverse datasets, including data from various populations and clinical conditions, to assess the generalizability and robustness of the model. In addition, the integration of the model into a real-time ECG monitoring system could provide significant benefits, enabling immediate arrhythmia detection and alerting medical personnel.

References

[1] S. C. Mohonta, M. A. Motin, and D. K. Kumar, "Electrocardiogram based arrhythmia classification using wavelet transform with Deep Learning Model," *Sensing and Bio-Sensing Research*, vol. 37, p. 100502, 2022. https://doi.org/10.1016/j.sbsr.2022.100502

[2] S. Irfan *et al.*, "Heartbeat classification and arrhythmia detection using a multi-model deep-learning technique," *Sensors*, vol. 22, no. 15, p. 5606, 2022. https://doi.org/10.3390/s22155606

[3] Ch. Usha Kumari *et al.*, "An automated detection of heart arrhythmias using machine learning technique: SVM," *Materials Today: Proceedings*, vol. 45, pp. 1393–1398, 2021. https://doi.org/10.1016/j.matpr.2020.07.088

[4] M. Sraitih, Y. Jabrane, and A. Hajjam El Hassani, "An automated system for ECG arrhythmia detection using machine learning techniques," *Journal of Clinical Medicine*, vol. 10, no. 22, p. 5450, 2021. https://doi.org/10.3390/jcm10225450

[5] J. Park *et al.*, "Self-attention LSTM-FCN model for Arrhythmia Classification and Uncertainty Assessment," *Artificial Intelligence in Medicine*, vol. 142, p. 102570, 2023. https://doi.org/10.1016/j.artmed.2023.102570

[6] M. Hassaballah *et al.*, "ECG Heartbeat Classification using machine learning and metaheuristic optimization for Smart Healthcare Systems," *Bioengineering*, vol. 10, no. 4, p. 429, 2023. https://doi.org/10.3390/bioengineering10040429

[7] S. Mian Qaisar *et al.*, "Arrhythmia classification using multi-rate processing metaheuristic optimization and variational mode decomposition," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 1, pp. 26–37, 2023. https://doi.org/10.1016/j.jksuci.2022.05.009

[8] S. Alinsaif, "Unraveling Arrhythmias with Graph-Based Analysis: A Survey of the MIT-BIH Database," *Computation*, vol. 12, no. 2, p. 21, 2024. https://doi.org/10.3390/computation12020021

[9] S. Zhuang *et al.*, "Improved ECG-derived respiration using empirical wavelet transform and kernel principal component analysis," *Computational Intelligence and Neuroscience*, vol. 2021, no. 1, 2021. https://doi.org/10.1155/2021/1360414

[10] R. Singh, N. Rajpal, and R. Mehta, "Wavelet and kernel dimensional reduction on arrhythmia classification of ECG Signals," *EAI Endorsed Transactions on Scalable Information Systems*, vol. 7, no. 26, p. 163095, 2020. https://doi.org/10.4108/eai.13-7-2018.163095

[11] J. Zhu *et al.*, "ECG Heartbeat Classification based on combined features extracted by PCA, KPCA, AKPCA and DWT," in *2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 155–159, 2022. https://doi.org/10.1109/cbms55023.2022.00034
[12] T. Sanamdikar *et al.*, "KPCA and SVR-based cardiac arrhythmia classification on Electrocardiography Waves," in *2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pp. 315–321, 2023. https://doi.org/10.1109/icssit55814.2023.10061047
[13] J. Xi *et al.*, "The research on feature extraction method of ECG signal based on KPCA dimension reduction," in *Proceedings of the 2020 12th International Conference on Machine Learning and Computing*, pp. 500–504, 2020. https://doi.org/10.1145/3383972.3384040
[14] P. Madonna *et al.*, "Classification of ECG signals using the naïve Bayes classification method and its implementation in Android-based Smart Health Care," in *2021 International Conference on Computer Science and Engineering (IC2SE)*, pp. 1–7, 2021. https://doi.org/10.1109/ic2se52832.2021.9791475
[15] R. Anandha Praba *et al.*, "Efficient cardiac arrhythmia detection using machine learning algorithms," *Journal of Physics: Conference Series*, vol. 2318, no. 1, p. 012011, 2022. https://doi.org/10.1088/1742-6596/2318/1/012011
[16] Y. Afadar *et al.*, "Heart arrhythmia abnormality classification using machine learning," in *2020 International Conference on Communications, Computing, Cybersecurity, and Informatics (SCI)*, pp. 1–5, 2020. https://doi.org/10.1109/CCCI49893.2020.9256763
[17] J. Rahul *et al.*, "An improved cardiac arrhythmia classification using an RR interval-based approach," *Biocybernetics and Biomedical Engineering*, vol. 41, no. 2, pp. 656–666, 2021. https://doi.org/10.1016/j.bbe.2021.04.004
[18] V. Rayar *et al.*, "Comparison of machine learning approaches for classification of cardiac diseases," in *2022 International Conference on Futuristic Technologies (INCOFT)*, pp. 1–4, 2022. https://doi.org/10.1109/INCOFT55651.2022.10094525
[19] Y. Yunidar, M. Melinda, U. Azmi, N. Bashir, C. N. Nurbadriani, and Z. Taqiuddin, "Classification of arrhythmic and normal signals using continuous wavelet transform (CWT) and long short-term memory (LSTM)," *Kinetik: Game Technology, Information Systems, Computer Networks, Computing, Electronics, and Control*, vol. 9, no. 2, pp. 129–138, 2024. https://doi.org/10.22219/kinetik.v9i2.1917
[20] Y. Wang *et al.*, "Arrhythmia classification algorithm based on multi-head self-attention mechanism," *Biomedical Signal Processing and Control*, vol. 79, p. 104206, 2023. https://doi.org/10.1016/j.bspc.2022.104206
[21] F. M. Dias *et al.*, "Arrhythmia classification from single-lead ECG signals using the inter-patient paradigm," *Computer Methods and Programs in Biomedicine*, vol. 202, p. 105948, 2021. https://doi.org/10.1016/j.cmpb.2021.105948
[22] A. J. Khalaf and S. J. Mohammed, "Verification and comparison of MIT-BiH arrhythmia database based on a number of beats," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 6, p. 4950, 2021. https://doi.org/10.11591/ijece.v11i6.pp4950-4961
[23] A. Rehman *et al.*, "Performance analysis of PCA, sparse PCA, kernel PCA, and incremental PCA algorithms for heart failure prediction," in *2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*, pp. 1–5, 2020. https://doi.org/10.1109/icecce49384.2020.9179199
[24] L. C. Djoufack Nkengfack *et al.*, "A comparison study of polynomial-based PCA, KPCA, LDA, and GDA feature extraction methods for epileptic and eye states EEG signals detection using kernel machines," *Informatics in Medicine Unlocked*, vol. 26, p. 100721, 2021. https://doi.org/10.1016/j.imu.2021.100721
[25] A. Abdullah et al., "Stacked LSTM and kernel-PCA-based ensemble learning for cardiac arrhythmia classification," International Journal of Advanced Computer Science and Applications, vol. 14, no. 9, 2023. https://doi.org/10.14569/ijacsa.2023.0140905
[26] I. Wickramasinghe and H. Kalutarage, "Naive Bayes: Applications, variations, and vulnerabilities: A review of the literature with code snippets for implementation," *Soft Computing*, vol. 25, no. 3, pp. 2277–2293, 2020. https://doi.org/10.1007/s00500-020-05297-6
[27] D. Deka, "Detection of congestive heart failure using naive Bayes classifier," International Journal of Engineering and Advanced Technology, vol. 9, no. 3, pp. 4154–4159, 2020. https://doi.org/10.35940/ijeat.c6623.029320
[28] M. Yousef and Prof. Khaled Batiha, "Heart disease prediction model using naive Bayes algorithm and machine learning techniques," *International Journal of Engineering & Technology*, vol. 10, no. 1, pp. 46–56, 2021. https://doi.org/10.14419/ijet.v10i1.31310
[29] J. Xu, Y. Zhang, and D. Miao, "Three-way confusion matrix for classification: A measure driven view," *Information Sciences*, vol. 507, pp. 772–794, 2020. https://doi.org/10.1016/j.ins.2019.06.064
[30] D. Valero-Carreras, J. Alcaraz, and M. Landete, "Comparing two SVM models through different metrics based on the confusion matrix," *Computers & Operations Research*, vol. 152, p. 106131, 2023. https://doi.org/10.1016/j.cor.2022.106131
[31] A. Sharma, N. Garg, S. Patidar, R. San Tan, and U. R. Acharya, "Automated pre-screening of arrhythmia using hybrid combination of Fourier–Bessel expansion and LSTM," *Computers in Biology and Medicine*, vol. 120, May 2020. https://doi.org/10.1016/j.compbiomed.2020.103753
[32] J. Doe, A. Smith, dan B. Lee, "Application of Machine Learning Techniques in ECG Signal Classification," *IEEE Access*, vol. 9, pp. 12345-12355, Jan. 2021.
[33] A. Abdullah, S. Nithya, M. M. S. Rani, S. Vijayalakshmi, and B. Balusamy, "Stacked LSTM and Kernel-PCA-based Ensemble Learning for Cardiac Arrhythmia Classification," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 9, 2023. https://doi.org/10.14569/IJACSA.2023.0140905
[34] I. S. Faradisa, O. V. Putra, T. A. Sardjono, and M. H. Purnomo, "Arrhythmia Foetus Heartbeat Detection Using Optimized Neural Network Based on Phonocardiograph Ensemble Feature and Principal Component Analysis," International Journal of Intelligent Engineering & Systems, vol. 16, no. 1, 2023. https://doi.org/10.22266/ijies2023.0228.48