Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 44 No. 6 (2023)

An Enhancement Performance Study to Predict Premature Ventricular Contraction Using Machine Learning Techniques

Mona Shaaban Hassan[1], Sami Ali Mostafa[1], Hamed A. Ibrahim [1]

[1] Faculty of Technology and Education, Department of Electrical, Suez University, Suez, Egypt.

Corresponding author: Mona Shaaban Hassan, master's student, Faculty of Technology and Education, Department of Electrical, Suez University, Suez 43713, Egypt,

Abstract

Premature ventricular contractions (PVC) are common cardiac arrhythmias in which heartbeats are initiated from the ventricles instead of the sinoatrial (SA) node. Detecting abnormal PVC beats is essential for the early detection and prevention of life-threatening heart diseases. This paper focuses on the detection of PVC arrhythmia using machine learning (ML) with the MITBIH arrhythmia database. Since the electrocardiogram (ECG) signal is inherently non-stationary, noise reduction is necessary to denoise the raw data during the pre-processing stage. The process of PVC arrhythmia detection involves several stages, including ECG signal feature extraction and selection. Features are extracted and selected from the ECG signal in the morphological stage as input for three different classifiers and the final classification stage. The focus of this paper is on the classification stage, which employs support vector machines (SVM), random forest (RF), and quadratic discriminant analysis (QDA). After tuning different parameters using GridSearchCV, the performance metrics of various models for each classification method were compared. The performance measures considered for comparative analysis are specificity (Spe), precision (Pre), F1-score, sensitivity (Sen), and accuracy (Acc). The SVM classifier achieves Spe 90.3%, Pre 90.7%, Sen 97.3%, Acc 93.8%, and F1-score 94%. The RF classifier results are: Spe 99.1%, Pre 99.1%, Sen 100%, Acc 99.5%, and F1-score 99.5%. As for the QDA classifier results, they are as follows: Spe 89.7%, Pre 89.9%, Sen 91.3%, Acc 90.5%, and F1-score 90.6%.
It is found from the experimental results that the RF classifier achieves a higher accuracy rate.

Keywords: PVC, ECG signal, Classification.

1. Introduction and Preview

Cardiovascular diseases (CVDs) are responsible for most deaths worldwide, with over 17.9 million deaths in 2019 [1]. The heart plays a significant role in circulating blood throughout the body. The heart's electrical function is initiated by pacemaker cells, which create the ECG signal detected by electrodes placed on the body. Changes in voltage within the ECG signal are a result of the activation and deactivation of the heart's cells through action potentials [2], causing the cells to contract. The heart's rhythm is characterized by the P peak, QRS complex, and T peak, as shown in Figure 1.

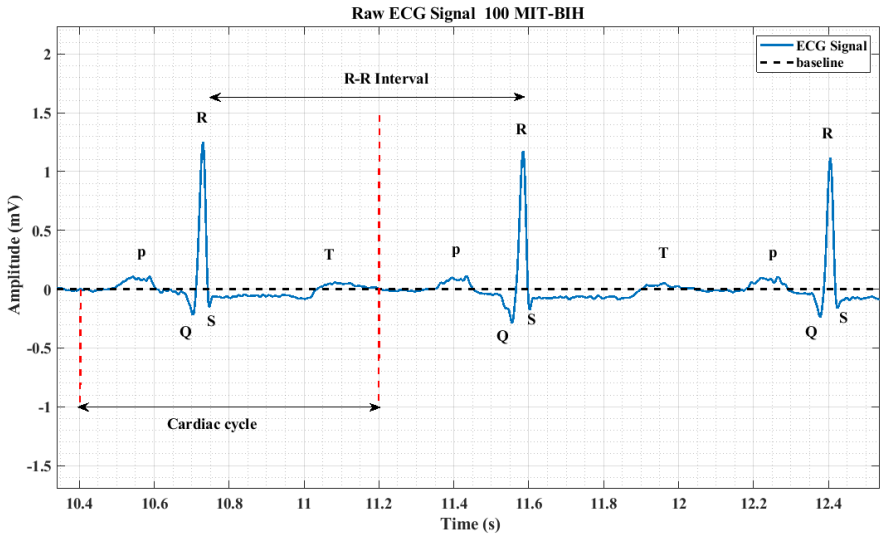

Figure 1. Normal cycle P, QRS, and T peak: raw 100 ECG signal from the MITHIB arrhythmia database

These waveforms are created by the depolarization and repolarization of the atria and ventricles. Abnormal heart rhythms, or arrhythmias, can lead to fluctuations in blood pressure and potentially fatal conditions like stroke or sudden death [3]. PVC is referred to as a type of abnormality that affects older people, occurring in around 17% of people and making up around 33% of all heart diseases worldwide. PVC has a different sequence than a normal beat, as when the ventricles contract before the atria due to the activation of Purkinje fibers instead of the pacemaker cells' SA node, it results in an abnormal beat known as ectopic rhythms. Frequent PVCs can lead to ventricular tachycardia or fibrillation [4]. The study has developed a traditional method for identifying PVC by automatically learning features from raw ECG signals. These techniques include algorithms like template matching [5], self-organizing maps [6], correlation coefficients [7], and discriminant analysis [8].

Recently, various ML techniques have been utilized to create predictive models for PVC. Sarshar and Mirzaei (2022) applied a deep learning model using a convolutional neural network to detect PVCs from ECG signals automatically [9]. They have a sensitivity of 99.2% and a specificity of 98.6% using a support vector machine (SVM) classifier. In 2021, Yu et al. used deep metric learning and K-nearest neighbors machine learning algorithms to detect PVCs and were able to detect PVCs with 96.9% sensitivity and a 92.1% positive predictive value (PPV) [10]. A one-dimensional convolutional neural network was proposed by Wang et al. (2020) with high accuracy of 99.6%, 96.97% sensitivity, and 99.8% specificity [11]. Xie et al. (2019) used a random forest model with classification and regression tree (CART) analysis to achieve an accuracy of 96.38% in classifying PVC from normal sinus rhythm from the MITBIH arrhythmia database [12]. Oliveira et al. (2019) introduced a novel set of features based on geometric shapes constructed from QRS complexes and evaluated several machine-learning algorithms for PVC classification [13]. Zhou et al. (2017) suggested a novel approach for PVC identification that employs a combination of inference rules based on signal rating systems and deep neural networks [14].

In the present work, it is found that the utilization of pre-processing and feature extraction steps prior to classification will enhance the system's performance. The imbalance problem is a common issue faced by most of the ECG databases used in the experimentation process. The paper is organized in the following manner: Section two outlines the research methodology used. Section three presents results analysis and discussion. At last, Section 4 summarizes the conclusions of the study.

2. Methodology

The methodology used is illustrated in Figure 2. Figure 2 describes the main steps, such as preprocessing, detection of the R wave in the ECG signal, feature extraction, feature selection, and classification.

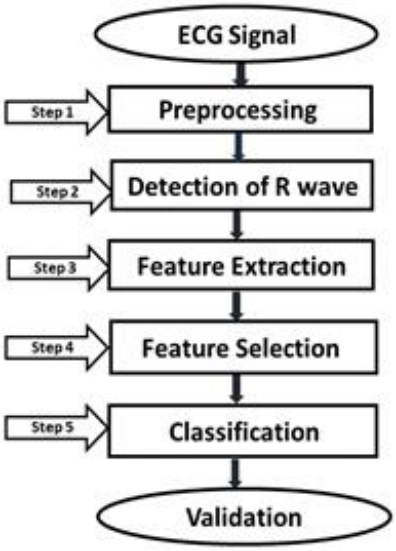

Figure 1. General flowchart diagram: five basic steps can be used to detect PVC arrhythmia.

2.1 Pre-processing

In the preprocessing stage, the non-stationary ECG signal consists of different types of artifacts, namely, baseline wander (BW), electromyography (EMG), and powerline interference (PLI). Noise affects the ECG signal, so it is necessary to eliminate this noise from the raw data.

The data used for analysis is available at the MITIBIH arrhythmia database [15], which includes normal and abnormal ECG signals from different people of all sexes and ages. A single signal MLII collected at a 360 HZ sample rate with a millivolt input range consists of the 48 ECG recordings in the database, each lasting 30 minutes. In this work, only 16 records are chosen to be used [100, 101, 103, 114, 115, 119, 122, 123, 201, 208, 213, 215, 219, 220, 223, and 234], which contain normal and PVC beats.

The remaining records contain kinds of arrhythmia and processing depending on limb lead II only, which are not considered in this work.

BW refers to the phenomenon where the baseline axis (X-axis) of individual heartbeats appears to deviate up or down instead of remaining straight. This effect is caused by low-frequency components, typically between 0 and 0.5 Hz, which are the result of respiration and body movement [16-17].Various techniques are utilized for the removal of BW, such as the median filter [18] and the fast Fourier transform [19].

This paper introduces a method to eliminate the baseline wander (BW) from the signal using discrete wavelet transform (DWT) with Daubechies wavelet (db4) at level 10. The sampling frequency is divided into the dyadic scale to identify low-frequency noise, resulting in a set of approximation coefficients (A10) and 10 sets of detail coefficients. By removing the approximation coefficients (A10), the filtered signal, devoid of BW, is obtained. This process is illustrated in Figure 3. The details of the DWT and the corresponding frequency range are provided in Table 1, which are used to recover the filtered signal.

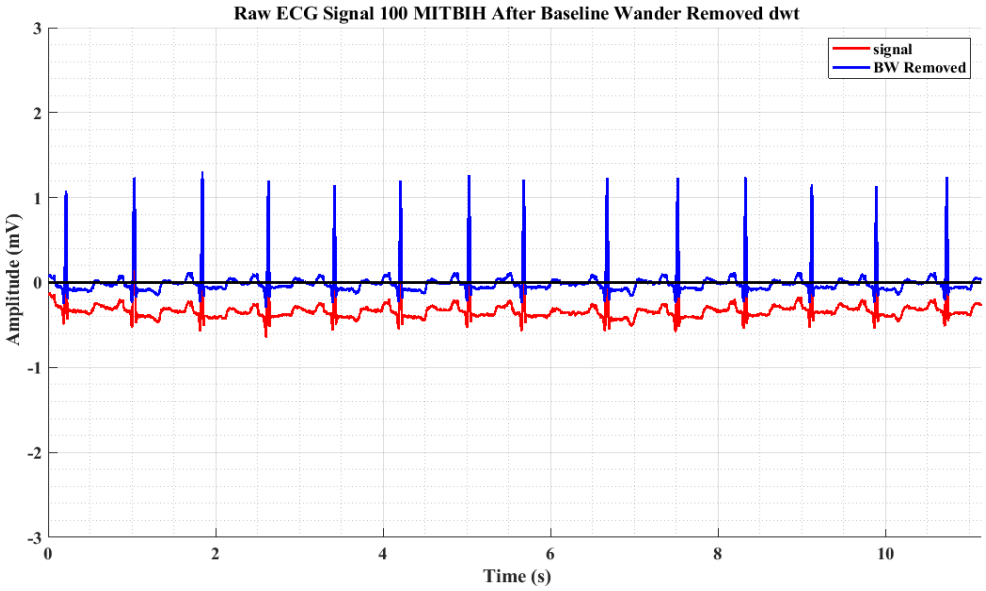

Figure 2. Removal of BW noise from MIT-BIH -100: ECG signals with and without BW noise.

Table 1 Detail coefficients DWT of level ten and the frequency range.

Frequency (Hz)	Detail Coefficients
360-180	D1
180-90	D2
90-45	D3
45 -22.5	D4
22.5- 11.25	D5
11.25- 5.62	D6
5.62-2.81	D7
2.81-1.40	D8
1.40-0.70	D9
0.70-0.35	**D10 (BW range)**

After removing the BW, the next step is to eliminate the high-frequency noise of the EMG and PLI. PLI is a type of interference that occurs at a high frequency, typically around 60 Hz, caused by the electromagnetic fields generated by power lines [16-17]. There are several techniques employed to eliminate PLI, such as wavelet packet transform (WPT) [20], notch filter, and short-time Fourier transform (STFT) [21].

To ensure the accuracy of an ECG signal for diagnostic purposes, it is crucial to eliminate PLI, which has a significant impact on the signal. This issue can be effectively addressed by employing a notch filter, which is essential for removing PLI from the ECG signal. The IIR notch filter only removes the unwanted noise without distorting the important components of the ECG signal. A notch filter with a bandwidth that corresponds to the power frequency can effectively eliminate the PLI from the ECG signal.

The bandwidth of the notch filter is typically chosen to be narrow enough to specifically target the power frequency while still allowing the rest of the ECG signal to pass through unaffected. The IIR notch filter is a band resection filter that we used to eliminate PLI, the high-frequency noise, see Figure 4.

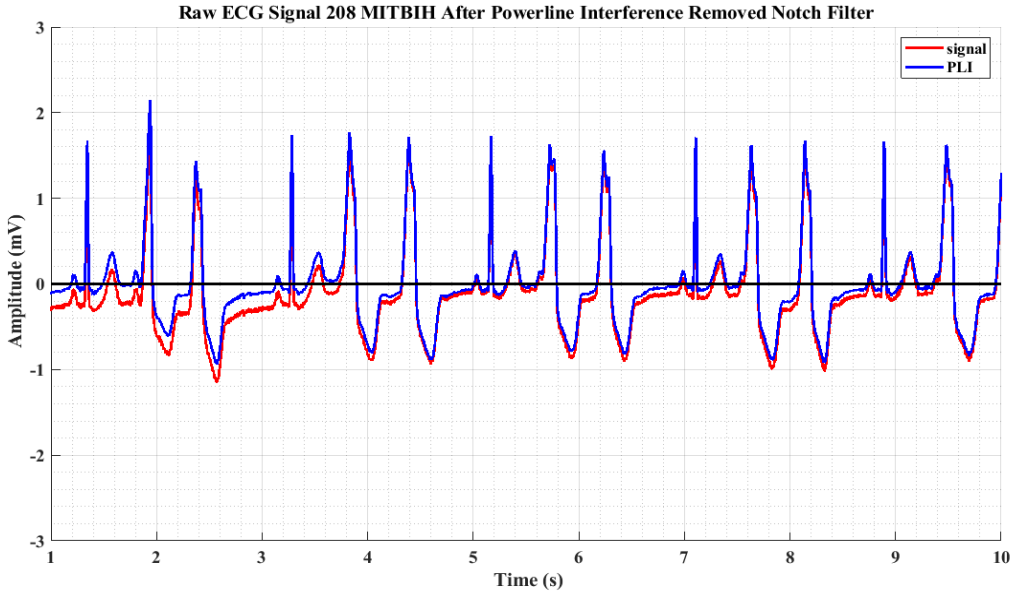

Figure 3. Removal of high-frequency PLI noise from MIT-BIH-208: displays the ECG signal before and after filtration was applied in the iirnotch filter to remove the PLI 60 Hz noise in a random period.

After removing the PLI from the ECG signal, the next process of elimination of the EMG, also known as muscle contractions, is typically around 100 Hz. The spectral content of muscle activity overlaps significantly with the ECG wave[16-17].
Using various methods to denoise the EMG noise from the ECG signal, such as the moving average filter [22] and the digital filter [23] will eliminate the noise using a low-pass (LP), Butterworth filter with an order of 20 and a cutoff frequency of 80 Hz, as depicted in Figure 5.

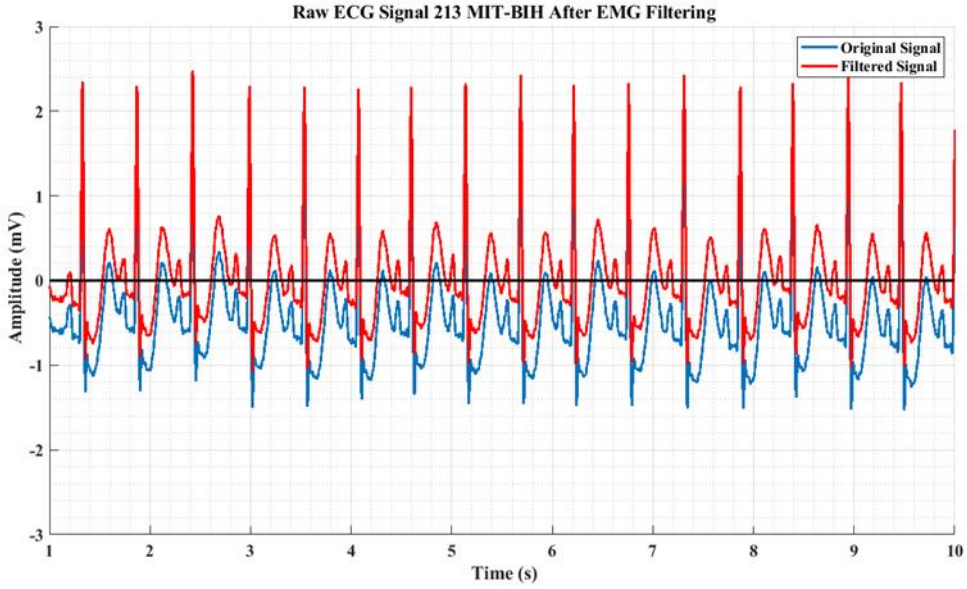

Figure 5. Removal of high-frequency EMG noise from MIT-BIH-213: displayed the elimination of EMG noise from the ECG signal by applying the Butterworth LP filter.

2.2 R peak detection

Next, in the preprocessing operation of ECG signals, the next step is R peak detection. The R-wave is a significant positive deflection that represents the depolarization of both ventricles. The R wave is typically the

largest and most easily identifiable peak in the ECG signal. Due to its high amplitude and dominant nature, the detection of R waves in ECG signals is often prioritized, particularly in lead II, where the QRS complex exhibits a higher frequency within a shorter period. Moreover, it serves as a reference point for measuring the duration and amplitude of other waves and intervals in the ECG signal, such as the PQS and T waves. Various methods have been explored for detecting the R wave in ECG signals, including the Pan-Tompkins algorithm [24], thresholding [25], differentiation methods [26], IIR Filter [27], deep learning [28], adaptive Kalman filter bank [29], and hybrid EMD-DWT [30].

In this case, R peaks in the ECG signal are detected by wavelet analysis using dwt (db4), and these wavelets were determined to be suitable due to the QRS complex.

DWT (db) level 4 is applied to filtered signals, and the output signals are decomposed to reduce the details and preserve the QRS complex. Once R peaks are detected, the other peaks are detected using minimum and maximum algorithms, and a window is used to search for signals for down-sampling purposes, as shown in Figure 6.

Finally, after R peak detection, the amplitudes of P, Q, S, and T are localized using a window of Rloc-50 to Rloc-50 to find the maximum for P peaks. a window of Rloc-40 to Rloc-10 to find the maximum for Q peaks. a window of Rloc+5 to Rloc+50 to find the maximum for S peaks and a window of Rloc+40 to Rloc +100 to find the maximum for T peaks, as displayed in Figure 7 normal beats.

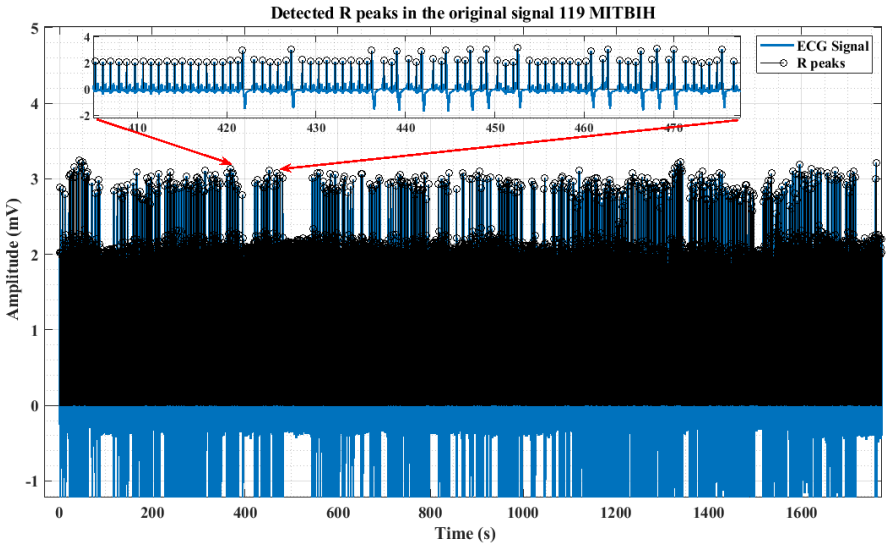

Figure 4. R Peak Detection from MIT-BIH-119: R Peak locations are shown in a better way by zooming in randomly.

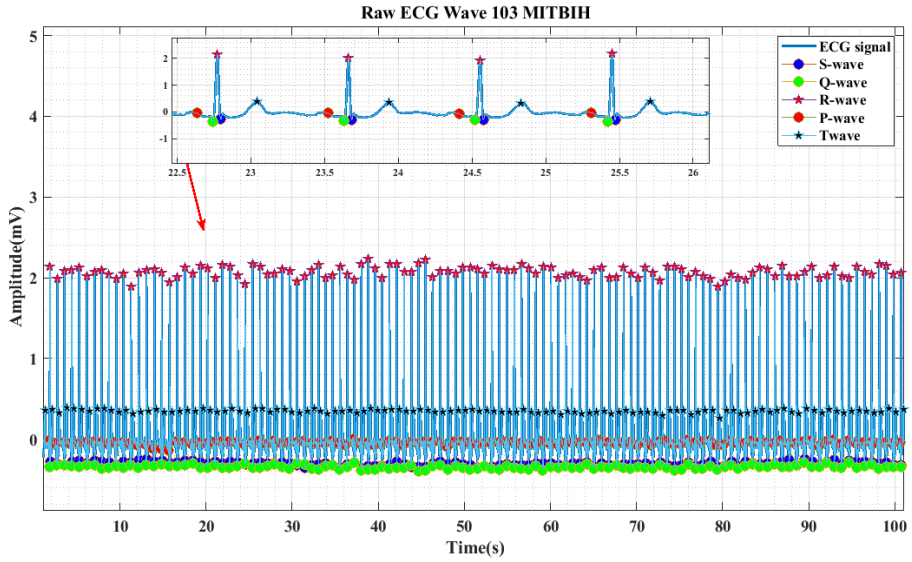

Figure 7. ECG Wave from MIT-BIH-103: As illustrated in **Figure 7,** the results are all peaks detected in the ECG signal. A random sample has been zoomed in.

2.3 Features Extraction

The natural next step after the preprocessing step is typically feature extraction, which is a significant step in biomedical signal classification. Its primary objective is to identify unique patterns and a set of characteristics that can summarize the discriminating information present in the raw ECG signal and allow for the use of this information at the classification stage. By analyzing and comparing these amplitude and interval features extracted from ECG signals, machine learning algorithms can be trained to classify ECG patterns as normal or abnormal.

Feature extraction is important for ECG signal analysis for several reasons. One of the main reasons is that it is non-stationary [31]. Also, noise and artifacts can significantly affect feature extraction, and the complexity of ECG signals is complex and contains multiple components, including P, QRS complexes, and T waves. Each of these components contains important information about the cardiac cycle.

There are many potential features that could be extracted from ECG signals, including time-domain, frequency-domain, morphological, and statistical features [32]. To extract ECG features, there are many techniques used, including DWT [33-34], independent component analysis (ICA) [18], the authors [35-36], feature statistic calculation (FSC), etc. It is used to extract different characteristics of denoised ECG signals.

In this paper, we used the morphological features extracted from an ECG signal, including the R wave and nine-time intervals such as the RR interval, which is the most crucial feature for PVC detection and monitoring ventricular activity, and the previous RR interval, which indicates timing differences between the current and pre-RR intervals. The post-RR interval, which is the time between the current and the next heartbeat, can be abnormal due to PVC, resulting in a shorter than normal post-RR interval. Other features extracted for PVC diagnosis include the QRS complex, which is abnormally wide in PVC; the PR interval, which reflects the duration between the electrical impulse passing from the SA node through the atrioventricular (AV) node and into the ventricles; and the QT interval, which reflects the duration of ventricular depolarization and repolarization. The ST interval reflects the initial ventricular repolarization, while the QR and RS lengths represent the electrical activity during ventricular depolarization. Finally, the R amplitude, which is larger than 2 mV in PVC, is also used as a feature, as shown in Figure 8.

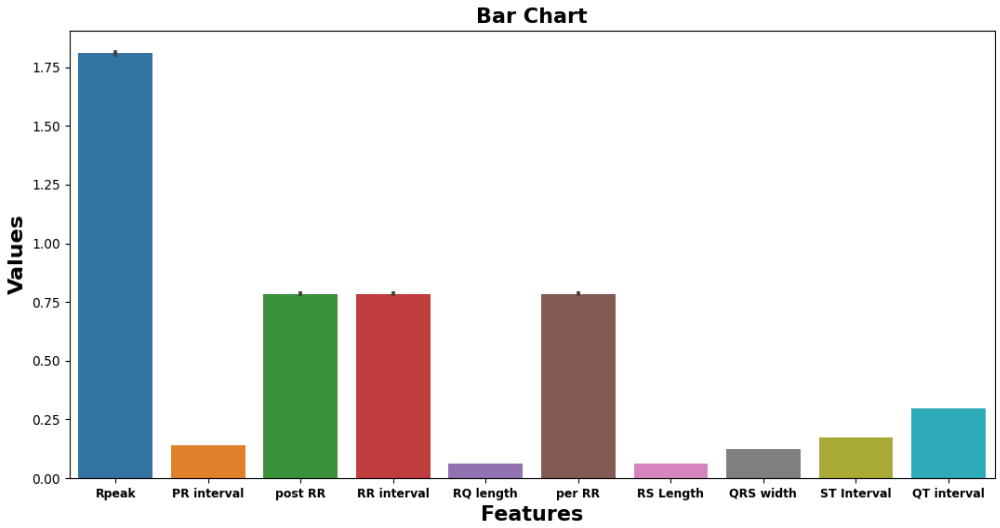

Figure 5. Feature extraction: Extracted time-domain features from ECG signal.

2. 4 Feature Selection

Feature selection is an essential stage in the process of developing predictive models or classification algorithms. It involves selecting a subset of relevant features from a larger set of available features to improve the accuracy and efficiency of the algorithm. Feature selection is a crucial step in the development of predictive models or classification algorithms. Proper feature selection can lead to various benefits, such as reducing the training duration, enhancing generalization by reducing over fitting, and improving prediction performance. Various techniques are used for feature selection, such as genetic algorithms, techniques for optimizing problems using high-parameter optimization [37], and a random forest under the category of embedded methods. They are implemented by algorithms with their own internal feature selection methods. [12], the author [33] uses principal component analysis (PCA), a statistical method that allows you to summarize the dimensions of a database and preserve the most important patterns and relationships between variables without knowledge of target variables, and the author [18] uses ICA to find a linear transformation of the data and generate an independent component set to reduce features.

In this paper, we focus on the wrapper method techniques as a feature selection technique. Wrapper Methods: This type of method selects the features by training a model and selecting the most effective combination of features that depend on the model's performance. Sequential forward selection (SFS) is a wrapper method that starts with no features and gradually adds features that increase the rating performance in each iteration until a specific number of features is reached. The stopping criteria for SFS can be a limit on the number of relevant features or the generation of all possible subsets. The average score accuracy of five cross-validation scores for the k-nearest neighbors model trained on the selected features is evaluated to be 96.74% on the test set, as shown in Figure 9.

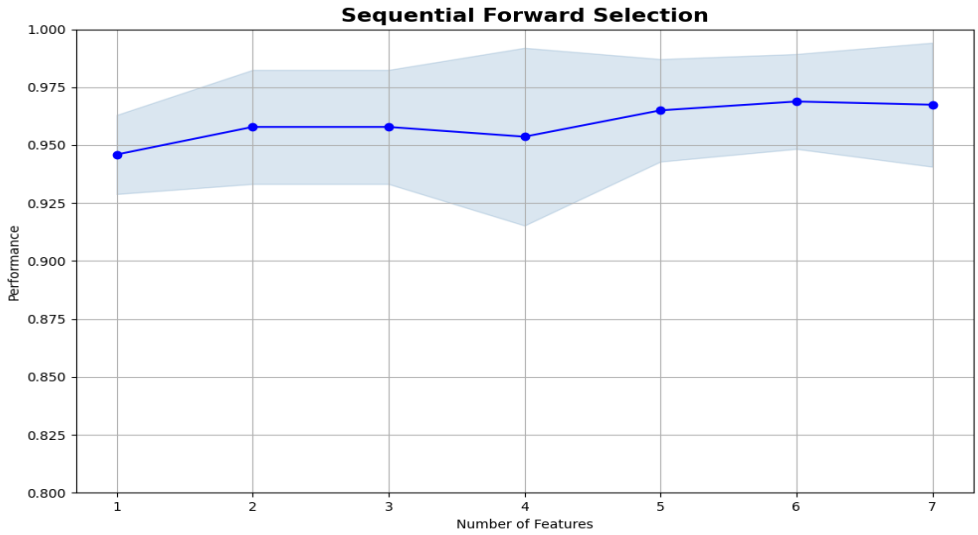

Figure 9. Performance of Selected Features: The Results After Testing the KNN Model

In Figure 10, it is stated that these seven features play a useful role in PVC classification. It was finally found that the k-nearest neighbors model trained on the selected features evaluated an average score of 96.74% on the test set.
Using these seven features can help achieve a good classification result on the test set. Therefore, the R amplitude (R peak), RR interval, QRS width, PR interval, per RR interval, RQ length, and RS length were applied for classification purposes.

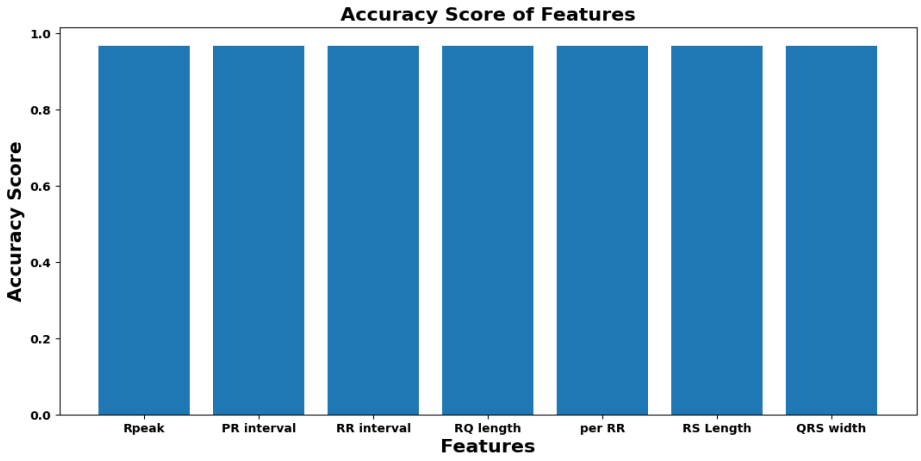

Figure 10. *Best Features Selected: The Results of the SFS Technique.*

2.5 Classification PVC Cardiac Arrhythmia

The goal of classification is to accurately estimate the label class for each instance in the original ECG signal. The algorithms can be classified based on supervised learning algorithms for making predictions. In this study on classification prediction using SVM, RF, and QDA, there are reasons to choose the classifiers for all popular supervised ML algorithms. SVM is a powerful algorithm with a wide range of kernel functions that allows it to model complex decision boundaries. It is also a regularization parameter that helps to control overfitting problems and allows it to generalize better on unseen data. RF is multiple decision trees combined using the

ensemble to make a robust model, make it more accurate, reduce overfitting, and it is also easy to use. QDA allows each class to have its own covariance matrix.

A. SVM is a model that can handle both linear and non-linear data using kernel functions. The classification process of an SVM involves creating a hyperplane and supporting vectors. The performance of the classifier is indicated by the margin between these support vectors. A higher performance is achieved when the margin of the support vectors is larger [38-39].

B. RF is a member of the family of ensemble learning algorithms. The mode of the classes is output as the prediction in this type of decision tree method, which builds many decision trees during training. The classifier uses voting to make the best decision between trees [40]. The most significant parameters affecting the model built by RF are the maximum split, the number of trees, and the maximum depth.

C. QDA is a classic generative probabilistic method for classification problems. It can learn nonlinear relationships between features. It estimates the parameters of the Gaussian distributions for each class and then uses Bayes' theorem to classify the new data points.

3. Results and Discussion

In this paper, we will take a closer look at the Results and Discussion section. We plot our MITBIH dataset for analysis and classification purposes. The ECG signal contains about 37009 beats. The data set is divided into two categories: normal at (0) and PVC beat at (1).

as shown in Figure 11 in the case of MITBIH datasets, which are known for their extremely imbalanced nature. The imbalanced distribution of the database creates a bias in the classification process toward the majority class.

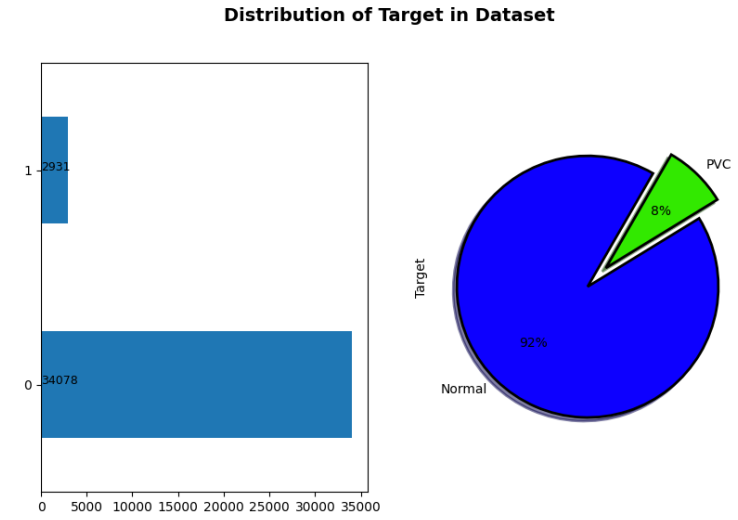

Figure 11. Distribution of Target in the Data: Data consisting of normal and PVC beats

To counteract this, resampling techniques are implemented to balance the number of values in the majority label with the minority label. We applied random oversampling to increase the number of instances in the minority label, which in this case were the normal beats and PVC, to make a balanced dataset, as shown in Figure 12.

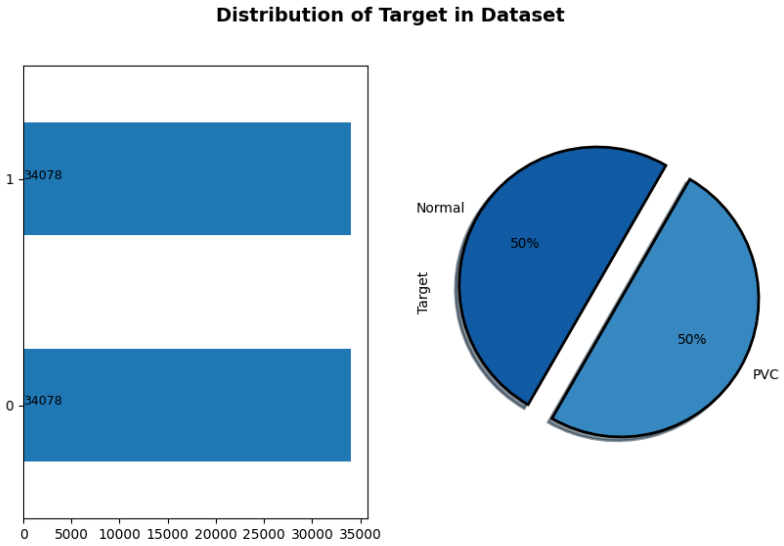

Figure 12. Distribution of Target in a Balanced Dataset:

The Results Oversampling to Equal Normal and PVC Beat.

Next, we divided our dataset into two parts. We used 70% of the dataset for training classifiers, and 30% of the remaining data was reserved for testing purposes. To reduce the risk of variance and bias in the data, we applied both a random train-test split and ten-fold cross-validations. One subset was used for testing, while the other nine subsets were used for training. However, using a larger value of k can help reduce the variance of the evaluation results.

The evaluation of the classifier's performance depends on the confusion matrix using the four fundamental metrics: false negative (FN) and false positive (FP) refer to the total number of items misclassified, and true negative (TN) and true positive (TP) refer to the total number of items correctly detected to fully represent the effectiveness of the test. The model's performance classifiers were evaluated by parameter accuracy (Acc). It should be noted that accuracy alone cannot provide a comprehensive evaluation of a model's performance. Achieving high accuracy does not guarantee that the model is not overfitting the testing set. Therefore, alternative metrics, such as the F1 score, specificity (Spe), sensitivity (Sen), and precision (Pre) [12] where:

$$\text{Acc} = (TP + TN)/(TN + TP + FP + FN)$$

$$\text{Spe} = \frac{TN}{(FP + TN)}$$

$$\text{Pre} = \frac{TP}{(FP + TN)}$$

$$\text{Sen} = \frac{TP}{(FN + TN)}$$

$$\text{F1 score} = \frac{TP}{\text{TP} + 0.5\,(\text{FN} + \text{FP})}$$

The last of the processes involves analyzing the results by evaluating the outcomes of each model. Initially, the primary focus is to examine the results of the tuning of hyperparameters to identify the optimal hyperparameters

for the algorithms. Finally, these results are compared to the most advanced classification studies that employed ML techniques.

The models were trained with three classifiers: SVM, RF, and QDA, written in a Python computing environment.

For each classifier, we employed a 10-fold cross-validation, where the dataset was first randomly selected, then divided into 10 equal sets, and then the dictionaries were trained using 9 folds and tested with the remaining one. To select the best hyperparameters for these ML algorithms, we employed GridSearchCV, a technique for hyperparameter optimization. GridSearchCV systematically searches through a predefined set of hyperparameters and finds the combination that produces the best performance for a given algorithm.

After optimization iteration, 10 k-fold cross-validation (70% train and 30% test) using grid search techniques for each classifier's performance evaluation: the best results for the RF classifier occur when the max features are set to 'auto', the n_estimators parameter is set to 250, and max_depth is set to 15 to achieve the highest average accuracy scores of 99%. Consequently, the RF model was fitted on the training set and evaluated on the test set, resulting in an accuracy of 99.55%, precision of 99.11%, and sensitivity of 100%.

The best results obtained on the test set are as follows, according to the confusion matrix, as depicted in Figure 13.

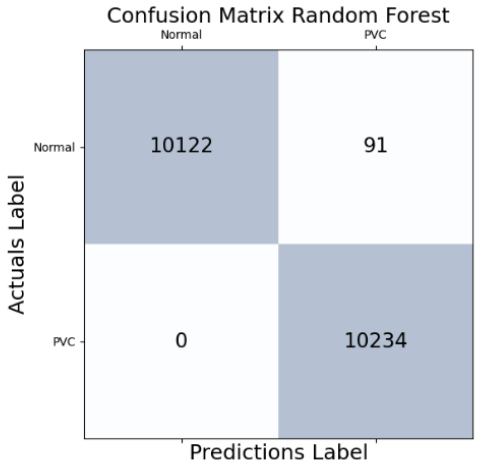

Figure 6. Confusion matrix for the RF model.

Similarly, in SVM, better results are obtained when the gamma parameter is set to 0.001, the C parameter is set to 15, and the ‘rbf’ kernel is used, resulting in the highest average accuracy score of 93.57%. Afterwards, the model was trained on the training set and evaluated on the test set. An accuracy of 93.85% on the testing set was achieved by the SVM algorithm.

The precision and sensitivity were 90.97% and 97.39%, respectively. The best results obtained in the test set; the confusion matrix gives the following results, as shown in Figure 14.

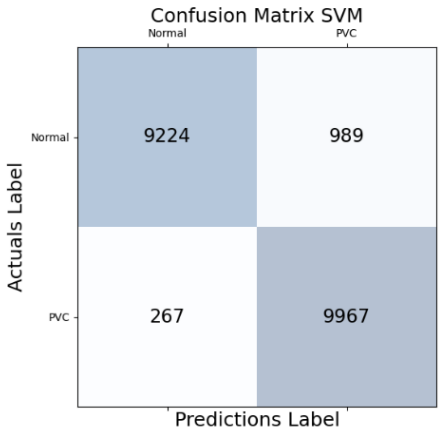

Figure 7. Confusion Matrix for the SVM Model.

In QDA, the best results are obtained when the regularization parameter is set to 0.2, resulting in the highest average accuracy score of 90.04%. Next, a QDA classifier was fitted to the training set and evaluated on the test set.

An accuracy of 90.56% was achieved on the test set by the classifier. The precision and sensitivity were 89.97% and 91.33%, respectively. The best results obtained on the test set are as follows in Figure 15, according to the confusion matrix.

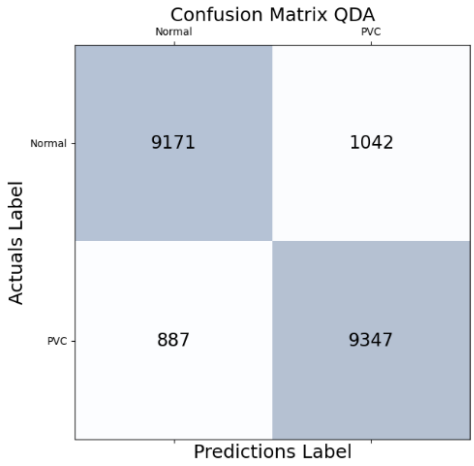

Figure 15: Confusion matrix of the QDA model

Figures 13, 14, and 15, illustrate the two-class confusion matrix for three machine learning classes. Correct predictions for each category are listed on the diagonal of the confusion matrix on the blue background, and all other values outside the diagonal count cases of incorrect predictions.

Overall, the analysis of the results involves carefully examining the performance metrics of each model after hyperparameter tuning to select the best configurations for the RF, SVM, and QDA algorithms.

The results are presented in Table 2, and Figure 16 provides a clearer depiction of the results.

Table 2 Comparison between three algorithms using datasets from MIT-BIH for testing.

Models	Acc (%)	Pre (%)	Sen (%)	Spe (%)	F1Score (%)
SVM	93.85	90.97	97.39	90.31	94.07
RF	**99.55**	**99.11**	**1.00**	**99.10**	**99.55**
QDA	90.56	89.97	91.33	89.79	90.64

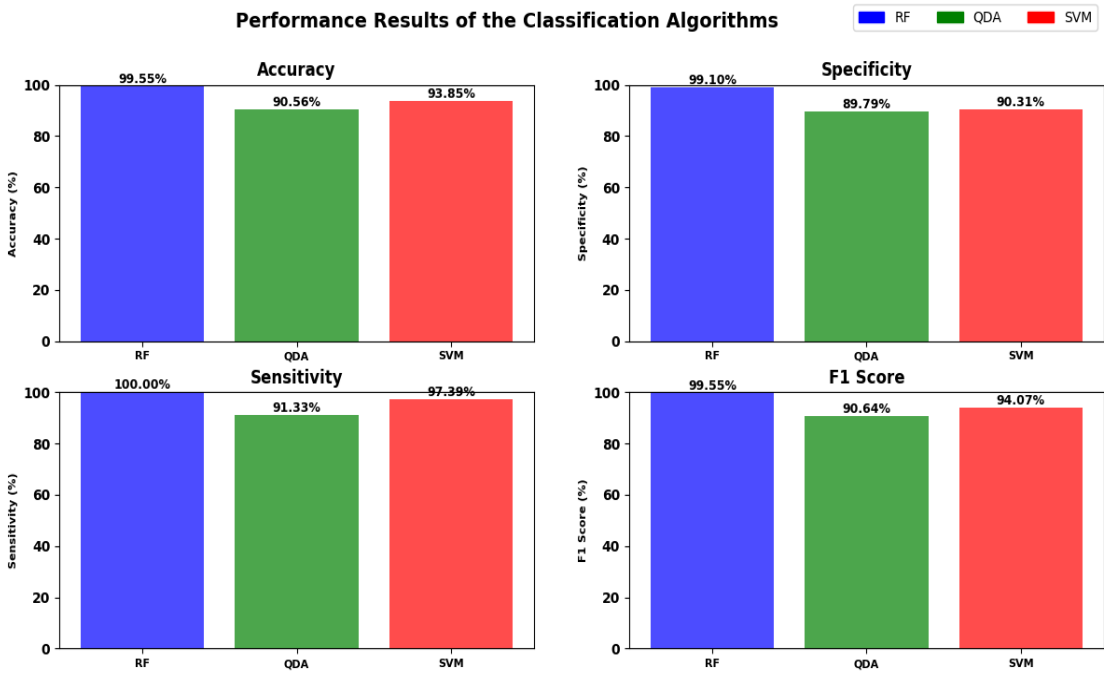

Figure 8. Performance results of the RF, SVM, and QDA classification models.

Table 2 and Figure 16 provide the performance evaluation of various ML classifiers (RF, SVM, and QDA) based on standard metrics. It clearly shows that the results obtained using the QDA algorithm are relatively the lowest.

The random forest algorithm is found to be the best among the other classifiers and has a peak of 99.5% accuracy for sensitivity, specificity 99.5%, precision 99.1%, and F1 score 99.5%.

Figure 17 shows the receiver operating characteristic curve representing the classification results. The area under the curve (AUC) is the direct indication of good classification performance, as proved by the performance metrics in Table 2.

A higher AUC indicates better accuracy and a better model, as displayed in Figure 17.

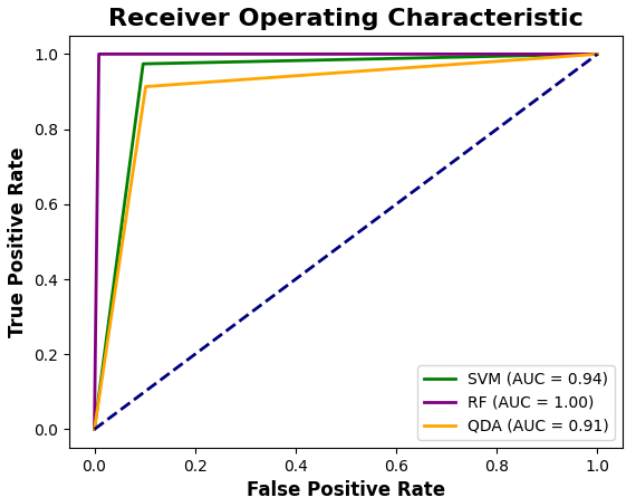

Figure 9. Receiver operating characteristic curve.

The most accurate prediction method would give us a point in the top left corner. for three different ML. The maximum accuracy value for each model and the corresponding threshold value are given as follows: RF model achieving 100% accuracy, SVM model accuracy reaching 94%, and QDA model with 91% accuracy.

The paper focuses mainly on the classification of PVC arrhythmia from ECG signals using ML classification. The attributes also represent different stages of PVC arrhythmia. RF, SVM, and QDA algorithms were tested, and the RF algorithm showed higher performance than other algorithms through overall accuracy, area under ROC curves, and confusion matrix measurements. The results have determined an effective method for the classification of PVC arrhythmia. Table 3 presents a comparative analysis of performance for the proposed study to classify the normal and PVC beats in ECG signals using the MITBIH arrhythmia database compared against existing methods. By comparison, the performance of the proposed method of classification of normal and PVC beasts in ECG signals using the MIT-BIH arrhythmia database with existing methods is given in Table 3.

Table 3 summarizes a comparison of the performance of the proposed method with related work.

References	Accuracy	Sensitivity	Specificity
[9]	-	99.2%	98.6%
[10]	99.7%	97.45%	99.87%
[11]	99.64%	96.97%	99.84%
[12]	96.38%	97.88%	97.56%
[13]	98%	91.1%	98.7%
[14]	99.41%	97.59%	99.54%
This paper (RF)	**99.5%**	**100%**	**99.10%**

The results of our proposed methodology are better when compared to the results in the literature [12] and [14] based on accuracy, and for [10] and [11], their research has exceeded 0.84% of our research accuracy. In terms of sensitivity, our research has achieved a 100% success rate. In the literature [10], [11] and [13], their research managed to outperform 2.15% of our research specificity. Our study involves extracting features and recognizing the heartbeat simultaneously. As a result, our model can determine the heartbeat directly from the MITBIH arrhythmia database.

4. Conclusions

In summary, PVC can have serious health implications if it frequently occurs and is not treated early. This study discusses the use of various ML techniques that can be used to classify ECG from a single lead into normal and PVC arrhythmia classes. involves several important steps. In the beginning, we removed all artifacts from ECG signals using signal processing techniques such as DWT, the IIR notch filter, and the Butterworth filter. We extracted different feature intervals from the morphological features, and feature selection involves choosing the most important features for the classification. Once the features had been extracted and selected, we used a ten-fold technique to divide the dataset into 10 subsets. We apply three different ML classification algorithms that can be used to detect PVC arrhythmia. Overall, the results of the study were evaluated using precision, sensitivity, F1 score, and accuracy metrics on an MIT-BIH Arrhythmia publicly available dataset. Based on the paper results, the RF algorithm outperformed other machine learning algorithms, achieving Spe, Pre, Sen, Acc, and F1-score rates of 99.1%, 99.1%, 100%, 99.5%, and 99.5%, respectively.

REFERENCES

[1] "Cardiovascular diseases (CVDs)." [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] D. Dubin, "Rapid interpretation of EKG's: an interactive course," *2000*.

[3] "Sudden Cardiac Arrest: Causes & Symptoms." Accessed: May 25, 2023. [Online]. Available: https://my.clevelandclinic.org/health/diseases/17522-sudden-cardiac-death-sudden-cardiac-arrest

[4] D. P. Zipes and H. J. J. Wellens, "Sudden Cardiac Death," *Circulation*, vol. 98, no. 21, pp. 2334–2351, Nov. 1998, doi: 10.1161/01.CIR.98.21.2334.

[5] V. Krasteva and I. Jekova, "QRS template matching for recognition of ventricular ectopic beats," *Ann Biomed Eng*, vol. 35, no. 12, pp. 2065–2076, Dec. 2007, doi: 10.1007/s10439-007-9368-9.

[6] M. R. Risk, J. F. Sobh, and J. P. Saul, "Beat detection and classification of ECG using self organizing maps," *Annual International Conference of the IEEE Engineering in Medicine and Biology - Proceedings*, vol. 1, pp. 89–91, 1997, doi: 10.1109/iembs.1997.754471.

[7] C. C. Chiu, T. H. Lin, and B. Y. Liau, "Using correlation coefficient in ECG waveform for arrhythmia detection," *Biomed Eng (Singapore)*, vol. 17, no. 3, pp. 147–152, Jun. 2005, doi: 10.4015/S1016237205000238.

[8] N. Uchaipichat, C. Thanawattano, and A. Buakhamsri, "Wavelet power spectrum analysis for PVC discrimination," in *Lecture Notes in Engineering and Computer Science*, 2013, pp. 1316–1319.

[9] N. T. Sarshar and M. Mirzaei, "Premature Ventricular Contraction Recognition Based on a Deep Learning Approach," *J Healthc Eng*, vol. 2022, 2022, doi: 10.1155/2022/1450723.

[10] J. Yu, X. Wang, X. Chen, and J. Guo, "Automatic Premature Ventricular Contraction Detection Using Deep Metric Learning and KNN," *Biosensors 2021, Vol. 11, Page 69*, vol. 11, no. 3, p. 69, Mar. 2021, doi: 10.3390/BIOS11030069.

[11] J. Yu, X. Wang, X. Chen, and J. Guo, "Searching for Premature Ventricular Contraction from Electrocardiogram by Using One-Dimensional Convolutional Neural Network," *Electronics 2020, Vol. 9, Page 1790*, vol. 9, no. 11, p. 1790, Oct. 2020, doi: 10.3390/ELECTRONICS9111790.

[12] T. Xie, R. Li, S. Shen, X. Zhang, B. Zhou, and Z. Wang, "Intelligent Analysis of Premature Ventricular Contraction Based on Features and Random Forest," *J Healthc Eng*, vol. 2019, 2019, doi: 10.1155/2019/5787582.

[13] B. R. de Oliveira, C. C. E. de Abreu, M. A. Q. Duarte, and J. Vieira Filho, "Geometrical features for premature ventricular contraction recognition with analytic hierarchy process based machine learning algorithms selection," *Comput Methods Programs Biomed*, vol. 169, pp. 59–69, Feb. 2019, doi: 10.1016/J.CMPB.2018.12.028.

[14] F. yan Zhou, L. peng Jin, and J. Dong, "Premature ventricular contraction detection combining deep neural networks and rules inference," *Artif Intell Med*, vol. 79, pp. 42–51, Jun. 2017, doi: 10.1016/J.ARTMED.2017.06.004.

[15] G. B. Moody and R. G. Mark, "The impact of the MIT-BIH arrhythmia database," *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, no. 3, pp. 45–50, 2001, doi: 10.1109/51.932724.

[16] Leif Sörnmo and Pablo Laguna, "BIOELECTRICAL SIGNAL PROCESSING IN CARDIAC AND NEUROLOGICAL APPLICATIONS," 2005. [Online]. Available: https://www.sciencedirect.com/book/9780124375529/bioelectrical-signal-processing-in-cardiac-and-neurological-applications#book-description

[17] D. Zhang, "Wavelet approach for ECG baseline wander correction and noise reduction," *Annual International Conference of the IEEE Engineering in Medicine and Biology - Proceedings*, vol. 7 VOLS, pp. 1212–1215, 2005, doi: 10.1109/IEMBS.2005.1616642.

[18] N. Z. N. Jenny, O. Faust, and W. Yu, "Automated classification of normal and premature ventricular contractions in electrocardiogram signals," *J Med Imaging Health Inform*, vol. 4, no. 6, pp. 886–892, Dec. 2014, doi: 10.1166/jmihi.2014.1336.

[19] A. Vishwa, M. K. Lal, S. Dixit, and P. Vardwaj, "Clasification Of Arrhythmic ECG Data Using Machine Learning Techniques," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 1, no. 4, p. 67, 2011, doi: 10.9781/ijimai.2011.1411.

[20] A. Almumri, E. Balakrishnan, and S. Narasimman, "Discrete Wavelet Transform Based Feature Extraction in Electrocardiogram Signals," 2021. [Online]. Available: http://www.ripublication.com/gjpam.htm

[21] C. Yang, G. W. Ku, J. G. Lee, and S. H. Lee, "Interval-based LDA algorithm for electrocardiograms for individual verification," *Applied Sciences (Switzerland)*, vol. 10, no. 17, Sep. 2020, doi: 10.3390/app10176025.

[22] R. Kher, "Signal Processing Techniques for Removing Noise from ECG Signals," 2019. [Online]. Available: http://creativecommons.org/licenses/by/3.0/,which

[23] S. Destercke, A. Rico, and O. Strauss, "Approximating general kernels by extended fuzzy measures: application to filtering," pp. 112–123, 2020, doi: 10.1007/978-3-030-50143-3_9ï.

[24] M. Olazabal *et al.*, "Analysis of Pan-Tompkins Algorithm Performance with Noisy ECG Signals," *J Phys Conf Ser*, vol. 1532, no. 1, p. 012022, Jun. 2020, doi: 10.1088/1742-6596/1532/1/012022.

[25] X. Xiaomin and L. Ying, "Adaptive Threshold for QRS Complex Detection Based on Wavelet Transform," *Conf Proc IEEE Eng Med Biol Soc*, vol. 2005, pp. 7281–7284, 2005, doi: 10.1109/IEMBS.2005.1616192.

[26] J. Arteaga-Falconi, H. Al Osman, and A. El Saddik, "R-peak detection algorithm based on differentiation," *WISP 2015 - IEEE International Symposium on Intelligent Signal Processing, Proceedings*, Jun. 2015, doi: 10.1109/WISP.2015.7139157.

[27] N. Manjula, N. P. Singh, and P. A. Babu, "An Efficient Designing of IIR Filter for ECG Signal Classification Using MATLAB," *Engineering Proceedings 2023, Vol. 34, Page 24*, vol. 34, no. 1, p. 24, Mar. 2023, doi: 10.3390/HMAM2-14154.

[28] A. Peimankar and S. Puthusserypady, "DENS-ECG: A deep learning approach for ECG signal delineation," *Expert Syst Appl*, vol. 165, p. 113911, Mar. 2021, doi: 10.1016/J.ESWA.2020.113911.

[29] H. D. Hesar and M. Mohebbi, "An Adaptive Kalman Filter Bank for ECG Denoising," *IEEE J Biomed Health Inform*, vol. 25, no. 1, pp. 13–21, Jan. 2021, doi: 10.1109/JBHI.2020.2982935.

[30] P. N. Malleswari, C. H. Bindu, and K. S. Prasad, "A hybrid EMD-DWT based algorithm for detection of QRS complex in electrocardiogram signal," *J Ambient Intell Humaniz Comput*, vol. 13, no. 12, pp. 5819–5827, Dec. 2022, doi: 10.1007/S12652-021-03268-9/METRICS.

[31] S. Krishnan, "Biomedical Signal Analysis for Connected Healthcare," *Biomedical Signal Analysis for Connected Healthcare*, pp. 1–325, Jan. 2021, doi: 10.1016/C2016-0-04436-4.

[32] A. K. Singh and S. Krishnan, "ECG signal feature extraction trends in methods and applications," *Biomed Eng Online*, vol. 22, no. 1, p. 22, Dec. 2023, doi: 10.1186/S12938-023-01075-1/FIGURES/11.

[33] Turker Ince, Serkan Kiranyaz, and Moncef Gabbouj, *Automated patient-specific classification of premature ventricular contractions*. I E E E, 2008. doi: 10.1109/IEMBS.2008.4650453.

[34] F. I. Alarsan and M. Younes, "Analysis and classification of heart diseases using heartbeat features and machine learning algorithms," *J Big Data*, vol. 6, no. 1, Dec. 2019, doi: 10.1186/s40537-019-0244-x.

[35] M. M. Casas, R. L. Avitia, F. F. Gonzalez-Navarro, J. A. Cardenas-Haro, and M. A. Reyna, "Bayesian Classification Models for Premature Ventricular Contraction Detection on ECG Traces," *J Healthc Eng*, vol. 2018, 2018, doi: 10.1155/2018/2694768.

[36] M. M. Casas, R. L. Avitia, M. A. Reyna, and A. Cárdenas, *Evaluation of Three Machine Learning Algorithms as Classifiers of Premature Ventricular Contractions on ECG beats.* 2016. doi: 10.1109/GMEPE-PAHCE.2016.7504615.

[37] Y. Kaya and H. Pehlivan, "Feature selection using genetic algorithms for premature ventricular contraction classification," *ELECO 2015 - 9th International Conference on Electrical and Electronics Engineering*, pp. 1229–1232, Jan. 2016, doi: 10.1109/ELECO.2015.7394628.

[38] Q. Wu and D.-X. Zhou, "Analysis of support vector machine classification," 2006. [Online]. Available: https://www.researchgate.net/publication/238652514

[39] M. Awad and R. Khanna, "Support Vector Machines for Classification," in *Efficient Learning Machines*, Apress, 2015, pp. 39–66. doi: 10.1007/978-1-4302-5990-9_3.

[40] L. Breiman, "Random forests," *Mach Learn*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324/METRICS.