International Journal of Basic and Applied Sciences, 14 (8) (2025) 540-550

International Journal of Basic and Applied Sciences

Website: www.sciencepubco.com/index.php/IJBAS
https://doi.org/10.14419/hcmee071
Research paper

Autonomous Detection of Cardiac Ailments Using ECG Signals and Random Forest Classifier

Gunturu Naga Lakshmi [1] *, S. Nagakishore Bhavanam [2], Vasujadevi Midasala [3], Dvsrk Chaitanya [4]

[1] Af Research Scholars, Department of Computer Science and Engineering, Acharya Nagarjuna University College of Engineering and Technology, Acharya Nagarjuna University
[2] Professor, Department of Computer Science and Engineering, Mangalayatan University Jabalpur, Jabalpur
[3] Associate Professor, Department of Computer Science and Engineering, Mangalayatan University Jabalpur, Jabalpur
[4] Department of Civil Engineering, Acharya Nagarjuna University College of Engineering and Technology, Acharya Nagarjuna University
**Corresponding author E-mail: gunturunagalakshmi@gmail.com*

Received: June 24, 2025, Accepted: October 24, 2025, Published: December 26, 2025

Abstract

The ECG signal is a vital resource in diagnosing various cardiac ailments, providing crucial information for accurate decision-making regarding different types of heart diseases. To achieve autonomous detection of cardiac ailments, several strategies have been proposed to extract critical features from the ECG signal with utmost precision. In this study, a state-of-the-art methodology is introduced for the automatic detection of cardiac ailments. The proposed methodology encompasses three main steps: pre-processing, feature extraction, and classi-fication. In the pre-processing step, a Butterworth third-order band-pass filter is utilized to refine the ECG signal. For feature extraction, a four-level maximal overlap discrete wavelet packet transform (MODWPT) technique is employed, utilizing the symlet wavelet as the mother wavelet. In the final stage of classification, five supervised machine learning algor2 ns are applied to classify the considered three cardiac ailments from the MIT-BIH database: Arrhythmia, Congestive Heart Failure, Atrial Fibrillation, and Normal Sinus Rhythm. The algorithms used include Support Vector Machine (SVM), K-nearest Neighbour (KNN). Naive Bayes (NB), Decision Tree (DT), and Random Forest (RF). These classifiers yield overall accuracies of 90.83%, 90.56%, 90.28%, 91.39%, and 91.94% respectively. Based on the results, it is evident that the Random Forest classifier exhibits superior accuracy among the proposed methodology's classifiers for the multiclass classi-fication of cardiac ailments.

Keywords: *ECG; Anamoly Detection; Classification; Random Forest.*

1. Introduction

Cardiac ailments have been prevalent among humans for many years, and early diagnosis is crucial for improving life expectancy. Cardiologists often use automatic detection of cardiac ailments through Electrocardiogram (ECG) analysis, particularly for long-term ECG records. The ECG provides an electrical representation of heart activity in the form of a signal, and processing this signal enables the discrimination of cardiac ailments. This discrimination can be automated through state-of-the-art techniques involving pre-processing, feature extraction, and classification. Various methods have been introduced to pre-process the ECG signal and eliminate motion artifacts such as electromyographic (EMG) noise, baseline wander, and power line interference. These methods include the Monte Carlo Filter, bandpass filters, adaptive filters, wavelet transform-based approaches, independent component analysis, and empirical mode decomposition. In the pre-processing stage, Jung and Lee utilized wavelet basis functions with Daubechies 4 as the mother wavelet to remove noise from the ECG signal.

Feature extraction methods commonly employed to capture complex features of the ECG signal include Discrete Wavelet Transform, Empirical Mode Decomposition, Complete Ensemble Empirical Mode Decomposition (CEEMD), Variational Mode Decomposition, Discrete Cosine Transform, and Maximal Overlap Discrete Wavelet Transform (MODWT). Techniques such as Artificial Neural Networks, Support Vector Machines. Decision Trees, K-nearest Neighbour, and Linear Discriminants are utilized for the classification of the extracted features. However, state-of-the-art automated ECG recognition systems often rely on pattern-recognition frameworks that represent the ECG signal as a sequence of stochastic patterns.

These systems require complex feature extraction methods and high sampling rates, making them time-consuming processes. To implement these systems in real-time scenarios at a reasonable cost, it is necessary to simplify the feature set and reduce the sampling rate. In this work, the Maximal Overlap Discrete Wavelet Packet Transform (MODWPT) technique is employed to detect characteristic waves and extract the necessary features for discriminating the considered cardiac ailments.

The structure of this paper is as follows: Section II presents a literature review conducted to support the implementation of this study. Section III describes the different stages of the proposed methodology and provides details on the implementation process. Section IV discusses and summarizes the results obtained when testing the proposed methodology on cardiac ailments databases obtained from the MIT-BIH Physio net website. Finally, Section V concludes by summarizing the key points highlighted in this paper.

2. Literature Review

2.1. Electrocardiogram (ECG)

The electrical activity of the heart can be visualized graphically through an Electrocardiogram (ECG) by placing 12-lead electrodes at different locations on the body. These electrodes are capable of capturing subtle electrical changes that occur during each cardiac cycle (heartbeat) as a result of cardiac muscle depolarization and repolarization [17]. The normal ECG waveform consists of distinct characteristic waves, namely the 'P' wave, 'QRS' complex, and 'T' wave. The 'P' wave represents atrial depolarization, which occurs before contraction. The 'QRS' complex signifies ventricular depolarization, indicating the onset of contraction. Both the 'P' wave and 'QRS' complex represent depolarization waves within the heart. Finally, the T' wave corresponds to ventricular repolarization, representing the recovery phase after depolarization. Thus, the electrocardiogram captures both depolarization and repolarization events. Figure 1, displayed below, illustrates the peaks and durations of these characteristic waves in an ECG signal.

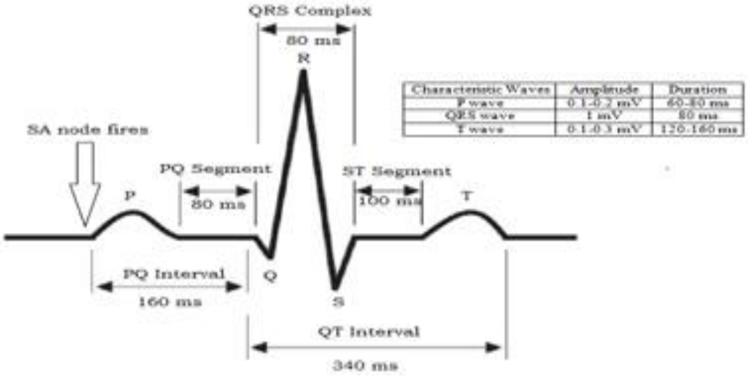

Fig. 1: Graphical Representation of ECG Signal.

2.2. Support vector machine

Support Vector Machine (SVM) is most broadly used for classification problems. It works very well for classifying higher-dimensional data (lots of features) [12]. The basic idea behind SVM is to detect the optimum maximum 1ginal hyper plane (MMH). To create this hyperplane SVM selects the extreme points/vectors. These extreme points are called as support vectors, and hence algorithm is called as Support Vector Machine. The Error-Correcting Output Codes (ECOC) is a method which works on multi-class classification problem by reframing it as multiple binary classification problems [18]. Thus, SVM can be applied for multiclass classification problem by using ECOC method.

2.3. k-Nearest neighbour

The k-Nearest Neighbour algorithm is a simple supervised machine learning algorithm which is an example-based learning group. It classifies the data based on similarity of data from the other [19]. The considered data is split into train and test datasets and the algorithm learns from the train data and groups into defined categories. For the test data to be classified into category we need to choose the number of k neighbours. Compute the k neighbours of the test data point according to some distance measure such as Euclidean distance given by

$$d_{xy} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{1}$$

Where (x_2, y_2) are training object coordinates and (x_1, y_1) is testing object coordinates.

The algorithm counts number of data points from each category among the k neighbours computed previously. The test data point is assigned to the category with most neighbours and the process continues on all test data points for chosen k neighbours.

2.4. Naive bayes classifier

Naive Bayes is a classification algorithm based on Bayes theorem which is simple and most efficient. It is similar to Bayesian network in which all attributes are independent of given class variable. This conditional independency of the attributes of Bayes theorem can be called as Naive bayes [20]. The probability function for Naive Bayes classifier for multiclass classification is given as;

$$p\{A = k|B_1, B_2, B_{3,} \ldots . . B_p\} = \frac{\pi(A=k)\prod_{j=1}^{p} p(B_j|A = k)}{\sum_{k=1}^{k} \pi(A=k)\prod_{j=1}^{p} p(B_j|A = k)} \tag{2}$$

In equation) which is used to compute the value of probability [21], k indicates number of classes for classification, A is the random variable corresponding to the class index of an observation, B1, B2, ..., BP are random predictors of an observation and (A=k) is the prior probability that a class index is k.

2.5. Decision tree

The decision tree classification algorithm is a supervised learning method that is widely used for solving both classification and regression problems. It is considered a relatively simple yet powerful approach in machine learning. The goal of the decision tree algorithm is to create a model that can predict the value of a target variable based on a set of input features. The decision tree algorithm builds tree-like structure where each internal node represents a feature or attribute, and each leaf node represents a class label or a predicted value. The algorithm learns decision rules by recursively partitioning the data based on the values of the input features. These decision rules are inferred from the training data, where the algorithm learns to make optimal splits at each internal node, resulting in homogeneous subsets of data at the leaf nodes.

During the training phase, the decision tree algorithm analyses the training dataset and determines the most informative features for making decisions. It evaluates different split points for each feature and selects the one that maximizes the separation of classes or reduces the variance within the subsets. This process is repeated recursively for each subset until a stopping criterion is met, such as reaching a maximum tree depth or when further splits do not provide significant improvements. Once the decision tree model is constructed, it can be used for predicting the target variable of unseen instances by traversing the tree based on the feature values of the instance. The decision rules encoded in the tree structure guide the prediction process, leading the assignment of a class label or a predicted value.

One of the advantages of the decision tree algorithm is its interpretability. The resulting decision tree can be visualized, allowing users to understand the decision-making process and the importance of different features. Decision trees also handle both categorical and numerical features, making them versatile for various types of datasets. However, decision trees are prone to overfitting, where the model becomes too complex and fits the training data too closely, resulting in poor generalization to unseen data. To mitigate this issue, techniques such as pruning, setting a minimum number of samples for splitting, or using ensemble methods like random forest can be employed.

2.6. Random forest

A random forest is a popular machine learning classifier that consists of multiple decision trees. It is known for its ability to provide accurate and stable predictions. To create a random forest, a collection of classification trees, called a training forest, is generated. The random forest algorithm employs a technique called bagging, which stands for bootstrap aggregating. Bagging involves creating multiple subsets of the training data by sampling with replacement. Each subset is used to train a separate decision tree in the forest. This approach introduces randomness and diversity into the learning process, as each decision tree is trained on a slightly different subset of the data.

During the training phase, each decision tree in the random forest is built using a subset of the features randomly selected from the available features. This further enhances the diversity of the trees in the forest. The trees are constructed by recursively partitioning the data based on the selected features, using techniques such as information gain or Gini impurity to determine the optimal splits at each node. Once the random forest is trained, it can be used to classify new, unseen data. When a classification decision is tired, each tree in the forest independently classifies the input data. In the case of classification, each tree assigns a class label to the input data based on the features it considers important. The final classification result is determined by aggregating the individual decisions made by the trees. This can be done through a voting mechanism, where the class that receives the most votes among the trees is selected as the predicted class. The use of multiple decision trees in a random forest provides several benefits. Firstly, it helps to reduce the risk of overfitting, as the averaging effect of multiple trees tends to make the predictions more robust and less sensitive to noise in the data. Secondly, the random forest can handle a large number of features and effectively capture complex relationships between the features and the target variable. Additionally, the random forest can provide estimates of feature importance, which can be useful for understanding the relative contribution of different features in the classification task.

3. Materials and Methods

The methodology proposed in this paper is shown in the Fig. 2. which has various stages involved for the multiclass classification of the cardiac ailments.

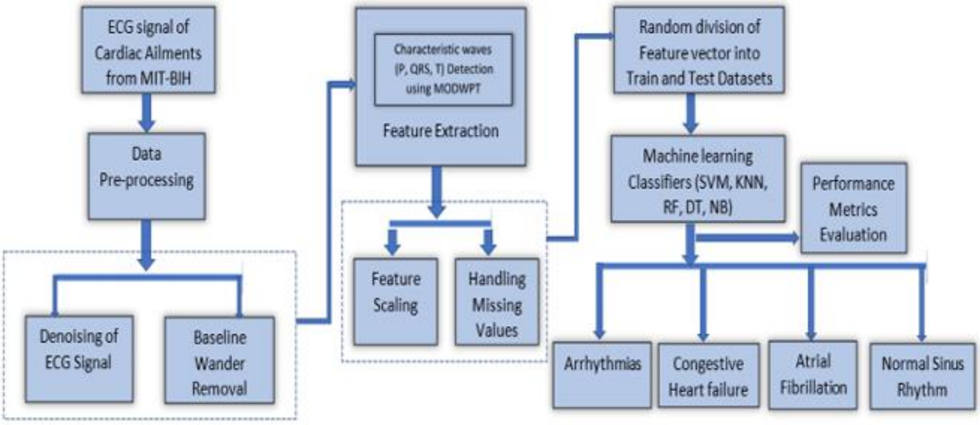
ECG signal of Cardiac Ailments from MIT-BIH
Data Pre-processing
Denoising of ECG Signal
Baseline Wander Removal
Characteristic waves (P, QRS, T) Detection using MODWPT
Feature Extraction
Feature Scaling
Handling Missing Values
Random division of Feature vector into Train and Test Datasets
Machine learning Classifiers (SVM, KNN, RF, DT, NB)
Performance Metrics Evaluation
Arrhythmias
Congestive Heart failure
Atrial Fibrillation
Normal Sinus Rhythm

Fig. 2: Process Flow Chart of the Proposed Methodology.

3.1. Dataset used

In this study, ECG signals representing four different cardiac conditions, namely Arrhythmia, Congestive Heart Failure, Atrial Fibrillation, and Normal Sinus Rhythm, were collected from the freely accessible MIT-BIH database online. The dataset used in this research consists of 300 signals for each cardiac condition, with each signal containing 4120 samples. A comprehensive dataset was constructed by combining a total of 1200 signals sourced from various databases available on the Physio net website [16] MIT-BIH Arrhythmia Database

- BIDMC Congestive Heart Failure Database
- MIT-BIH Atrial Fibrillation Database
- MIT-BIH Normal Sinus Rhythm Database

3.1.1. BIDMC congestive heart failure database

This database includes long-term ECG recordings from 15 subjects (11 men, aged 22 to 71, and 4 women, aged 54 to 63) with severe congestive heart failure (NYHA class 3–4). This group of subjects was part of a larger study group receiving conventional medical therapy before receiving the oral inotropic agent, milrinone. Further details about the larger study group are available in the first reference cited above. Several additional studies have utilized these recordings; see the references below for further details. The individual recordings are each about 20 hours in duration and contain two ECG signals, each sampled at 250 samples per second with 12-bit resolution over a range of ±10 millivolts. The original analog recordings were made at Boston's Beth Israel Hospital (now the Beth Israel Deaconess Medical Center) using ambulatory ECG recorders with a typical recording bandwidth of approximately 0.1 Hz to 40 Hz. Annotation files (with the suffix .ecg) were prepared using an automated detector and have not been corrected manually.

Key features of the BIDMC CHF database:

a) Data Source: Long-term ECG recordings from patients with severe congestive heart failure.
b) Number of Subjects: 15 (11 men, 4 women).
c) Patient Severity: NYHA class 3-4 (severe CHF).
d) Recording Duration: 20 hours per recording.
e) ECG Signals: Two signals sampled at 250 samples/second, 12-bit resolution.
f) Annotation Files: Automated beat annotations (suffix. ecg), not manually corrected.
g) Availability: Publicly available on PhysioNet.

3.1.2. MIT-BIH atrial fibrillation database

The source of the ECGs included in the MIT-BIH Arrhythmia Database is a set of over 4000 long-term Holter recordings that were obtained by the Beth Israel Hospital Arrhythmia Laboratory between 1975 and 1979. Approximately 60% of these recordings were obtained from inpatients. The database contains 23 records (numbered from 100 to 124 inclusive with some numbers missing) chosen at random from this set, and 25 records (numbered from 200 to 234 inclusive, again with some numbers missing) selected from the same set to include a variety of rare but clinically important phenomena that would not be well-represented by a small random sample of Holter recordings. Each of the 48 records is slightly over 30 minutes long. The first group is intended to serve as a representative sample of the variety of waveforms and artifacts that an arrhythmia detector might encounter in routine clinical use. A table of random numbers was used to select tapes, and then to select half-hour segments of them. Segments selected in this way were excluded only if neither of the two ECG signals was of adequate quality for analysis by human experts.

Records in the second group were chosen to include complex ventricular, junctional, and supraventricular arrhythmias and conduction abnormalities. Several of these records were selected because features of the rhythm, QRS morphology variation, or signal quality may be expected to present significant difficulty to arrhythmia detectors; these records have gained considerable notoriety among database users.The subjects consisted of 25 men, aged 32 to 89 years, and 22 women, aged 23 to 89 years. (Records 201 and 202 came from the same male subject.)

3.1.3. MIT-BIH normal sinus rhythm database

The MIT-BIH Normal Sinus Rhythm Database contains 18 long-term ECG recordings of individuals in normal sinus rhythm, meaning their hearts beat with a regular rhythm and rate. These recordings were taken from patients at Boston's Beth Israel Hospital and are intended to serve as a reference for normal heart activity. The database includes ECG waveform data as well as beat annotation files.

Key Details:

a) Purpose: The database serves as a benchmark for evaluating algorithms that analyze heart rhythms, particularly in detecting arrhythmias.
b) Data: It includes 18 recordings of 30 minutes each, digitized at 128 samples per second with 12-bit resolution.
c) Subjects: The recordings are from 5 men (aged 26-45) and 13 women (aged 20-50) who were found to have no significant arrhythmias.
d) Availability: The database is available on PhysioNet and can be used for research purposes.
e) Data Format: The data includes ECG waveform data in a variety of formats, as well as beat annotation files that identify the timing of key events in the heart's electrical cycle.
f) Use: Researchers use this data to develop and test algorithms for detecting and classifying heart rhythm abnormalities.

3.2. Data pre-processing

ECG signals are usually combined with noise and various artifacts such as baseline wander (due to muscle movement), high frequency noise. To have a best accuracy these undesirable noises need to be eliminated. For denoising of ECG signal different types of filters are used. To discard the baseline, wander and high frequency noise a Butterworth third der band pass filter with 5 Hz and 15 Hz as low and high cut off frequencies respectively is designed. As we are using four different types of databases the BPF's cut off frequency and order is altered according to the database's sampling frequency.

3.3. Feature extraction

Feature extraction plays a crucial role in the machine learning algorithms, enabling the classification of ECG signals as either normal or abnormal. In this study, a total of 54 features were extracted, incorporating both detected and calculated attributes derived from the detected features. The extraction process was performed using the maximal overlap discrete wavelet packet transform (MODWPT) technique implemented in MATLAB 2021a version software. Specifically, a four-level symlet wavelet with four vanishing moments was employed to identify the characteristic waves (P, QRS, and T waves) within the ECG signal. Since its four level we get 24 16 coefficients of the signal of which first four are used to reconstruct the signal using inverse MODWPT and the maximum value of this reconstructed signal gives R wave. The remaining characteristic waves, after identifying the P, QRS, and T waves, are obtained through the application of an appropriate moving window technique. These detected characteristic waves serve as the basis for extracting 54 features, including but not limited to:

1) Amplitude features: Minimum amplitude, maximum amplitude, mean amplitude, and standard deviation of the amplitudes.
2) Duration features: Duration of the waves, such as P wave duration, QRS complex duration, and T wave duration.
3) Morphological features: Shape-related features like the slope, curvature, and area under the curve of the waves.
4) Frequency-domain features: Power spectrum analysis to extract frequency-related characteristics, such as dominant frequency and spectral entropy.
5) Time intervals: Intervals between different waves, such as RR interval, PR interval, and QT interval.
6) Waveform-based features: Statistical measures like skewness, kurtosis, and energy of the waveforms.
7) Heart rate variability features: Parameters that describe the variability between consecutive heartbeats, such as standard deviation of RR intervals and frequency domain features (e.g., low-frequency power, high-frequency power, and their ratio).
8) These features are calculated and derived from the detected characteristic waves, providing valuable information for the classification of ECG signals. The Fig. 3. below show characteristic waves detected in ECG signal using MODWPT.

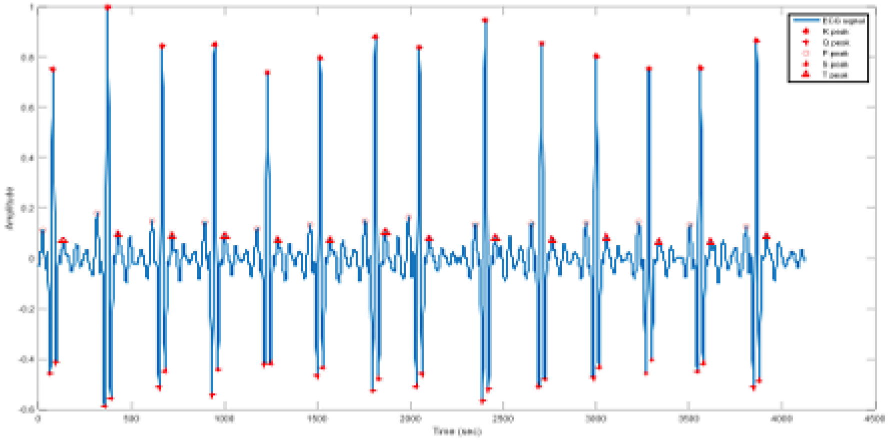

Fig. 3: Characteristic Waves Detection in ECG Signal Using MODWPT.

3.4. Refine the feature set

Once the feature extraction process is complete, the resulting data from the 1200 signals is used to construct a feature vector of size 1200x54. This feature set comprises values that span different ranges, and it is possible for some values to be missing or contain NaN (Not a Number) values. Prior to proceeding to the next stage, which is classification, it is necessary to address these issues by scaling and handling the feature values appropriately. There are two common approaches to handle this scaling: normalization and standardization. In this study, standardization is employed due to the presence of outliers within the dataset. Standardization, also known as Z-score normalization, is utilized to bring the values onto a standardized scale.

Features may be derived from time-domain, frequency-domain, or nonlinear methods. Since they represent different properties, their scales naturally differ. Some signals may not contain enough data to compute a given feature. Example: If a segment is too short, you can't compute spectral entropy. Certain feature extraction functions may fail when data is constant (zero variance → undefined correlation, entropy, etc.). Sometimes preprocessing removes parts of the signal (artifact removal, filtering), leaving empty windows.

$$Z = \frac{[\text{data value} - \text{mean(data value)}]}{[\text{standard deviation(data value)}]} \tag{3}$$

By applying the Z-score normalization (standardization) technique to the feature set, all the data is brought onto the same scale, making it suitable for the subsequent classification stage. In addition to handling scaling, it is important to address missing values that may occur within the dataset. Some values within the feature set can be missing or contain NaN (Not a Number) values due to the feature calculation and are two common approaches to handling missing values: imputation and feature reduction. In this particular dataset, the second approach is employed. Any feature column containing NaN values is completely removed, resulting in a reduction of the feature set from 54 features to 45 features. This process ensures that no incomplete or unreliable feature columns are included in the analysis. As a result, the reduced feature set, now with a size of 1200x45, can be effectively applied to the classification stage, where machine learning algorithms can be used to train models and make predictions based on the available features.

3.5. Classification

After refining the feature set it is fed to various supervised machine learning algorithms for training. The normalized data is randomly divided into train and test data sets in 70% and 30% respectively. The trained data set is given to machine learning algorithms which are implemented in MATLAB 202la version software to fit themselves into a model. This model is used to predict the class for the given test data set. Theme procedure is applied for considered five models of machine learning algorithms i.e., SVM, KNN, Naive Bayes, Decision tree and Random Forest and the model's Confusion Matrix, Accuracy and other performance metrics are evaluated and discussed in the results section.

3.6. Performance metrics evaluation

To evaluate the performance of machine learning models. Confusion matrix is an important tool used which is an N x N matrix where N is number of target classes. In this work N is 4 The main parameters and their measurement using confusion matrix for a multiclass classification problem is shown in the Fig 4 below.

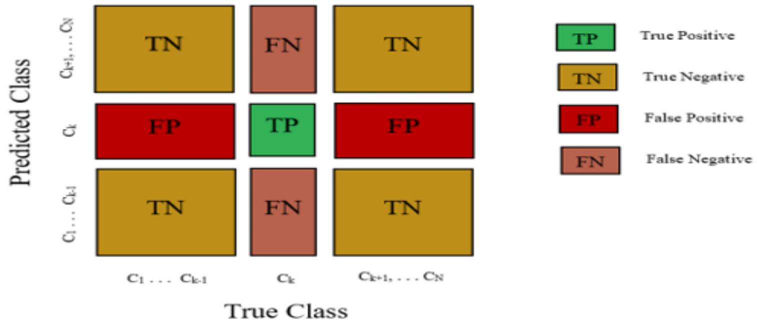

Fig. 4: Confusion Matrix and Its Parameters Measurement for Multiclass Classification.

In multiclass classification problems, unlike binary class classification, the measurement of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) parameters is performed for each individual class. As a result, the measurement of these parameters can vary and is specific to a particular class. True Positive (TP) refers to the scenario in which the predicted class and the true class align for a specific class, denoted as Cx. In other words, TP represents the instances where the model correctly identifies and assigns data points to the class Ck.

Performance metrics are essential in evaluating the effectiveness of deep learning models. These metrics provide insights into how well the model is performing and can guide further improvements. Several common performance metrics used in deep learning include:

1) Accuracy: Accuracy measures the proportion of correctly predicted labels to the total number of Predictions. It provides an overall assessment of the model's correctness.
2) Precision: Precision calculates the proportion of true positive predictions to the total number of 6 positive predictions. It helps evaluate the model's ability to correctly identify positive instances.
3) Recall: Recall, also known as sensitivity or true positive rate, measures the proportion of true positive predictions to the total number of actual positive instances. It assesses the model's ability to identify all positive instances.
4) F1 Score: The F1 score is the harmonic mean of precision and recall. It provides a balanced measure 6 of the model's performance, particularly when dealing with imbalanced datasets.
5) Area Under the ROC Curve (AUC-ROC): The AUC-ROC metric evaluates the model's ability to distinguish between positive and negative instances across various classification thresholds. It quantifies the overall performance by considering the trade-off between true positive rate and false Positive rate.
6) Mean Squared Error (MSE): MSE is commonly used in regression tasks and calculates the average squared difference between predicted and actual values. It measures the model's accuracy in 21imating continuous variables.
7) Mean Absolute Error (MAE): MAE calculates the average absolute difference between predicted and actual values. It provides a measure of the model's accuracy in estimating continuous variables, similar to MSE but without squaring he differences.
8) R-squared (R2): R-squared evaluates the proportion of the variance in the dependent variable that is explained by the model. It ranges from 0 to 1, where 1 indicates a perfect fit.

These performance metrics enable deep learning practitioners to assess the model's accuracy, precision, recall, F1 score, and the ability to handle imbalanced datasets or regression tasks. By considering these metrics, researchers can make informed decisions about model performance and fine-tune their deep learning models accordingly.

4. Results and Discussions

As the cardiac ailments are classified using five different supervised machine learning algorithms, hence the results are discussed in five different sections.

4.1. Training SVM

12 The features set data which is randomly divided into 70% for training and 30% for testing is given to train the SVM model with radial basis kernel function (RBF) which is used to map the training data into kernel space. Error-Correcting Output Codes (ECOC) method is used to fit the data into model and classify the cardiac ailments as it is a multiclass application. A confusion matrix is obtained for this model as shown in the Fig. 5.

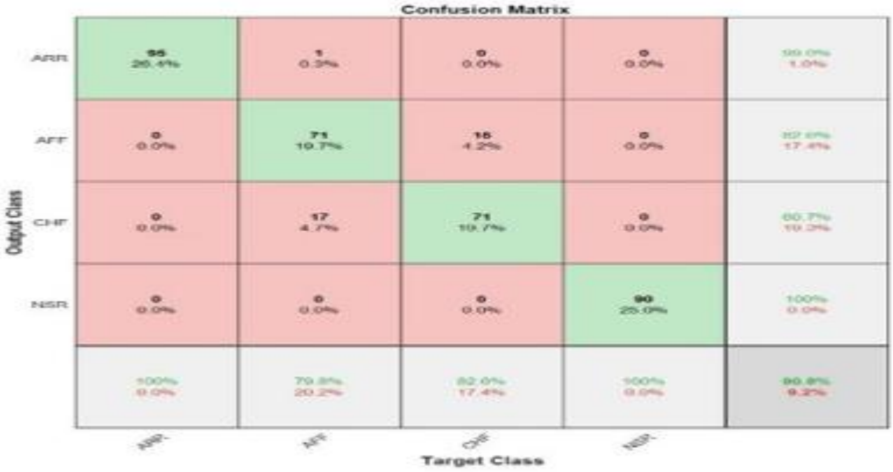

Fig. 5: A Confusion Matrix Is Used for the Classification of Cardiac Ailments Using Support Vector Machine (SVM) Algorithms.

32 From the obtained confusion matrix true positive (TP), true negative (TN), false positive (FP), false negative (FN) are calculated for each class and using this values performance metrics are computed using the formulae given in Table 1. As the application is multiclass the performance metrics are obtained for each class individually as shown in Table 2 below.

Table 2: Performance Metrics of Cardiac Ailments Using SVM

Performance Metrics	Arrhythmia	Atrial Fibrillation	Congestive Heart Failure	Normal Sinus Rhythm
Class Accuracy	0.9972	0.9083	0.9111	1.0
Sensitivity	1.0	0.7978	0.8256	1.0
Specificity	0.9962	0.9446	0.9380	1.0
Precision	0.9896	0.8256	0.8068	1.0
F1 Score	0.9948	0.8114	0.8161	1.0
Mathews Correlation Coefficient	0.9929	0.7511	0.7576	1.0

By using support vector machine, prediction of Arrhythmia disease is done with a class accuracy of 99.72%, Atrial fibrillation with 90.83%, Congestive heart failure with 91.11% and normal sinus rhythm with 100% and MCC value is 1 for normal sinus rhythm data which shows that the SVM model best predicts NSR signals.

4.2. Training KNN

The features set data which is randomly divided into 70% for training and 30% for testing is given to train the KNN model with k=18 which indicates 18 nearest neighbours with Euclidean distance is considered to verify the test data set. As the number of features are high 18 neighbours give optimum results. With these specifications the train data is fit into model and test data applied to the model to classify the four cardiac ailments. A confusion matrix is obtained for this model as shown in the Fig. 6.

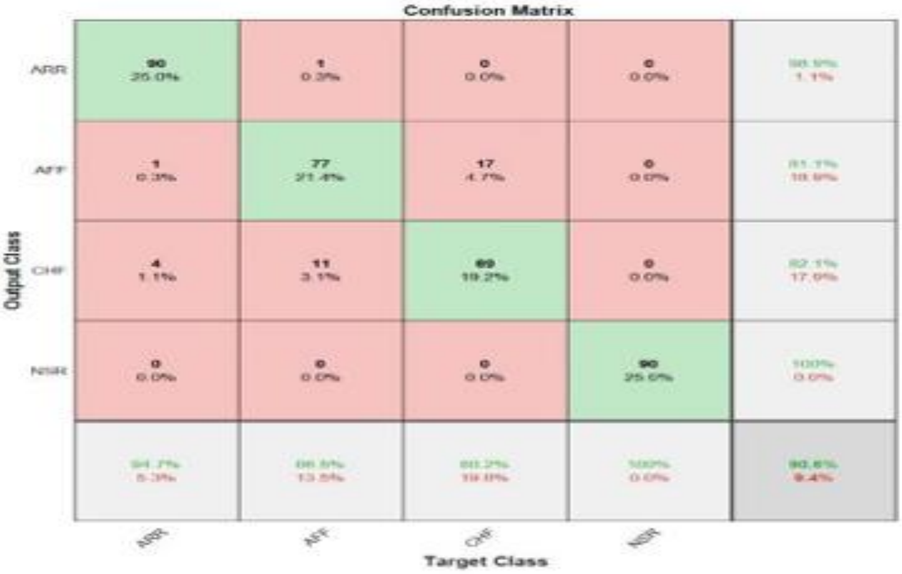

Fig. 6: A Confusion Matrix Is Used for the Classification of Cardiac Ailments Using KNN Algorithms.

As discussed earlier, the performance metrics obtained for each class are shown in Table 3. below.

Table 3: Performance Metrics of Four Cardiac Ailments Using KNN

Performance Metrics	Arrhythmia	Atrial Fibrillation	Congestive Heart Failure	Normal Sinus Rhythm
Class Accuracy	0.9833	0.9167	0.9111	1.0
Sensitivity	0.9474	0.8562	0.8023	1.0
Specificity	0.9962	0.9336	0.9453	1.0

Precision	0.9870	0.8105	0.8214	1.0
F1 Score	0.9677	0.8370	0.8118	1.0
Mathews Correlation Coefficient	0.9569	0.7818	0.7537	1.0

By using K nearest neighbor algorithm, prediction of Arrhythmia disease is done with a class accuracy of 98.33% which is less than SVM, Atrial fibrillation with 91.67% which is more than SVM, Congestive heart failure with 91.11% and normal sinus rhythm with 100% both are similar in case of SVM and MCC value is 1 for normal sinus rhythm data which shows that the KNN model best predicts NSR signals but for arrhythmia data its value is 0.9569 which is relatively low when compared to SVM.

4.3. Training naive bayes

The randomly divided feature set data into train and test of 70% and 30% respectively is given to fit the Naive Bayes model which predicts the data with the probability that is calculated using equation (2) given above. A confusion matrix is obtained for the predicted data using this model as shown in the Fig. 7.

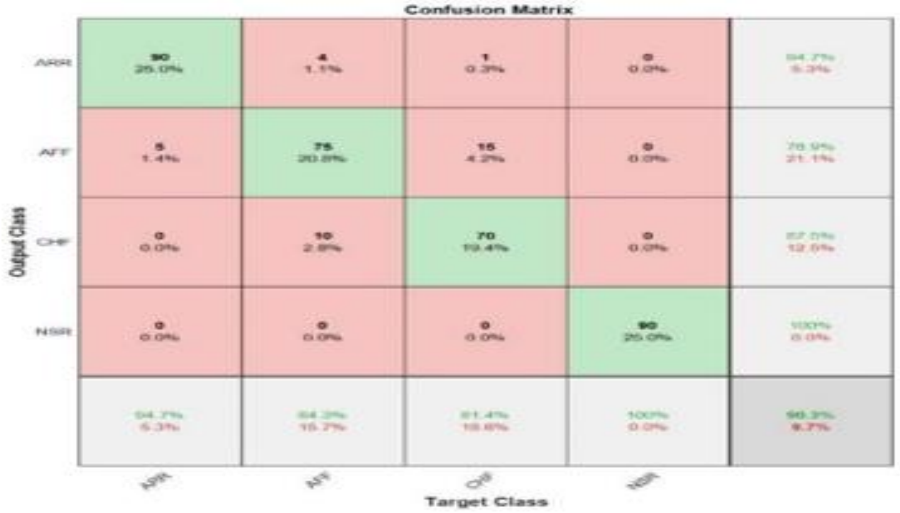

Fig. 7: A Confusion Matrix Is Used for the Classification of Cardiac Ailments Using Naive Bayes.

As discussed earlier, the performance metrics obtained for each class are shown in Table 4 below.

Table 4: Performance Metrics of Four Cardiac Ailments Using Naive Bayes

Performance Metrics	Arrhythmia	Atrial Fibrillation	Congestive Heart Failure	Normal Sinus Rhythm
Class Accuracy	0.9722	0.9056	0.9278	1.0
Sensitivity	0.9474	0.8427	0.8140	1.0
Specificity	0.9811	0.9262	0.9635	1.0
Precision	0.9474	0.7895	0.8750	1.0
F1 Score	0.9474	0.8152	0.8434	1.0
Mathews Correlation Coefficient	0.9285	0.7526	0.7974	1.0

By using Naive Bayes algorithm, prediction of Arrhythmia disease is done with a class accuracy of 97.22% which is less than both SVM and KNN, Atrial fibrillation with 90.56% which is less than both SVM and KNN, Congestive heart failure with 92.78% which is more than both SVM and KNN and normal sinus rhythm with 100% which is similar to SVM and KNN. MCC value is 1 for normal sinus rhythm data which shows that the Naive Bayes model best predicts NSR signals but for arrhythmia data its value is 0.9285 which is relatively low when compared to SVM and KNN.

4.4. Training decision tree

The randomly divided feature set data into train and test of 70% and 30% respectively is given to fit the Decision tree model with maximum number of splits as optimum specification. A confusion matrix is obtained for the predicted data using this model as shown in the Fig. 8.

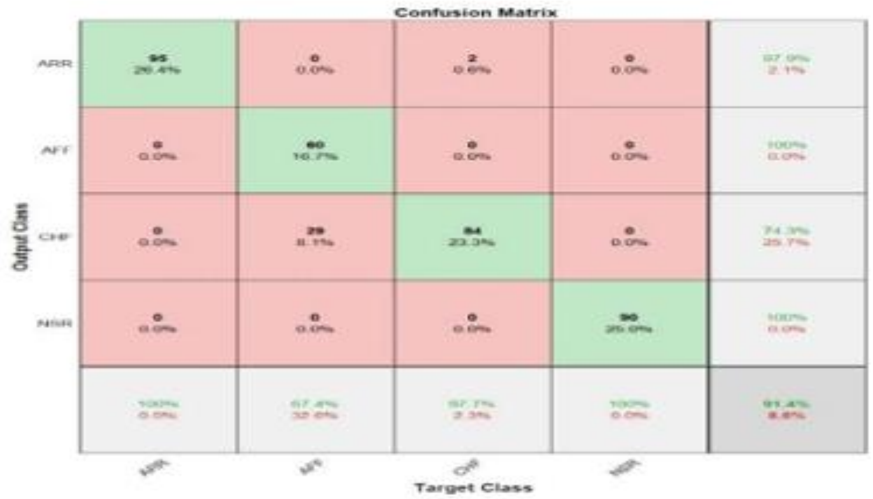

Fig. 8: A Confusion Matrix is Used for the Classification of Cardiac Ailments Using Decision Tree.

As discussed earlier, the performance metrics obtained for each class are shown in Table 5 below.

Table 5: Performance Metrics of Four Cardiac Ailments Using Decision Tree

Performance Metrics	Arrhythmia	Atrial Fibrillation	Congestive Heart Failure	Normal Sinus Rhythm
Class Accuracy	0.9944	0.9194	0.9189	1.0
Sensitivity	1.0	0.6742	0.9767	1.0
Specificity	0.9925	1.0	0.8942	1.0
Precision	0.9794	1.0	0.7434	1.0
F Score	0.9896	0.8054	0.8442	1.0
Mathews Correlation Coefficient	0.9859	0.7804	0.8002	1.0

By using Decision tree algorithm, prediction of Arrhythmia disease is done with a class accuracy of 99.44% which is more than both KNN and Naive bayes but less than SVM and sensitivity is 100% which is not the case in above three algorithms, Atrial fibrillation with 91.94% which is more than previous case, Congestive heart failure with 91.89% and normal sinus rhythm with 100% which is similar to SVM, KNN and Naïve Bayes. MCC value is 1 for normal sinus rhythm data which shows that the Decision tree model best predicts NSR signals similar to previous models. A special observation for Atrial fibrillation data is that it has 100% for both specificity and Precision which is not the case in previous algorithms.

4.5. Training random forest

The features dataset is randomly divided into train and test of 70% and 30% respectively. The train dataset is used to fit the Random Forest model which is multiple decision tree model with Bagging method and 4 maximum number of splits as optimum specification. Once the model is created the test data is applied to the model to predict the cardiac ailments and a confusion matrix is obtained as shown in the Fig 9.

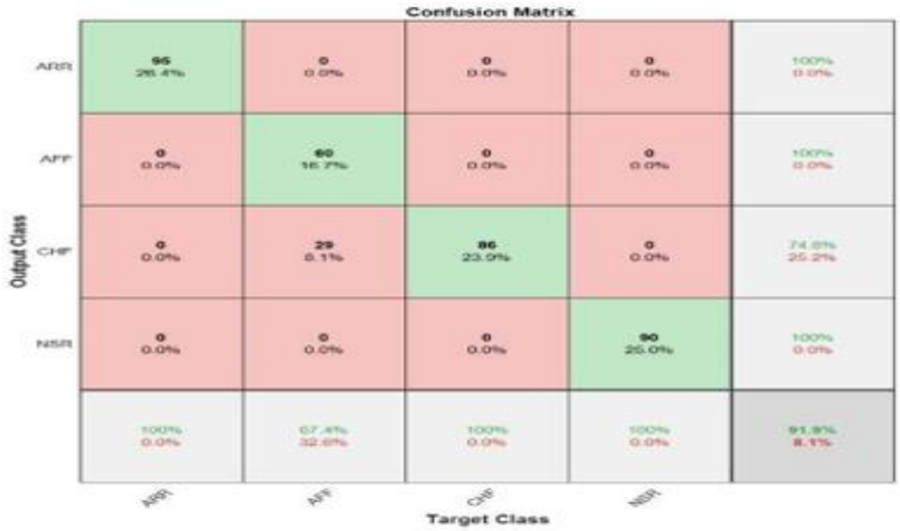

Fig. 9: A Confusion Matrix Is Used for the Classification of Cardiac Ailments Using Random Forest

As discussed earlier, the performance metrics obtained for each class are shown in Table 6 below.

Table 6: Performance Metrics of Four Cardiac Ailments Using Random Forest

Performance Metrics	Arrhythmia	Atrial Fibrillation	Congestive Heart Failure	Normal Sinus Rhythm
Class Accuracy	1.0	0.9194	0.9194	1.0
Sensitivity	1.0	0.6742	1.0	1.0
Specificity	1.0	1.0	0.8942	1.0

Precision	1.0	1.0	0.7478	1.0
F1 Score	1.0	0.8054	0.8557	1.0
Mathews Correlation Coefficient	1.0	0.7804	0.8117	1.0

By using Random Forest algorithm, prediction of Arrhythmia disease is done with a class accuracy of 100% which is more than all the four algorithms. Also, sensitivity, specificity, precision, F1score, and MCC is 100% which is not the case in above four algorithms for arrhythmia data, Atrial fibrillation with 91.94% which is more than previous case, Congestive heart failure with 91.94% and normal sinus rhythm with 100% which is similar to previous models. Also, for NSR data all the six metrics value is 100% which shows Random Forest model best predicts NSR along with Arrhythmia signals. A special observation for Atrial fibrillation data is that it has 100% for both specificity and Precision which is similar to decision tree algorithm and sensitivity for Congestive heart failure data is 100% which not the case in any previous algorithms.

The Overall performance metrics such as accuracy of model, mean MCC Weighted F1 Score and Classification error of the model for classifying of four cardiac ailments using SVM, KNN, Naive Bayes, Decision Tree and Random Forest is given in Table 7 below.

Table 7: Overall Performance Metrics for Classification with Five Machine Learning Models

Overall Performance Metrics	SVM	KNN	Navie Bayes	Decision Tree	Random Forest
Overall Accuracy of Model	90.83	90.56	90.28	91.39	91.94
Mean of MCC	87.54	87.31	86.96	89.16	89.95
Weighted Fi Score	90.86	90.49	90.25	91.59	92.15
Classification Error	9.17	9.44	9.72	8.61	8.06

Clearly, from Table 7 we can say that Random Forest overall accuracy is 91.94% in classifying multiclass data which is high when compared to other machine learning models.

Most of the previous study used binary classes as show in Table 8. The proposed work considered four classes which makes the classification of model complex which in turn effects the accuracy of the model. But Random Forest model is best when compared to other models in previous work as well as in proposed work in terms of accuracy.

Table 8: Comparison of Obtained Results of Proposed Work with the Previous Study

Reference Name	Pre-processing	ECG features	Classifier	Database	Accuracy
Dakun Lai [27]	Median filter, Band pass filter	Morphological features	SVM, KNN. Naive Bayes, Decision Tree, Random Forest	MIT-BIH SDDB, NSRDB	98.7%, 98.91%, 97.46%. 98.99%, 99.49%
Melgani [28]		Morphological, Temporal, RR	SVM-PSO, SVM, KNN	MITDB	89.72%. 85.98%, 88.70%
Proposed work	Band pass filter	Morphological, Fiducial, HRV. statistical	SVM, KNN, Naive Bayes, Decision Tree, Random Forest	MIT-BIH Arrhythmia, Atrial Fibrillation, NSRDB, BIDMC CHFDB	90.83%, 90.56%, 90.28%. 91.39%. 91.94%

5. Conclusion

This paper proposed five different machine learning algorithms- SVM, KNN, Naive Bares. Decision Tree and Random Forest for automatic detection of four Cardia ailments i.e... Arrhythmia, Atrial Fibrillation, Congestive heart failure and Normal sinus rhythm signals obtained from MIT-BIH physio Net database by using various morphological, Fiducial, statistical and HRV features which are obtained by MOD-WPT algorithm. A Butterworth third order band pass filter is used for pre-processing step. The obtained feature set of size 1200x54 is refined to remove some irrelevant data and the size of feature set after handling the missing values is 1200x45. This feature set is given to the considered classifiers which gives overall accuracy of the model as 90.83%, 90.56%, 90.28%, 91.39%, 91.94% for SVM, KNN, Naive Bayes. Decision tree and Random Forest respectively. The obtained accuracy is high for Random Forest algorithm for multiclass classification problem. This can be further improved using advanced neural network like deep neural network (DNN) and convolutional neural network (CNN).

References

[1] S. Banerjee and G. K. Singh, "Monte Carlo Filter-Based Motion Artifact Removal from Electrocardiogram Signal for Real-Time Telecardiology System," in IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1-10, 2021, Art no. 4006110, https://doi.org/10.1109/TIM.2021.3102737.

[2] A. Gotchev, N. Nikolaev and K. Egiazarian, "Improving the transform domain ECG denoising performance by applying inter-beat and intra-beat decorrelating transforms", Proc. IEEE Int. Symp. Circuits Syst. (ISCAS), pp. 17-20, May 2001. https://doi.org/10.1109/ISCAS.2001.920995.

[3] S. Poungponsri and X.-H. Yu, "An adaptive filtering approach for electrocardiogram (ECG) signal noise reduction using neural networks", Neurocomputing, vol. 117, no. 87, pp. 206-213, Oct. 2013. https://doi.org/10.1016/j.neucom.2013.02.010.

[4] W.-H. Jung and S.-G. Lee, "ECG identification based on non-fiducial feature extraction using window removal method", Appl. Sci., vol. 7, no. 11, Nov. 2017. https://doi.org/10.3390/app7111205.

[5] T. He, G. Clifford and L. Tarassenko, "Application of independent component analysis in removing artefacts from the electrocardiogram", Neural Comput. Appl., vol. 15, no. 2, pp. 105-116, Apr. 2006. https://doi.org/10.1007/s00521-005-0013-y.

[6] M. Rakshit and S. Das, "An efficient ECG denoising methodology using empirical mode decomposition and adaptive switching mean filter", Biomed. Signal Process. Control, vol. 40, pp. 140-148, Feb. 2018. https://doi.org/10.1016/j.bspc.2017.09.020.

[7] W.-H. Jung and S.-G. Lee, "ECG identification based on non-fiducial feature extraction using window removal method", Appl. Sci., vol. 7, no. 11, Nov. 2017. https://doi.org/10.3390/app7111205.

[8] H. Chen and K. Maharatna, "An Automatic R and T Peak Detection Method Based on the Combination of Hierarchical Clustering and Discrete Wavelet Transform," in IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 10, pp. 2825-2832, Oct. 2020, https://doi.org/10.1109/JBHI.2020.2973982.

[9] U. Satija, B. Ramkumar and M. S. Manikandan, "A New Automated Signal Quality-Aware ECG Beat Classification Method for Unsupervised ECG Diagnosis Environments," in IEEE Sensors Journal, vol. 19, no. 1, pp. 277-286, 1 Jan.1, 2019, https://doi.org/10.1109/JSEN.2018.2877055.

[10] M. Nazari and S. M. Sakhaei, "Variational Mode Extraction: A New Efficient Method to Derive Respiratory Signals from ECG," in IEEE Journal of Biomedical and Health Informatics, vol. 22, no. 4, pp. 1059-1067, July 2018, https://doi.org/10.1109/JBHI.2017.2734074.

[11] N. Karimian, Z. Guo, M. Tehranipoor and D. Forte, "Highly Reliable Key Generation from Electrocardiogram (ECG)," in IEEE Transactions on Biomedical Engineering, vol. 64, no. 6, pp. 1400-1411, June 2017, https://doi.org/10.1109/TBME.2016.2607020.

[12] L. B. Marinho, N. D. M. M. Nascimento, J. W. M. Souza, M. V. Gurgel, P. P. R. Filho and V. H. C. de Albuquerque. "A novel electrocardiogram feature extraction approach for cardiac arrhythmia classification", Future Gener. Comput. Syst., vol. 97, pp. 564-577. Aug. 2019. https://doi.org/10.1016/j.future.2019.03.025.

[13] M. Thomas, M. K. Das and S. Ari, "Automatic ECG arrhythmia classification using dual tree complex wavelet-based features", AEU-Int. J. Electron. Commun., vol. 69, no. 4, pp. 715-721. Apr. 2015. https://doi.org/10.1016/j.aeue.2014.12.013.

[14] Y. Kaya, H. Pehlivan and M. E. Tenekeci, "Effective ECG beat classification using higher order statistic features and genetic feature selection", Biomed. Res., vol. 28, no. 17. pp. 7594-7603, Aug. 2017.

[15] 1. Christov, V. Krasteva, I. Simova, T. Neycheva and R. Schmid, "Multi-parametric analysis for atrial fibrillation classification in the ECG", Proc. Comput. Cardiol. Conf. (CinC), pp. 1-4, Sep. 2017. https://doi.org/10.22489/CinC.2017.175-021.

[16] https://physionet.org/about/database/ (accessed in October 2020).

[17] S. T. Alam, M. M. Hossain, M. D. Rahman. M. K. Islam," Towards Development of a Low Cost and Portable ECG Monitoring System for Rural/Remote Areas of Bangladesh", International Journal of Image, Graphics and Signal Processing (IJIGSP), Vol. 10, No.5, pp. 24-32, 2018. https://doi.org/10.5815/ijigsp.2018.05.03.

[18] Mikel Galar, Alberto Fernández, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. 2011. An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. Pattern Recogn.44, 8 (August, 2011), 1761-1776. https://doi.org/10.1016/j.patcog.2011.01.017.

[19] T. Denoeux, "A k-nearest neighbor classification rule based on Dempster-Shafer theory", IEEE Trans. Syst. Man. Cybern., vol. 25, no. 5, pp. 804-813, May 1995. https://doi.org/10.1109/21.376493.

[20] S.-B. Kim, K.-S. Han, H.-C. Rim and S. H. Myaeng, "Some effective techniques for Naive Bayes text classification", IEEE Trans. Knowl. Data Eng., vol. 18, no. 11, pp. 1457-1466, Nov. 2006. https://doi.org/10.1109/TKDE.2006.180.

[21] Hastie, T., R. Tibshirani, and J. Friedman. The Elements of Statistical Learning, Second Edition. NY: Springer, 2008.

[22] J. R. Quinlan, "Simplifying decision trees", Int. J. Man-Mach. Stud., vol. 27, no. 3, pp. 221-234, Sep. 1987. https://doi.org/10.1016/S0020-7373(87)80053-6.

[23] A. Liaw and M. Wiener, "Classification and regression by random forest", R Newslett., vol. 2, no. 3, pp. 18-22, 2002.

[24] M. Ingale, R. Cordeiro, S. Thentu, Y. Park and N. Karimian, "ECG Biometric Authentication: A Comparative Analysis," in IEEE Access, vol. 8, pp. 117853-117866, 2020, https://doi.org/10.1109/ACCESS.2020.3004464.

[25] M. Nabian, Y. Yin, J. Wormwood. K. S. Quigley. L. F. Barrett and S. Ostadabbas, "An Open-Source Feature Extraction Tool for the Analysis of Peripheral Physiological Data," in IEEE Journal of Translational Engineering in Health and Medicine, vol. 6, pp. 1-11, 2018, Art no. 2800711, https://doi.org/10.1109/JTEHM.2018.2878000.

[26] Rácz, A., Bajusz, D., & Héberger, K. (2019). Multi-Level Comparison of Machine Learning Classifiers and Their Performance Metrics. Molecules (Basel, Switzerland), 24(15), 2811. https://doi.org/10.3390/molecules24152811.

[27] D. Lai, Y. Zhang, X. Zhang, Y. Su and M. B. Bin Heyat. "An Automated Strategy for Early Risk Identification of Sudden Cardiac Death by Using Machine Learning Approach on Measurable Arrhythmic Risk Markers," in IEEE Access, vol. 7. pp. 94701-94716, 2019, https://doi.org/10.1109/AC-CESS.2019.2925847.

[28] F. Melgani and Y. Bazi, "Classification of Electrocardiogram Signals With withort Vector Machines and Particle Swarm Optimization," in IEEE Transactions on Information Technology in Biomedicine, vol. 12, no. 5. pp. 667-677, Sept. 2008, https://doi.org/10.1109/TITB.2008.923147.