Classification of Cardiac Arrhythmia using Kernelized SVM

Yogita Bhatia, Akanksha Mittal, Shefali Athavale, Tanya Mohanani
Depart of Computer Engineering,
Vivekanand Education Society's Institute of Technology
Mumbai, India

{yogitab2798, akankshamittal1998, shefaliathavale1, belletanu}@gmail.com

Dr.Mrs. Gresha Bhatia
Department of Computer Engineering,
Vivekanand Education Society's Institute of Technology
Mumbai, India
gresha.bhatia@ves.ac.in

Abstract- **Cardiovascular diseases are one of the major causes of death in the world. These diseases include abnormalities in the smooth functioning of the heart causing cardiac arrest, blockages, and other related problems. One such ailment is the irregularities in the heartbeat of the person. Due to this, the movements of the heart are not operating at the normal pace causing palpitations and cardiac arrest. Though Electrocardiogram (ECG) is one of the most popular and widely used methods for monitoring the heart's electrical activity, it becomes quite strenuous for understanding the ECG reports which is a manual approach. So, there is a need to develop a system that could determine the condition a prior and classify them according to its severity. This paper focuses on the ECG deflections, cardiac arrhythmia, and its types. The paper further dwells into the development of an automated system to detect and classify arrhythmia. Various Machine Learning algorithms like Support Vector Machine (SVM), Random Forest Classifier (RF) are analyzed that lead to the identification of the optimized machine learning algorithm for classification of cardiac arrhythmia to distinguish the patient with arrhythmia. Kernelized SVM has been identified as the most accurate model.**

Keywords — Cardiac arrhythmia; electrocardiogram; ECG signal; machine learning

I. INTRODUCTION

The word "arrhythmia" means a change in the normal sequence of electrical impulses. The electrical impulses can be too fast, too slow, or erratic – causing the heart to beat irregularly. When the heart doesn't beat properly (irregular heartbeats), it can't pump blood effectively throughout the body. When this happens, organs like lungs, brain, etc can't work properly and may shut down or be damaged permanently.

An ECG (Electrocardiogram) is the medical test conducted to measure electrical activity or pulses of the heart. Each heartbeat causes an electrical impulse that travels through the heart. This electric impulse causes the heart muscle to squeeze and pump blood throughout the body. A normal heartbeat on the ECG result shows the timing of the top and lower chambers of the heart.

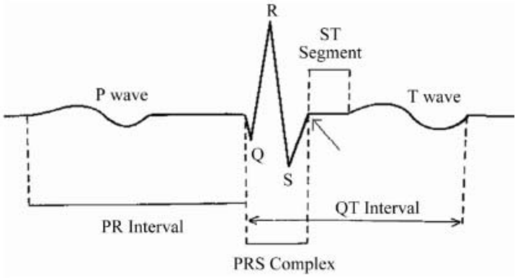


Fig 1.1: Diagrammatic representation of the basic electrocardiography deflections.

There are three main components of an ECG reading: the P wave that is the representation of the depolarization of the atria; the QRS complex representing the ventricle's depolarization; and the T wave, which is the representation of the repolarization of the ventricles.

To treat a patient effectively for arrhythmia, early diagnosis of this heart problem is essential. Good classification and performances can be achieved by utilizing computer-assisted methods and techniques. However, many require long computation times, complex classification mechanisms, and a lot of computational power. To add to this further, experts find it strenuous to study the long-duration ECG recordings and find the minute irregularities in them.

Different parameter values that are necessary can be extracted from the ECG signals and be used along with other information about the patient like age, gender, medical history, etc to detect and classify cardiac arrhythmia using a computer-based system. This research work is experiment with various supervised machine learning algorithms like Random Forest Classifier and Support Vector Machine for the detection and classification of arrhythmia.

II. LITERATURE SURVEY

In [1], the authors use the dataset available in UCI's Machine Learning Repository. Data preprocessing and Feature Selection is applied to avoid overfitting and find the important features in the dataset. They have implemented Naive Bayes binomial and multinomial Classifiers. In [2] the authors present an ECG classification strategy of a feature reduction method combining a probabilistic neural network classifier and PCA with LDA which differentiates against the eight types of arrhythmia. In [3], An algorithm for classification of arrhythmia is proposed by the author, in which the feature dimensions are reduced with the help of LDA i.e. linear discriminant analysis and SVM i.e. support vector machine-based classifier. The author also has done the procedure of cross-validation in which the comparison of SVM classifiers was done with Multilayer Perceptrons (MLP) and Fuzzy Inference System (FIS) classifiers. In [4], various machine learning methods were applied to classify arrhythmia, and evaluation of the most accurate learning methods was done. In [5], the author puts forward a productive automated ANN-based system that classifies Cardiac Arrhythmia from ECG signal data into multiple classes. A Neural network model with backpropagation has been proposed to classify the arrhythmia into abnormal and normal classes.

In [6], the authors qualitatively compare the following classifiers, SVM along with (K–A) training algorithm and MLP along with (BP) training algorithm. Feature extraction methods or data reduction methods are not employed by concerning the training performance, testing performance, and training time. In [7], the authors recommend the use of various trends to record the ECG, such as off the person, to get better accuracy in detection. In [8], the authors inferred that the most suitable combination for identifying various types of cardiac dysrhythmia from ECG signals is using the method of feature extraction based on the Naive–Bayes classifier and Higher-Order Statistics.

In [9], the authors have developed a comprehensive diagnostic system using which various types of cardiovascular diseases are identified with the help of deep learning methods. The algorithms showed a good amount of accuracy and successfully detected all disease states in each ECG signal by using MLP and CNN algorithms. In [10], For the VEB class, the method proposed by the authors shows great classification results on the AHA and MIT-BIH AR databases, thus outperforming the existing algorithms related to single lead classification in the detection of ventricular arrhythmia. The ESN approach presented is suited to process long-term recordings and large databases since the feature extraction and the algorithms have very few computational requirements.

In [11], a unique ECG recognition system is developed which is based on multi-domain feature extraction using Kernel-independent component analysis (KICA). It is also proposed that combining it with a discrete wavelet transform (DWT), it can be used to classify five types of ECG heartbeat reading. A new refined threshold wavelet method for ECG preprocessing is introduced to eliminate the noise influence. In [12], the classification performance of 3 algorithms is compared and the author concludes that the SCG method gives better accuracy than the other 2 algorithms. In [13], Using the MIT-BIH database, the authors propose a novel classification method based on a bijective soft set for ECG signal classification. This proposed methodology consists of 3 modules which are: signal acquisition, feature extraction, and classifier.

In [14], a new method is proposed which is based on the fractal dimension of the ECG signal. Concerning the chaotic system of the heart, it represents the electrical activity of the heart. In complex signals, minor changes can be examined by fractal dimension. In [15], the authors have designed a system where raw ECG data is given as input. The data are first preprocessed and later feature extraction is performed to obtain the output from the softmax layer. The output analysis and comparison is done for different activation functions.

III. SYSTEM DESIGN FRAMEWORK

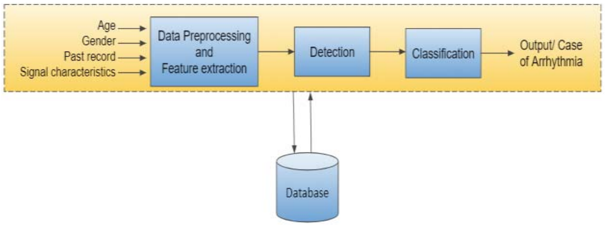


Fig 3.1: System block diagram

A. Data Collection

The dataset used is available publicly on UCI's Repository of Machine Learning. It contains 452 instances with 279 attributes each. Every instance contains the class of arrhythmia. The dataset consists of 16 classes and these classes denote the arrhythmia type. Class 01 represents the 'normal' class of ECG i.e. normal arrhythmia, classes 02 to 15 represent the various other classes of arrhythmia and class 16 represents the rest of unclassified classes of arrhythmia.

B. Data Preprocessing and Dataset Preparation

A CSV file Dataset with 279 attributes from the UCI's machine learning repository is used for the classification of arrhythmia into various classes. Data preprocessing techniques are applied to this dataset since data is unclean and consists of missing values for many of the instances. Thus, the dataset is first cleaned to remove all the missing values. The median value of the data is used for replacing missing values. In cases where there are a large number of missing values, the attribute is not considered for training. The dataset is divided in such a way that 339 instances are used for training purposes and 113 instances for testing purposes. The complexity of the dataset is extremely high due to the number of attributes per instance. To reduce the complexity and computation time, key features of the dataset are identified. For this purpose, Data Mining techniques such as feature extraction and feature selection are used. The proposed work has made use of principal component analysis (PCA) and SelectKBest as feature extraction and feature selection techniques respectively to identify those features which have a major impact on the output.

Principal Component Analysis also termed as PCA has been used as a feature extraction technique out of all the other available techniques. PCA finds out a set of new dimensions from the dataset such that they have very high variance or the data in those dimensions is more spread out. The variance in PCA tells us about how the data is spread out and is computed as follows:

$$Var(x)=\frac{\sum (xi-\bar{x})^2}{N}$$

Here, xi is the value of x in the ith dimension, and x bar is the median value of the entire dimension.

SelectKBest technique has been used as a feature selection technique with the help of which the features are selected in the data that contribute the most to the target variable. It is a dimensionality reduction technique that selects the k best dimensions from the dataset. With the combination of both of these techniques, the problem of overfitting can be overcome and achieved an improved accuracy with a reduced training time.

C. Detection and Classification of Arrhythmia

After applying the preprocessing and data mining techniques on the dataset, it is used to train and test the model to predict the class of arrhythmia out of the various classes taken into consideration.

Feature selection and extraction techniques are used to extract those features which have a major impact on the output given.

In [1], the author has used various algorithms and has done a comparative analysis of all of them concerning accuracy. Hence, concerning that, the support vector machine has been used for the classification of arrhythmia.

Support Vector Machine also called SVM is a supervised Machine Learning Algorithm which is primarily used for classification problems. In this algorithm, each data item is taken into consideration and plotted as a point in n-dimensional space with n being in the range of 0 to the number of features available. A hyperplane has been detected that best will differentiate the classes. The sole reason behind using SVM for classification is its high accuracy which is as shown in Fig.4.3.

It can also be observed that in [1], the author has directly used the classification algorithms on the dataset without making use of the data mining activities. In the current scenario, with 279 attributes it becomes complex to classify the instance precisely. To reduce this complexity of data, feature extraction and selection methods should be used. Hence, a slight improvisation can be made to the idea proposed by the author in [1] by applying the aforementioned techniques on the dataset before using it for training the SVM.

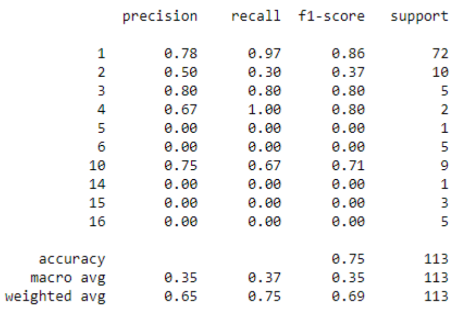
Initially, Linear SVM along with applying feature extraction and feature selection techniques on the dataset was used to classify Arrhythmia into its various classes and the corresponding accuracy recorded is as shown in Fig.4.3.

In the next step, Kernelized SVM was used for classification. In kernelized SVM, a kernel function is used which can convert any non-separable problem into a separable one to help achieve improved accuracy. There are both linear and polynomial types of kernels and in our case, the "Radial Basis Function" kernel which is a polynomial kernel has been used.

The training and testing of the algorithms have been done in python with the help of a Jupyter Notebook.

IV. RESULTS AND OUTPUTS

A. Classification Output

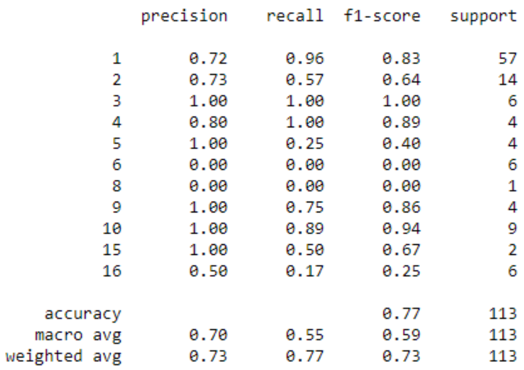


```
              precision    recall  f1-score   support

           1       0.78      0.97      0.86        72
           2       0.50      0.30      0.37        10
           3       0.80      0.80      0.80         5
           4       0.67      1.00      0.80         2
           5       0.00      0.00      0.00         1
           6       0.00      0.00      0.00         5
          10       0.75      0.67      0.71         9
          14       0.00      0.00      0.00         1
          15       0.00      0.00      0.00         3
          16       0.00      0.00      0.00         5

    accuracy                           0.75       113
   macro avg       0.35      0.37      0.35       113
weighted avg       0.65      0.75      0.69       113
```

Fig 4.1: Performance of Kernelized SVM with Feature Extraction (PCA)



```
              precision    recall  f1-score   support

           1       0.72      0.96      0.83        57
           2       0.73      0.57      0.64        14
           3       1.00      1.00      1.00         6
           4       0.80      1.00      0.89         4
           5       1.00      0.25      0.40         4
           6       0.00      0.00      0.00         6
           8       0.00      0.00      0.00         1
           9       1.00      0.75      0.86         4
          10       1.00      0.89      0.94         9
          15       1.00      0.50      0.67         2
          16       0.50      0.17      0.25         6

    accuracy                           0.77       113
   macro avg       0.70      0.55      0.59       113
weighted avg       0.73      0.77      0.73       113
```

Fig 4.2: Performance of Kernelized SVM with Feature Selection

The terms Precision, Recall, F1-score, and Support are used as evaluation measures to determine the accuracy of the model on test data. Figures 4.1 and 4.2 represent the classification report of the model. It denotes the precision, recall, f1-score, and support for each test case. Based on all the test cases, the macro average is the average of all test cases for precision and recall. The macro average of f1-score is the harmonic mean of the test cases. Weighted Average is the summation of (x1*w1) / w, where x1 stands for precision, recall, or f1-score of each test case; w1 stands for the respective support value of test case and w stands for the total support i.e. 113. The output of Kernelized SVM is shown above to understand the evaluation measures used. Similar classification reports are generated for other algorithms. The evaluation measures are explained in detail in the further sections.

B. Performance Comparison

Algorithm	**Accuracy**
Random Forest Classifier	67%
Random Forest Classifier with PCA	71%
Linear SVM	70%
Linear SVM with PCA	73%
Linear SVM with Feature Selection	65%
Kernelized SVM	71%
Kernelized SVM with PCA	75%
Kernelized SVM with Feature Selection (SelectKBest)	**77%**

Fig 4.3: Comparative analysis of different algorithms which can be used for classification

The algorithms used are supervised machine learning algorithms. Training data is provided to the algorithms which learn from it. The trained models are tested on the testing data which contains new instances similar to training data. Accuracy is determined by computing the number of instances correctly tested from the total number of training instances. An accuracy of 77% means that out of 100 samples, 77 data instances were classified correctly by the trained model.

C. Success Rate

Our system uses Kernelized SVM with Feature Selection and the success rate of the same is 77%. The algorithm used by our system provides the maximum accuracy compared to SVM and other algorithms used by the author in [1].

D. Evaluation Measures

The accuracy can be defined as the percentage of correctly classified instances.
Accuracy = (TP + TN) / (TP + TN + FP + FN)
where,
TP = Positive data instances classified correctly
FN = Negative data instances classified incorrectly
FP = Positive data instances classified incorrectly
TN = Negative data instances classified correctly
Precision means the number of positive class predictions that actually belong to the positive class.
Precision = TP / (TP + FP)

Recall means the number of positive class predictions made out of all positive examples in the dataset.

Recall = TP / (TP + FN)

F1-score is a measure used to determine the accuracy of classification models. It is computed to be the harmonic mean of Precision and Recall.

It is computed using the following formula:

F score=(2*Precision*Recall) / (Precision+Recall)

Support means the number of samples of the true response that lie in the class.

V. CONCLUSION AND FUTURE SCOPE

The proposed work has aimed to develop a fully automated and highly accurate system for the doctors to help them with multi-class Cardiac Arrhythmia classification from ECG signal data and other attributes. Our system will be user-friendly which will classify whether or not the person is suffering from arrhythmia. The system will take input in the form of ECG parameters which will then predict whether or not the patient is suffering from arrhythmia and if yes which class of arrhythmia will be given as an output. Also, along with developing this system, a combination of different algorithms will be used to determine the best available framework for the process and increase the test accuracy of this system. The algorithms can be given a bigger training dataset to increase efficiency and accuracy.

VI. REFERENCES

[1] Vasu Gupta, Sharan Srinivasan, Sneha S Kudli, “Prediction & Classification of Cardiac Arrhythmia” of Stanford University, 2013

[2] Jeen-Shing Wang, Wei-Chun Chiang, Ya-Ting C. Yang, Yu-Liang Hsu, “An Effective ECG Arrhythmia Classification Algorithm”

[3] Mi Hye Song, Jeon Lee, Sung Pil Cho, Kyoung Joung Lee, Sun Kook Yoo, “Support Vector Machine Based Arrhythmia Classification Using Reduced Features”

[4] Thara Soman, Patrick Bobbie, “Classification of Arrhythmia Using Machine Learning Techniques” of Southern Polytechnic State University (SPSU)

[5] Abhinav Vishwa, Mohit K. Lal, Sharad Dixit, Dr Pritish Varadwaj, “Classification Of Arrhythmic ECG Data Using Machine Learning Techniques”

[6] Majid Moavenian, Hamid Khorrami, “A qualitative comparison of Artificial Neural Networks and Support Vector Machines in ECG arrhythmias classification”

[7] Eduardo Joseda A.Luz, William Robson Schwartz, Guillermo Camara Chavez, David Menotti, “ECG-based Heartbeat Classification for Arrhythmia Detection: A Survey”

[8] Leandro B. Marinho, Navar de MM Nascimento, Joao Wekington M. Souza, Mateus Valentim Gurgel, Pedro P. Reboucas Filho, Victor Hugo C. de Albuquerque, “A novel electrocardiogram feature extraction approach for cardiac arrhythmia classification”

[9] Shalin Savalia, Vahid Emamian, “Cardiac Arrhythmia Classification by Multi-Layer Perceptron and Convolution Neural Networks”

[10] Miquel Alfaras, Miguel C. Soriano, Silvia Ortín, “A Fast Machine Learning Model for ECG-Based Heart Rate Classification And Arrhythmia Detection”

[11] Hongqiang Li, Danyang Yuan, Youxi Wang 1, Dianyin Cui, Lu Cao, “Arrhythmia Classification Based on Multi-Domain Feature Extraction for an ECG Recognition System”

[12] Manoj Kumar Senapati, “Cardiac Arrhythmia Classification of ECG Signal using Morphology and Heart rate signal”

[13] S. Udhaya Kumar and H. Hannah Inbarani, “Classification of ECG Cardiac Arrhythmias Using Bijective Soft Set

[14] Kourosh Kiani, Farzane Maghsoudi, “Classification of 7 Arrhythmias from ECG Using Fractal Dimensions”

[15] Rajkumar. A, Ganesan. M, Lavanya. R, “Arrhythmia classification on ECG using Deep Learning”