Heart Disease Prediction Using ECG Features Extraction and Classification

[1]Pranay Narendra Targe, [2] Prof. S. V. Patil,[3] Dr. K.S. Bhagat

[1]P. G. Student,[2] Associate Professor, [3]P.G. Coordinator

[1] Electronics and Telecommunication engineering Department, J. T. Mahajan College of engineering, Faizpur, India

[2] Electronics and Telecommunication engineering Department, J. T. Mahajan College of engineering, Faizpur, India

[3] Electronics and Telecommunication engineering Department, J. T. Mahajan College of engineering, Faizpur, India.

Abstract : Electrocardiogram (ECG) plays an imperative role in heart disease diagnostics, Arrhythmia is a cardiological disorder with prevalence of the world's population.The purpose of this research to discusses the electrocardiogram analysis because this problems concerning health issue which encourage the present research. The main objective of our research is to analyze the acquired ECG signals using signal processing tools such as SVM and KNN classifier, classify them. Total 187 ECG data subjects were analyzed. These data were grouped in two classes i.e, Normal class, and Arrhythmia class respectively. In order to achieve this we have applied a SVM and KNN classification technique. Features are extracted in two different groups. In first part various ECG features such as heart rate variability, mean RR interval, Root Mean Square Distance of Successive R-R interval, Number of R peaks in ECG that differ more than 50 millisecond, percentage NN50, Standard Deviation of R-R series, Standard Deviation of Heart Rate,sampleEntropy and Power Spectral Entropy were extracted. The second features set consist of wavelet based features consisting of power, spectrum, entropy and standard deviation for the positive alpha, positive beta, positive gamma, positive theta and positive delta wavelets separately. Then the extracted features data is analyzed and classified using KNN and SVM. The proposed algorithm is implemented and also tested in MATLAB software. The proposed system successfully classifies the Normal, Arrhythmia signal for all three datasets under consideration with the overall accuracy of 100%. The KNN classification with wavelet features gives the best performance.

*IndexTerms - **ECG, Arrhythmia, SVM, KNN, Wavelet Features, Accuracy**.*

I. INTRODUCTION

Monitoring Heart activity through the electrocardiogram (ECG) has become an important tool in the diagnosis of arrhythmia. Electrocardiogram (ECG) is the electrical manifestation of the contractile activity of the heart that can be recorded fast and automatically. It is a noninvasive diagnostic tool, meaning that ECG signal can be measured without entering the body at all. A natural electrical system causes the heart muscle to contract and pump blood through the heart to the lungs and the rest of the body. An ECG translates the heart electrical activity into line tracings on paper which shown in Figure 1. The spikes and dips in the line tracings are called waves which are P wave, QRS complex, T wave and PR and ST segment

***P wave**:* When the electrical impulse is conducted from the SA node towards the AV node and spreads from right to left atrium, the depolarization (contraction) of the atria occurs. The depolarization of atria results the P Wave in the ECG [1].

T wave: Ventricular repolarization results the preceding of ST segment and the T wave.

U wave: The origin of U wave is not clear and it is rarely seen. It is probably produced due to the repolarisation of the papillary muscles [2].

The PR interval: The PR interval begins with the onset of the P wave (Pi) and ends at the onset of the Q wave (Qi). It represents the duration of the conduction through the atria to the ventricles. Normal measurement for PR interval is 120ms-200ms.

***QRS complex**:* The QRS complex consists of three waves, sequentially known as Q, R and S. The rapid depolarization of both the ventricles results this complex. The muscles of the ventricles have large muscle mass than that of atria, hence its amplitude is much larger than that of P wave.

The ST segment: The ST segment represents the time between the ventricular depolarization and the repolarisation. The ST segment begins at the end of the QRS complex (called J point) and ends at the beginning of the T wave. Normally, the ST segment measures 0.12 second or less. The precise end of depolarization (S) is difficult to determine as some of the ventricular cells are beginning to repolarise.

The QT interval: The QT interval begins at the onset of the Q wave (Qi) and ends at the endpoint of the T wave (Tt), representing the duration of the ventricular depolarization/repolarisation cycle. The normal QT interval measures about 0.38 second, and varies in males and females and with age. As a general rule, the QT interval should be about 40 percent of the measured *R-R interval.*

The shapes of the ECG waveforms of different persons are different, so the differences of the waveform can be used to identify the different individual's characteristic. The ECG signal can varies from person to person due to the differences in position, size anatomy of the heart, age, sex relative body weight and also chest configuration.

The shape of ECG conveys very important hidden information in its structure. The amplitude and duration of each wave in ECG signals are often used for the manual analysis. Thus, the volume of the data being enormous and the manual analysis is tedious and very time-consuming task. Naturally, the possibility of the analyst missing vital information is high. Therefore, medical diagnostics can be performed using computer-based analysis and classification techniques [3]. In the case of arrhythmia this rhythm becomes irregular, that is either too slow or too fast. Several algorithms have been proposed to classify ECG heartbeat patterns based on the features extracted from the ECG signals to increase the accuracy and sensitivity.

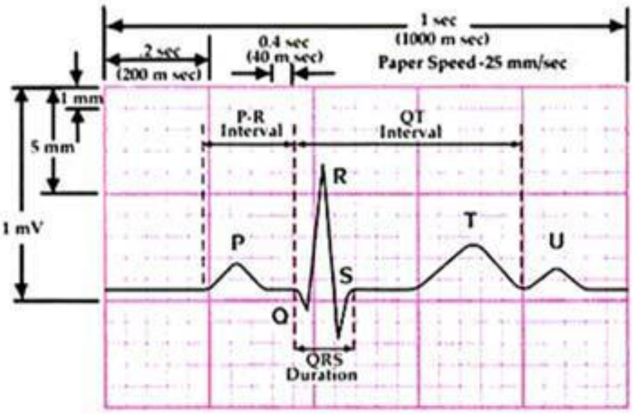

Figure 1 Normal ECG Waveform

II. DATABASE

The Physiological ECG signal use of this work provided by the MIT BIH arrhythmia database [2]. Three types of arrhythmiyas database are considered for classification of arrhythmiya.

- MIT-BIH ECG Data
- CU Ventricular Tachyarrhythmia Database
- Supraventricular Arrhythmia Database

The MIT BIH arrhythmias data base includes 48 ECG recording; these recordings are half an hour long. Signals are sampled 250 Hz. In the database each signal has an annoted file that includes the beat, rhythm and other information. This database is used by researchers to test their algorithms for detection of arrhythmias and classifications. In present research work we use of full length to classify into normal or abnormal class, abnormal for arrhythmia ECG signal. The online PhysioNet library is used for converting the MIT BIH data into matlab readable format.

III. RELATED WORK

In [7], authors proposed an automated arrhythmia classification system using different feature of ECG signals. Authors discussed the classification result of SVM shows the ECG signal feature can be used as a reliable indication of cardiac proble ms. Generally results are not achieved with 100 percent accuracy. The accuracy of the suggested system depends upon several factors like quality and size of training data and also used parameters. However the present method shows the result with 98 percent accuracy. The paper [8] presents a suitable and efficient implementation of a feature extraction algorithm (Pan Tompkins algorithm) on electrocardiography (ECG) signals, for detection and classification of four cardiac diseases: Sleep Apnea, Arrhythmia, Supraventricular Arrhythmia and Long Term Atrial Fibrillation (AF) and differentiating them from the normal heart beat by using pan Tompkins RR detection followed by feature extraction for classification purpose .The paper also presents a new approach towards signal classification using the existing neural networks classifiers.

Electrocardiogram (ECG) plays an imperative role in heart disease diagnostics, Arrhythmia is a cardiological disorder with prevalence of the world's populati The purpose of [9] to discusses the electrocardiogram analysis because this problems concerning health issue which encourage the present research. The main objective of [9] is to analyze the acquired ECG signals using signal processing tools such as wavelet transform and Neuro classifier, classify them.. Total 62 ECG data subjects were analyzed. These data were grouped in two classes i.e, Normal class, and Arrhythmia class respectively. In order to achieve this authors have applied a back-propagation based neural network classifier. DWT coefficients are used to extract the relevant information(statistical feature) from the ECG input data which are Energy, Variance, power Spectrul Density, Mean and Standard Deviation and morphological feature from the ECG input data which are mean and standard deviation of Magnitude of P,Q,R,S,T peak & PQ, QR,RS,ST,PR,QRS interval & RR interval. Then the extracted features data is analyzed and classified using Adaptive Neuro System (ANNS) as a Neuro classifier. The proposed algorithm is implemented and also tested in MATLAB software. The ECG signal are being selected and tested from PhysioNet Database using MIT-BIH Arrhythmia Database, and & Normal Sinus Rhythm (NSR) database. 14 subjects from Normal set and 48 subjects from arrhythmia set were analyzed for feature extraction and classification and data were divided in training, testing and validation of proposed algorithm.

In [10] authors have proved that using an ANN is more accurate than taking the FFT of an ECG signal for this case. ANN's are a great way to solve some common things in this world just by using some artificial intelligence and training it for the right answers. By analyzing the data points in an ECG signal, many characteristics of the heart can be determined and analyzed. ANN's have much potential especially for cardiac arrhythmias.

IV. METHODOLOGY

The method for arrhythmia detection is implemented in MATLAB 2013a. The flowchart of the proposed methodology for detection of arrhythmia data from normal data is shown in figure 2. From the figure the whole methodology is divided into three basic parts: QRST wave detection, feature extraction and classification. It can be seen that the raw ECG signal is offered for detection of QRST waves. The original ECG signal should be pre-processed with the purpose of detecting Q R S T peaks of ECG and preparing this processed signal for the next stage. The next stage of the proposed model is feature extraction that is preparing the input which best characterize the original signal. Final step of the method is to classify the processed signal into the normal and arrhythmia class.

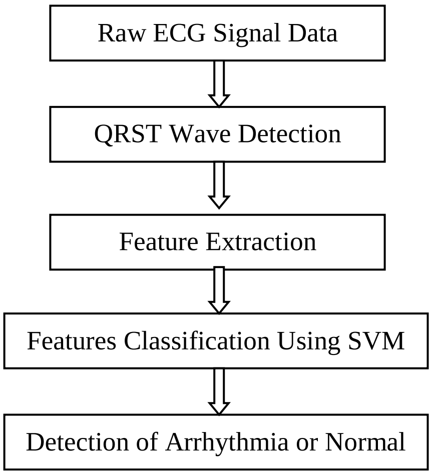

Figure 2 Steps of Detection of Arrhythmia Using ECG

QRST wave detection

The process of QRST wave detection uses the state-machine logic to determine different peaks in an ECG signal. It has the ability to confront noise by canceling out the noise by high pass filtering and baseline wander by low pass. Besides, check out criterion to stop detection of spikes.

The steps for QRST wave detection are as follows

1. Noise cancelation (Filtering)
   - cuttoff low frequency to get rid of baseline wander
   - cuttoff frequency to discard high frequency noise
   - cutt off based on fs
   - bandpass filtering
2. Enter state 1(putative R wave)
3. Locate R by finding highest Peak
4. Check if Sig drops below the threshold to look for S wave
5. Enter S wave detection state3 (S detection)
6. Enter state 4 possible T wave detection
   - See if the signal drops below mean
7. Set a double threshold based on the last detected S wave and baseline of the signal and look for T wave in between these two threshold

Features Extraction

Features extraction is done for two sets of features. In first set we have computed the features from the QRST wave.

Various features used for classification are

- Heart Rate Variability
- Mean RR Interval
- Root Mean Square Distance Of Successive R-R Interval
- Number Of R Peaks in ECG that Differ More than 50 Millisecond,
- Percentage NN50,
- Standard Deviation Of R-R Series,
- Standard Deviation Of Heart Rate,
- Sample entropy And
- Power Spectral Entropy

In the second set of features, we have considered the wavelets features for classification.

We have used

- Energy
- Spectral
- Power
- Standard deviation

The features are considered for positive Alpha, Beta, Delta, Gamma and Theta wavelets.

Support Vector Machines

In machine learning, support-vector machines (SVMs, also support-vector network) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.
When data is unlabelled, supervised learning is not possible, and an unsupervised learning approach is required, which attempts to find natural clustering of the data to groups, and then map new data to these formed groups.

KNN Classification

In pattern recognition, the k-nearest neighbors algorithm (k- NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression:
In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.
In k-NN regression, the output is the property value for the object. This value is the average of the values of k nearest neighbors.
k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all machine learning algorithms.
Both for classification and regression, a useful technique can be used to assign weight to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of 1/d, where d is the distance to the neighbor.
The neighbors are taken from a set of objects for which the class (for k-NN classification) or the object property value (for k-NN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required.
A peculiarity of the k-NN algorithm is that it is sensitive to the local structure of the data.

IV. RESULTS AND DISCUSSION

Using MATLAB R2013a the overall classification was done using SVM and KNN Classifier.
The overall samples are divided as-

i. Training Data-50 % of total dataset.
ii. Testing Data- 50 % of total dataset.

Figure 3 shows the sample QRST wave detection for ECG signal

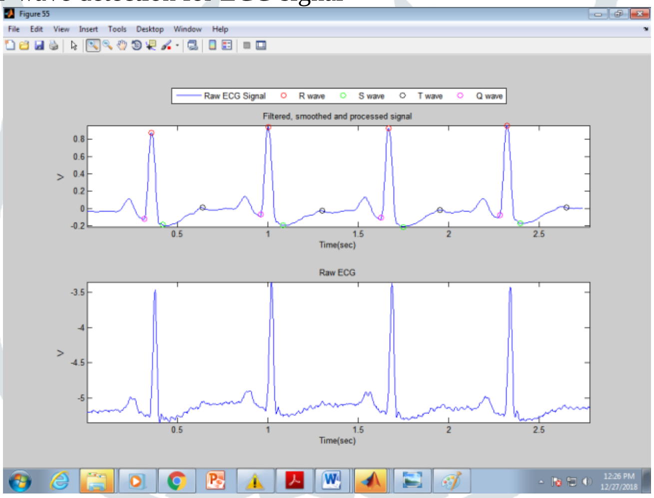

Figure 3 QRST wave detection

Table 1 depicts the features extracted for the sample ECG signals from the Arrhythmia database and table 1depicts the features extracted for the normal sinus database.

Table 1 depicts the classification performance for MIT-BIH Arrhythmiya Dataset with 55 ECG samples.

Table 3: Performance Evaluation for MIT-BIH Dataset

	Precision	Recall	Fmeasure	Accuracy
SVM	1	1	1	100
KNN	0.90	0.9783	0.9692	98.18
SVM with Wavelet Features	0.9778	0.9091	0.9714	98.18
KNN with Wavelet Features	1	1	1	100

The KNN classifier with wavelet features is able to classify the ECG signal from MIT BIH dataset with 100% accuracy.
Table 4 depicts the performance for CU venticular tachyarrhythmiya database (45 ECG Samples)

Table 4: Performance Evaluation for CU venticular tachyarrhythmiya database (45 ECG Samples)

	Precision	Recall	Fmeasure	Accuracy
SVM	1	1	1	100
KNN	0.90	0.9722	0.9677	97.77
SVM with Wavelet Features	1	1	1	100
KNN with Wavelet Features	1	1	1	100

The classification accuracy is 100% for SVM, SVM with wavelet features as well as KNN with wavelet features.

Table 6 depicts the performance for supraventricular arrhythmia database (87 ECG Signals). The SVM classifier with features set1 gives the 100% accuracy. The KNN classifier with wavelet features also able to classify the ECG signals with 100% accuracy.

Table 5: Performance Evaluation for supraventricular arrhythmia database (87 ECG Signals)

	Precision	Recall	Fmeasure	Accuracy
SVM	1	1	1	100
KNN	0.9610	0.7692	0.9301	96.55
SVM with Wavelet Features	0.8182	0.4167	0.7963	83.90
KNN with Wavelet Features	1	1	1	100

ECG File Name	Heart Rate Variability				Mean RR Interval	Root Mean Square Distance Of Successive R-R Interval	Number Of R Peaks in ECG	Percentage of Number Of R Peaks in ECG	Standard Deviation Of R-R Series	Standard Deviation Of Heart Rate
100	77	293	293	284	236.75	254.1078	2	50	47.25	471.6586231
101	314	315	138	97	216	237.8413	3	75	119	255.1968903
102	294	287	292	97	242.5	256.6505	3	75	145.5	48.0672392
103	253	311	301	97	240.5	255.3135	2	50	143.5	82.51497494
104	297	299	286	97	244.75	259.2369	3	75	147.75	47.6873036
105	197	261	250	256	241	242.3665	2	50	15	137.2287043
106	294	374	250	256	293.5	297.6357	2	50	37.5	139.6441355
107	245	297	316	118	244	255.9365	2	50	126	206.2146893
108	355	349	316	118	284.5	300.6684	3	75	166.5	12.61482727
109	107	235	228	235	194.4	200.884	3	60	27.4	353.5293944
111	197	292	314	235	241	247.2986	3	60	74	147.3337909
112	125	257	263	247	211.8	218.9708	3	60	44.8	240.5761165
113	412	384	263	247	294.6	308.3073	4	80	127.6	100.0530584

Table 1 Features Extraction for Sample ECG Signals from Normal Sinus database

ECG File Name	Heart Rate Variability				Mean RR Interval	Root Mean Square Distance Of Successive R-R Interval	Number Of R Peaks in ECG	Percentage of Number Of R Peaks in ECG	Standard Deviation Of R-R Series	Standard Deviation Of Heart Rate
16272	125	123	123	122	123.8571	123.8646	6	85.71429	2.142857	395.9738402

16420	81	79	82	82	82.63636	82.65536	10	90.90909	0.636364	1041.562967
16483	81	79	79	80	80.33333	80.33887	11	91.66667	0.666667	1141.345427
16773	106	107	100	103	98.25	98.74335	11	91.66667	17.25	608.4656085
16795	107	112	114	123	107.5	108.765	11	91.66667	26.5	479.8394454
17052	114	115	110	110	104.75	105.6862	11	91.66667	23.75	460.7478292
17453	93	89	90	93	91.58333	91.72468	11	91.66667	10.58333	802.8387669
18184	97	103	104	98	92.5	92.78829	11	91.66667	11.5	735.0602136
19088	128	128	128	128	110.9167	112.8107	11	91.66667	29.91667	304.6107332
19830	128	129	127	128	110.9167	112.8114	11	91.66667	29.91667	308.6419753

Table 2 Features Extraction for Sample ECG Signals from Arrithmiya database

V. CONCLUSON

In our research we have designed a SVM and KNN based expert system for classification of Normal and Arrhythmia ECG signals of 187 samples of data. The overall conclusion can be summarized in following points-:

- An Automated Arrhythmia Classification system is developed using Statistical and morphological features extraction as well as wavelet based features and SVM and KNN based classification tool
- Three types of arrhythmiyas database are considered for classification of arrhythmiya namely; MIT-BIH ECG Data, CU Ventricular Tachyarrhythmia Database, Supraventricular Arrhythmia Database.
- Two types of features set were analysed for performance. In first features set total 12 features for statistical and temporal features were selected to develop features input vector for classifier. In the second feature set wavelet based features like energy, power, spectral and standard deviation are selected.
- Classification process is carried out using SVM and KNN Classifier.
- Overall efficiency of 100 percent is achieved in the classification process. The KNN classification with wavelet based feature set gives the accuracy of 100% for almost all ECG datasets under consideration with 50% training data.

REFERENCES

[1] Http://En.Wikipedia.Org/Wiki/Electrocardiography

[2] M. Gabriel Khan. “Rapid ECG Interpretation.” 3rd Edition, Humana Press, New York, ISBN- 978-1-58829-979-6, 2008

[3] Maedeh Kiani Sarkaleh and Asadollah Shahbahrami “Classification of ECG Arrhythmias using discrete wavelet transform and neural networks” IJCSEA Vol.2, No.1, February 2012

[4] Http://www.physionet.org/physiobank/database/mitdb

[5] B.U. Kohler, C. Henning, and R. Orglmeister, “The principles of software QRS detection,” IEEE Eng. Med. Biol. Vol. 21, pp. 42–57, 2002

[6] T.Ince, S. Kiranyaz, and M. Gabbouj, “A generaric and robust system for automated patient-specific classification of ECG signals,” IEEE Trans. Biomed. Eng. vol. 56, pp. 1415-1426, 2009.

[7] Gorav Kumar Malik, Dr.Yatindra Kumar, Dr. Manoj Panda, “Cardiah arrhythmia detehtion in ECG signals by feature Extrahtion and support vehtor mahhine”, Second International Conference on Research in Intelligent and Computing in Engineering pp. 241–244 DOI: 10.15439/2017R63 ACSIS, Vol. 10 ISSN 2300-5963

[8] R karthik, Dhruv Tyagi, Amogh Raut, Soumya Saxena, Rajesh Kumar M, “Implementation of Neural Network and feature extraction to classify ECG signals”, 2017 International Conference on Mechanical,System and Control Engineering ©2017 IEEE

[9] Mangesh Singh Tomar,Manoj Kumar Bandil, D B V Singh, “ECG feature extraction and classification for Arrhythmia using wavelet & Scaled Conjugate-Back Propagation Neural Network”, International Journal of Advanced Research in Computer Science, Volume 4, No. 11, Nov-Dec 2013

[10] E. Roland Adams, Anthony Choi, “Using Neural Networksto Predict Cardiac Arrhythmias”, 2012 IEEE International Conference on Systems, Man, and Cybernetics October 14-17, 2012, COEX, Seoul, Korea