Classification of Arrhythmia

Saleha Samad, Shoab A. Khan, Anam Haq, and Amna Riaz
National University of Sciences and Technology, College of E&ME, Rawalpindi, Pakistan
Email: {salehasamad69, shoabak}@ce.ceme.nust.edu.pk, anam111@live.com, aminariaz_1988@hotmail.com

***Abstract*—The electrocardiogram (ECG) signal has great importance in diagnosing cardiac arrhythmias. In this paper we have compared three classifiers on the basis of their accuracies for the detection of arrhythmia. The algorithms that are used for classification are supervised machine learning algorithm. The performance of the classifier depends upon its accuracy rate. The classifiers used are Nearest Neighbors, Naive Bayes', and Decision Tree classifier. The dataset used is publically available on UCI Machine Learning Repository. The calculated accuracies by our classifier are 66.9645%, 59.7696%, and 45.8487% for k-NN, Descion Tree and Naïve Bayes' Classifier respectively. k-NN gives the maximum accuracy while the previously calculated accuracy of k-NN was 53%.**

***Index Terms*—accuracy, arrhythmia, descion tree classifier, k-NN classifier, na ïve bayes classifier**

I. Introduction

He various diseases from which human suffers are related to heart. Heart diseases are still the most problematic one to be known. With timely detection and proper medical treatment of these diseases we can save many lives. The "heart's peacemaker" is the node from which the electrical signal for heart beat are stimulated. These electrical signals are actually originated from the Sino Atrial node which is present at the top of the hearts right chamber known as Atrium [1].

If there is any sort of disruption in the peacemaker, the heartbeat will become abnormal which directly affects the flow of blood in the body. The abnormality is the heart beat indicates that the patient is suffering from arrhythmia. It is diagnosed using an electrocardiogram procedure. An ECG wave can be divided into three types of wave called as P wave, QRS complex and T wave. A classic ECG signal is shown in Fig. 1.

Heart patients are inspected through the parameters that include the QRs duration, RR, PR and QT intervals with some other information like sex, age, weight and the decision of the cardiologist. The irregular beat phases are called as arrhythmia and some type of arrhythmias are very serious for the patient.

The ECG of a patient provide two main kinds of information; one is the measuring of the time interval in the ECG which facilitates in finding out the time period of the electrical signal passing through the heart, the result of this is that we can easily find out the rate of electrical activity i.e. it's irregular, slow or fast. Second if we find out the total electrical activity that is passing through the heart muscle it helps out to determine the parts of heart that are excessively large or overburdened. Physician interprets ECG signal and determines whether his/her heart beat is normal or belongs to the class of arrhythmia.

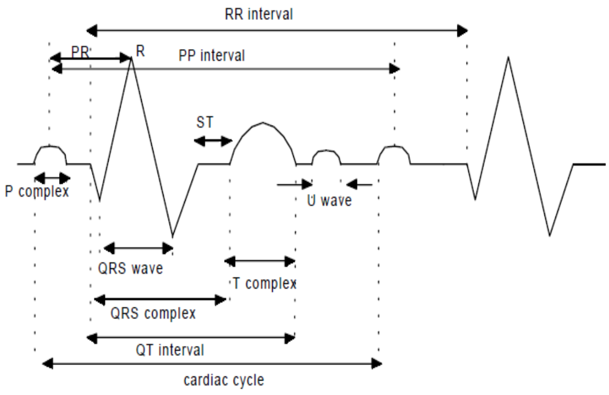

Figure 1. A cardiac cycle of an ECG.

A lot of research is done in the past few years regarding the classification of arrhythmia. HA Guvenir created a new algorithm which is called as VF15 to classify arrhythmia. He shows that this new algorithm performs very well as compared to basic classification algorithm [2].

NarendaKholi in his paper applied support vector machine classifier to detect arrhythmia [3].

MiHye Song and Jeon Lee in their journal applied linear discriminate analysis to reduce the feature set and uses SVM and wavelet transform to classify arrhythmia [4].

Cayonggio, Michael Madden and Des Chamber in their paper uses Bayesian Artificial Neural Network classifier to detect arrhythmia [5].

There is large volume of research work done in this field and still there is a possibility for improvement. In this paper we have compared the three classification algorithm i.e. knn, Na ïve Bayes' and Decision tree to determine whether the patient is suffering from arrhythmia or not.

The feature set is picked from the website http://archive.ics.uci.edu/ml/datasets/Arrhythmia.

In the Section II the Attributes and classes of the dataset are explained. In the Section III system model and the classifiers used in that model are explained. In the Section IV the results are showed. In the Section V summarizes the conclusion.

Manuscript received June 29, 2013; revised September 10, 2013.

doi: 10.12720/ijoee.2.1.57-61

II. Data-Set Description

The basic purpose is to differentiate the presence and absence of arrhythmia. The class of arrhythmia is further divided into 15 different types. In the dataset contains the record of 452 patients and 279 feature values. The classes and the type of arrhythmia are described in the Table I.

TABLE I. Description of Types of Arrhythmia

CLASS	DESCRIPTION
1	NORMAL ECG
2	ARRHYTHMIA
Arrhythmia can be any of the following types	
1	ISCHEMIC CHANGES
2	OLD FRONTAL MYOCARDIAL
3	OLD LOWER MYOCARDIAL
4	SINUS TACHYCARDIA
5	SINUS BRADYCARDY
6	VENTRICULAR PREMATURE CONTRACTION
7	SUPRARENTRICULAR PREMATURE CONTRACTION
8	LEFT BUNDLE BRANCH BLOCK
9	LEFT BUNDLE BRANCH BLOCK
10	1st DEGREE ATRIO VENTRICULAR BLOCK
11	2nd DEGREE ATRIO VENTRICULAR BLOCK
12	3 DEGREE ATRIO VENTRICULAR BLOCK
13	LEFT VERTICLE HYPERTROPHY
14	ATRIAL FIBRILLATION
15	REST

There are total 279 features, and the description of each feature is mentioned in the Table II.

TABLE II. Description of Features

Features	Description
f_1	Age
f_2	Gender
f_3	Height
f_4	Weight
f_5	QRS interval average (msec)
f_6	The normal interval between start of P and Q waves(msec)
f_7	The standard interval between onset of Q and offset of T waves (msec)
f_8	The normal distance between two consecutive T waves (msec)
f_9	The average distance between two consecutive P waves (msec)
f_{10}	The vector angle in degree on the front plane of QRS
f_{11}	The vector angle in degree on the front plane of T
f_{12}	The vector angle in degree on the front plane of P
f_{13}	The vector angle in degree on the front plane of QRST
f_{14}	The vector angle in degree on the front plane of J
f_{15}	Heart Beats per minute
f_{16}	Average width of Q waves in msec
f_{17}	Average width of R waves in msec
f_{18}	Average width of S waves in msec
f_{19}	Average width of R' waves in msec
f_{20}	Average width of S' waves in msec
f_{21}	No. of intrinsic deviation
f_{22}	Presence of dysphasic R waves (Boolean)
f_{23}	Presence of notched R waves (Boolean)
f_{24}	Presence of notched P waves(Boolean)
f_{25}	Presence of dysphasic P waves(Boolean)
f_{26}	Presence of notched T waves(Boolean)
f_{27}	Presence of dysphasic T waves (Boolean)
f_{28}—f_{39}	DII
f_{40}—f_{51}	DIII
f_{52}—f_{63}	AVR
f_{64}—f_{75}	AVL
f_{76}—f_{87}	AVF
f_{88}—f_{99}	V1
f_{100}—f_{111}	V2
f_{112}—f_{123}	V3
f_{124}—f_{135}	V4
f_{136}—f_{147}	V5
f_{148}—f_{159}	V6
f_{160}	J point depression(millivolts)
f_{161}	Amplitude of Q wave(millivolts)
f_{162}	Amplitude of R wave(millivolts)
f_{163}	Amplitude of S wave(millivolts)
f_{164}	Amplitude of R' wave(millivolts)
f_{165}	Amplitude of S' wave(millivolts)
f_{166}	Amplitude of P wave(millivolts)
f_{167}	Amplitude of T wave(millivolts)
f_{168}	QRSA (Total of the areas of all segments divided by 10)
f_{169}	QRSTA(equal to QRSA+0.5 x width of T wave x 0.1 x Height of T wave.)
f_{170}—f_{179}	DII
f_{180}—f_{189}	DIII
f_{190}—f_{199}	AVR
f_{200}—f_{209}	AVL
f_{210}—f_{219}	AVF
f_{220}—f_{229}	V1
f_{230}—f_{239}	V2
f_{240}—f_{249}	V3
f_{250}—f_{259}	V4
f_{260}—f_{269}	V5
f_{270}—f_{279}	V6

III. System Model

The system model used for the classification uses three different classifiers that are: i) k-Nearest Neighbor classifier, ii) Naïve Bayes' Classifier, and iii) Decision tree classifier.

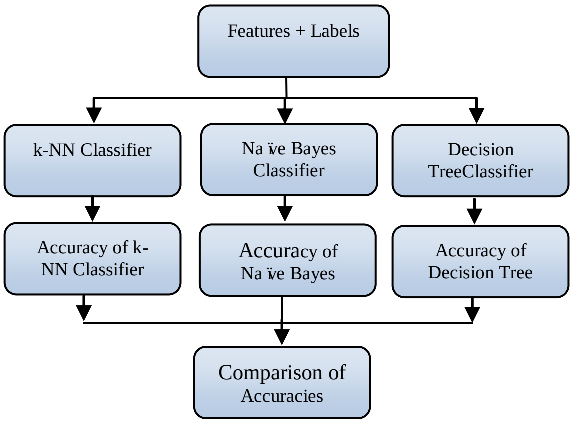

Figure 2. System model.

The sytem model for classification is shown in Fig. 2. The feature set and labels are passed through each of the classifiers. Then the accuracy of each classifier is calculated and the result is evaluated on the basis of accuracy i.e. we will rely on the result of the classifiers which gives maximum accuracy. The short detail of each classifier used in our model is given:

A. K-Nearest Neighbor Classifier

The *k*-nearest neighbor algorithm (*k*-NN) technique is used to label objects based on minimum distance criteria from the training samples that are present in feature vector. This in all machine learning algorithms is the simplest algorithm. The most common class amongst the *k*-nearest neighbors is labeled to the object, while*k*is a positive integer.The neighbors are selected from the object sets whose class label is already known. This will be the training data set for the algorithm. The training samples are vectors in a multidimensional feature space, each with a class label. The feature vectors and the class information are stored in the training session of the algorithm [6].

In testing part consists of the following steps:

- Measure Euclidian distance to each training point.
- Look for the k nearest points.
- Identify the most common class among those k closet points.
- Assign that class.

The choice of “k” is determined through the data. We have to set k according to our requirements e.g. if noise is present in the data, larger value of k is selected which will minimize the effect of noise on the classification. But because of high value of “k” the boundaries between classes become less distinct. If the Value of k is set 1 this is the special case in which the classifier is called as nearest neighbor classifier. If the features are not consistent with their significance the accuracy of the k-NN algorithm can be sternly despoiled by the occurrence of noisy or unrelated features [7].

B. Naive Bayes’ Classifier

Naive Bayes’ classifier works on the principle that absence or presence of a particular feature does not depend upon the absence or presence of other feature given the class label. Various studies have shown that Naive Bayes’ classifier is competitive with other advance classifiers [8].

Naive Bayes’ classifier follows the conditional model which is:

$$\mathrm{P(\,C|F_1, \dots .\, F_n)}$$

where C is the dependent class variable, the problem with it is that when there is huge data or when feature vector is too large, the use of conditional model is not feasible [1]. So we formulate the model to make it more efficient by using Bayes’ theorem, i.e.

$$P(\,C|F_1, \dots .\, F_n) = \frac{p(C)p(F_1|C)p(F_2|C), \dots ., p(F_n|C)}{Evidence}$$

$$\begin{aligned} Evidence = {} & p(C_1)p(F_1|C_1)..p(F_n|C_1) \\ & + p(C_2)p(F_1|C2)..p(F_n|C_2) \\ & + \, p(C_n)\, p(F_1|C_n) \dots p(F_n|C_n). \end{aligned}$$

$$Posterior\ Probability = \frac{(Prior\ Probability\ x\ Class\ Conditional\ Probability)}{Evidence}$$

In training of this classifier the mean and variance of each attribute is find. These values are then used in the testing procedure. In testing procedure posterior probability of that testing sample is found against each class. The test sample belongs to that class whose posterior probability is greater [9].

C. Decision Tree

Decision tree learning is a scheme for learning functions in which discrete valued target functions are approximated, by using a decision tree [10]. In a decision tree each branch node symbolizes a selection between number of options, and each leaf node characterizes a conclusion. Samples are classified through Decision trees by moving across root node to leaf node. Starting from root node, attributes are tested and specified by thisnode then according to the attribute value moving down the tree branch in the given set. This process is then repeated at the sub-tree level.ID3 is an easy decision tree learning algorithm developed by Ross Quinlan in 1983. The vital plan of ID3 algorithm is to employing a top-down search to construct the decision tree to test each attribute through the given sets at every tree node. That is most useful for classifying a given sets in order to select the attribute [11]. In decision tree at each node the entropy and gain is calculated. In training the tree is created by starting from the root node, the condition giving us the maximum entropy is set as a root node. From the root node two others nodes are created. The node satisfying the condition is put in the left side ($\mathrm{N_L}$) and the other one in the right side ($\mathrm{N_R}$). The tree is created until all the data is purified.The formulas to find out the entropy and the gain are given:

$$Entropy\ (i) = \, -(p_{\oplus} \log_2 p_{\oplus} + p_{\ominus} \log_2 p_{\ominus})$$

where $\mathrm{p_{\oplus}}$ is the probability of “i “belonging to class ⊕ and $\mathrm{p_{\ominus}}$ is the probability of “i “belonging to class⊖.

$$Gain\ (\triangle\ i) = i - P_L i(N_L\,) - \ P_R i(N_R\,)$$

here “i” is the total entropy, $\mathrm{P_L P_R}$ are the probability of left and right node and $\mathrm{i(N_L\,)i(N_R\,)}$ are the entropies of left and right nodes.

Cases to be considered:

Case 1: if the data is such that the prior probabilities of the classes are equal then in this case the entropy is maximum (the data highly impure).

Case 2: if the data is such that all the samples belongs to only one class in this case the entropy turns out to be zero meaning that the data is already pure.

IV. Results and Graphs

Following (Table III) are the results and graphs that are obtained in our simulation:

TABLE III. ACCURACY VALUES OF DIFFERENT CLASSIFIERS IN ARRHYTHMIA DETECTION

Classifiers	Accuracy
knn	66.9645%
Decision Tree	59.7696%
Naive Bayes'	45.8487%

Following are the accuracy graphs of the three Classifiers:

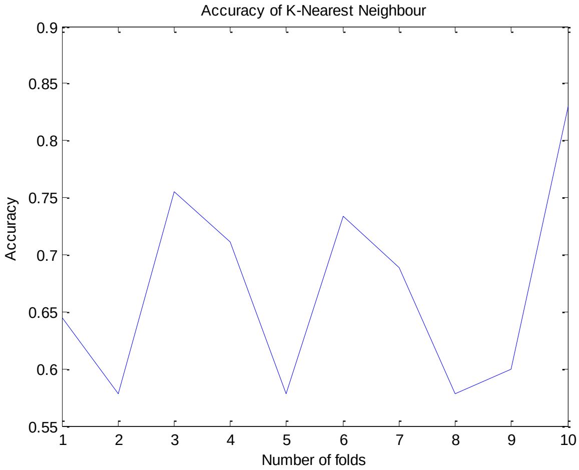

Figure 3. Accuracy curve of K-Nearest neighbor.

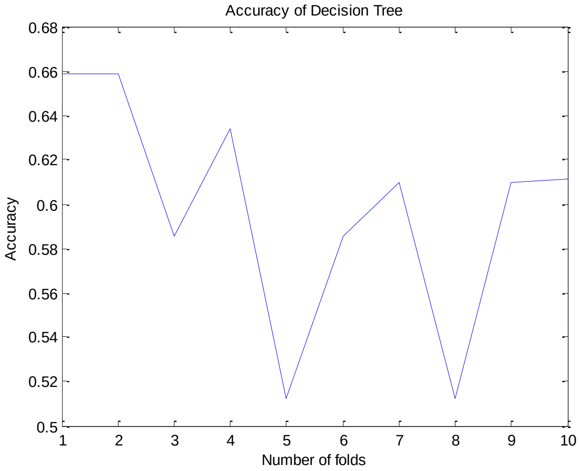

Figure 4. Accuracy curve of decision tree.

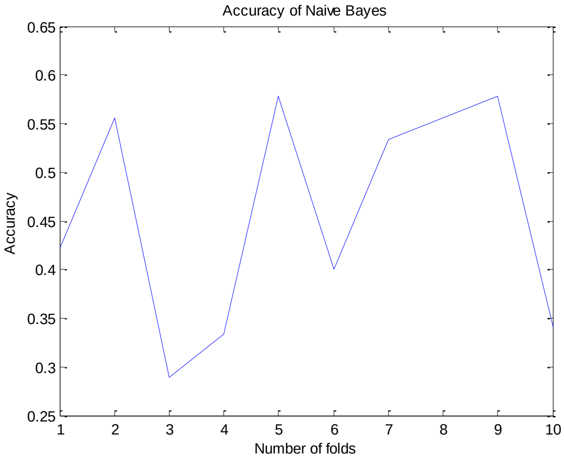

Figure 5. Accuracy curve of naive bayes.

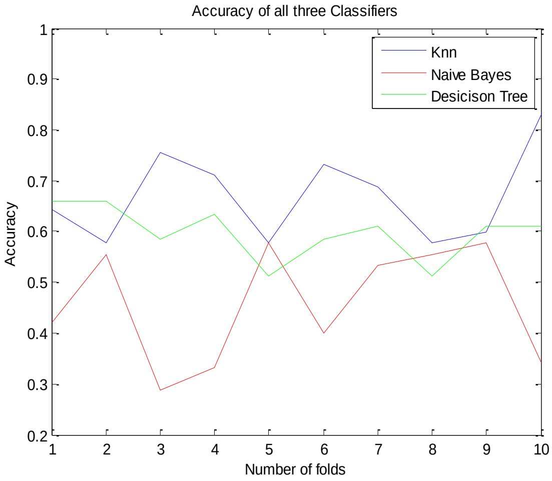

Figure 6. Comparison of accuracies of the classifiers.

V. CONCLUSION

From the results shown above it is clear that the k-Nearest Neighbor gives us the maximum accuracy in detecting arrhythmia. After k-NN decision tree gives us the best accuracy whereas the Naive Bayes' accuracy is the lowest of all three. The previously calculated accuracy for k-NN was 53% which is increased to 66.9645%. In future more work can be done to improve the classification of arrhythmia like building a hybrid model that classifies on the basis of highest accuracy of different classifiers.

REFERENCES

[1] M. G. Tsipouras, Y. Goletsis, and D. I. Fotiadis, "A method for arrhythmic episode classification in ECGs using fuzzy logic and Markov models," in *Proc. IEEE conf. on Computers in Cardiology*, 2004, pp. 361-364.

[2] H. A. Guvenir, S. Acar, G. Demiroz, and A. Cekin, "A supervised machine learning algorithm for arrhythmia analysis," in *Proc. IEEE conf. on Computers in Cardiology*, 1997, pp. 433-436.

[3] N. Kohli, N. K. Verma, and A. Roy, "SVM based methods for arrhythmia classification in ECG," in *Proc. IEEE Interantional conf. on Computer and Communication Technology*, 2010, pp. 486-490.

[4] M. H Song, J. Lee, S. P. Cho, K. J Lee, and S. K. Yoo, "Support vector machine based arrhythmia classification using reduced features," *International Journal of Control, Automation, and Systems*, vol. 3, no. 4, pp. 571-579, December 2005.

[5] D. Gao, M. Madden, D. Chambers, and G. Lyons, "Bayesian ANN classifier for ECG arrhythmia diagnostic system: A comparison study," in *Proc. IEEE International Joint Conference on Neural Networks*, 2005, pp. 2383-2388.

[6] D. Bremner, E. Demaine, J. Erickson, J. Iacono, S. Langerman, P. Morin, and G. Toussaint, "Output-sensitive algorithms for computation of nearest-neighbor decision boundaries," *International Journal of Discrete and Computational Geometry*, vol. 33, pp. 593-604, April 2005.

[7] G. Toussaint, "Geometric proximity graphs for improving nearest neighbor methods in instance-based learning and data mining," *International Journal of Computational Geometry and Applications*, vol. 15, pp. 101–150, April 2005.

[8] T. Soman and P. O. Bobbie. Classification of arrhythmia using machine learning techniques. Available: http://cse.spsu.edu/pbobbie/SharedFile/ECGDiagnosis_ICOSSE_2005_VFinal.pdf

[9] M. Wiggins, A. Saad, B. Litt, and G. Vachtsevanos, "Evolving a bayesian classifier for ECG-based age classification in medical

applications," *Journal of Applied Soft Computing*, vol. 8, no. 1, pp. 599-608, 2008.

[10] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Journal of Machine Learning*, vol. 29, pp. 131-163, 1997.

[11] B. Anuradha and V. C. Veera Reddy, "ANN for classification of cardiac arrhythmias," *Journal of Engineering and Applied Sciences*, vol. 3, no. 3, pp. 1-6, June 2008.

SalehaSamad was born in Bahawalpur, Pakistan in 1989. She received her BS degree in computer engineering from Islamia University of Bahawalpur (IUB), in 2010. She is currently pursuing the MS degree in computer engineering with the department of Computer Engineering, NUST, College of Electrical & Mechanical Engineering, Rawalpindi Pakistan. Her research fields include pattern recognition and image processing. SalehaSamad is a registered engineer by Pakistan Engineering Council (PEC) under PEC registration number COMP/7835.

Dr. Shoab Ahmed Khan was born in Lahore, Pakistan in 1965. He has done is PhD in Digital Signal Processing from Georgia Institute of Technology, Atlanta, GA. Dr. Khan has more than 17 years of industrial experience in companies like Scientific Atlanta, Picture Tel, Cisco Systems, and Avaz Networks. He has also been a recipient of National Education Award 2001 in the category of "Outstanding Services to Science and Technology", NCR National Excellence Award in the category of IT Education.He has published 16 ISI indexed Journal publications and more than 130 referred conference publications with 6 US patents have been awarded to his name. Currently, he is working as Head of Computer Engineering Department, EME College. His book on digital design titled "Digital Design of Signal Processing System" by John Willey & Sons has been released in Feb 2011. He has been awarded Tamgha-e-Imtiaz, the Presidential Award for his contribution in the field of Engineering.

Anum Haq was born in Islamabad, Pakistan in 1988. She received BS degree in computer engineering from Comsats institute of Information Technology Islamabad in 2010. She is currently pursuing the MS degree in computer engineering with the department of Computer Engineering, NUST, College of Electrical & Mechanical Engineering, Rawalpindi Pakistan. Her research fields include Aged related macular degeneration using Oct images, biomedical engineering.

Amina Riaz was born in Rawalpindi, Pakistan in 1988. She received her Bs degree in computer enginieering from Comsats institute of Information Technology Wah in 2010. She is currently pursuing the MS degree in computer engineering with the department of Computer Engineering, NUST, College of Electrical & Mechanical Engineering, Rawalpindi Pakistan. Her research fields include Traffic congestion recognition using motion vector statistical features.