International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 02 Issue: 05 | Aug-2015 www.irjet.net p-ISSN: 2395-0072

An Enhanced Approach on ECG Data Analysis using Improvised Genetic Algorithm

V.Priyadharshini[1], S.Saravana kumar[2]

ABSTRACT- *The analysis of the ECG can benefit in diagnosing most of the heart diseases. The electrocardiogram (ECG) provides almost all information about electrical activity of the heart. . The cardiac arrhythmias can be divided into number of categories according to the arrhythmia's mechanism of origin: irregular rhythm, escape, premature and tauchy - arrhythmias. Changes in the normal rhythm of a human heart may result in different cardiac arrhythmias, which may be immediately fatal or cause irreparable damage to the heart sustained over long periods of time. The ability to automatically identify arrhythmias from ECG recordings is important for clinical diagnosis and treatment. The basic objective is to come up with a simple method having less computational time without compromising with the efficiency. This paper proposes an enhanced technique for the arrhythmia classification and extraction of parameters from the ECG signal which is used for data acquisition and classification system. The problem of missing value is handled by average and imputation using boosted K-nn method. The preprocessed dataset is then classified using three different machine learning classification algorithms namely Naïve Bayes, C4.5 and genetic algorithm. The result shows that Genetic Algorithm outperforms the remaining two algorithms in analyzing ECG dataset*.

Keywords*: Electrocardiogram, Arrhythmia, Naive Bayes, C4.5 and Genetic Algorithm*

1. INTRODUCTION

Electrocardiogram (ECG) is a diagnosis tool that reported the electrical activity of heart recorded by skin electrode. The morphology and heart rate reflects the cardiac health of human heart beat. The cardiac arrhythmias can be divided into number of categories according to the arrhythmia's mechanism of origin: irregular rhythm, escape, premature and tauchy-arrhythmias. Changes in the normal rhythm of a human heart may result in different cardiac

Arrhythmias, which may be immediately fatal or cause irreparable damage to the heart sustained over long periods of time. The ability to automatically identify arrhythmias from ECG recordings is important for clinical diagnosis and treatment. The state of cardiac heart is generally reflected in the shape of ECG waveform and heart rate. ECG, if properly analyzed, can provide information regarding various diseases related to heart. However, ECG being a non-stationary signal, the irregularities may not be periodic and may not show up all the time, but would manifest at certain irregular intervals during the day. Clinical observation of ECG can hence take long hours and can be very tedious. Moreover, visual analysis cannot be relied upon and the possibility of the analyst missing the vital information is high. Hence, computer based analysis and classification of diseases can be very helpful in diagnosis. Various contributions have been made in literature regarding beat detection and classification of ECG signal. Most of them use either time or frequency domain representation of the ECG waveforms, on the basis of which many specific features are defined, allowing the recognition between the beats belonging to different classes.

The most difficult problem faced by today's automatic ECG analysis is the large variation in the morphologies of ECG waveforms. Moreover, it is necessary to consider the time constraints as well.

The basic objective is to come up with a simple method having less computational time without compromising with the efficiency. This objective has motivated me to search and experiment with various classification techniques

This paper focuses on some of the techniques proposed earlier for the arrhythmia classification and extraction of parameters from the ECG signal which is used for data acquisition and classification system. The raw dataset has to be preprocessed before performing any analysis. The missing value handling treatment consist of three method they are e1liminating the missing value, replacing the missing value with average and the

imputation based on machine learning algorithm. This work handles the problem of missing value imputation using boosted K-nn method. The preprocessed data is then classified using Genetic Algorithm, C4.5 and Naive Bayes to classify arrhythmia from ECG medical data sets. The aim of the study is to automatically classify cardiac arrhythmias as a part of a ongoing embedded medical device research, and to study the performance of machine learning algorithms.

1.1 Related Work

An approach for effective feature extraction from ECG signal was described in Saxena et al in [9]. Their paper deals with an efficient composite method which has been developed for data compression, signal retrieval and feature extraction of ECG signals. They carried out detailed studies and by training different topologies of error-back-propagation (EBP) artificial neural network (ANN) with respect to variations in number of hidden layers and number of elements in each hidden layer, the best topology with two hidden layers and four elements in each hidden layer has been finalized for ECG data compression using a Military Hospital (MH) data base.

After signal retrieval from the compressed data, it has been found that the network not only compresses the data, but also improves the quality of retrieved ECG signal with respect to elimination of high-frequency interference present in the original signal. The compression ratio (CR) in ANN method increases with increase in number of ECG cycles. The features extracted by amplitude, slope and duration criteria from the retrieved signal match with the features of the original signal. The test results at each stage are consistent and reliable and prove beyond doubt that the composite method can be used for efficient data management and feature extraction of ECG signals in many real-time applications. Castro et al. in [10] described a wavelet transforms approach for ECG feature extraction. Their paper presented an algorithm, based on the wavelet transform, for feature extraction from an electrocardiograph (ECG) signal and recognition of abnormal heartbeats. A method for choosing an optimal mother wavelet from a set of orthogonal and bi-orthogonal wavelet filter bank by means of the best correlation with the ECG signal was developed. The ECG signal is first denoised by a soft or hard threshold with limitation of 99.99 reconstructs ability and then each PQRST cycle was decomposed into a coefficients vector by the optimal wavelet function. The coefficients, approximations of the last scale level and the details of the all levels, were used for the ECG analyzed. The coefficients of each cycle were divided into three segments, which were related to the P-wave, QRS complex and T-wave, and summed to obtained a features vector of the signal cycles. Alexakis et al. in [11] used automatic extraction of both time interval and morphological features, from the Electrocardiogram (ECG) to classify ECGs into normal and arrhythmic.

Classification was implemented by artificial neural networks (ANN) and Linear Discriminant Analysis (LDA). The ANN gave more accurate results. Average training accuracy of the ANN was 85.07% compared with 70.15% on unseen data. Ramli et al. in [12] investigate the use of signal analysis technique to extract the important features from the 12 lead system (electrocardiogram) ECG signals. Lead II is chosen for the whole analysis due to it representative characteristics for identifying the common heart diseases. The analysis technique chosen is the cross-correlation analysis. Cross-correlation analysis measures the similarity between the two signals and extracts the information present in the signals. Results show that the parameters signal analysis technique extracted could clearly differentiate between the types of heart diseases analyzed and also for normal heart signal. There has been much work in the field of classification and most work has been based on neural networks, Markov chain models and support vector machines (SVMs).

2. PROPOSED METHODOLOGY

In this paper, going to classify cardiac arrhythmias based on signal matching using DTW and existing classification techniques in data mining to drive best technique that could build classification model which gains high accuracy when applied on real time applications without any consuming time. The main goal is to use signal matching instead of feature extractions is to decrease the learning time and increase the accuracy of classifying cardiac arrhythmias.

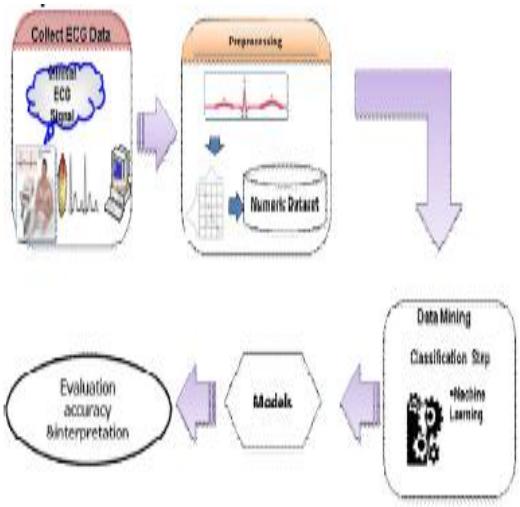


Figure 2.a Frame work steps.

In order to Collect ECG signals from human bodies usually use the leads. The standard ECG is composed of 12 leads, six limb leads are recorded by using arm and leg electrodes, and other six chest leads are recorded using electrodes at six different positions on chest. Now days there are many resources for ECG data such as physionet [14], Framingham Heart Study [15], Risk Assessment Tool [16].. This framework uses MIT-BIH Arrhythmai Database as the dataset. This framework is consisted of two steps, preprocessing step and classification step as shown in figure 2.a.

2.1 Data preprocessing step

The goal of this step is to convert the ECG in suitable format in order to be used in classification step. Most of Previous works was using features for classifying heart arrhythmia beats. In order to use signal matching, first need to determine the window size for reading heart beats accurately from ECG signal based on the heart rate variability to make the window size dynamic.

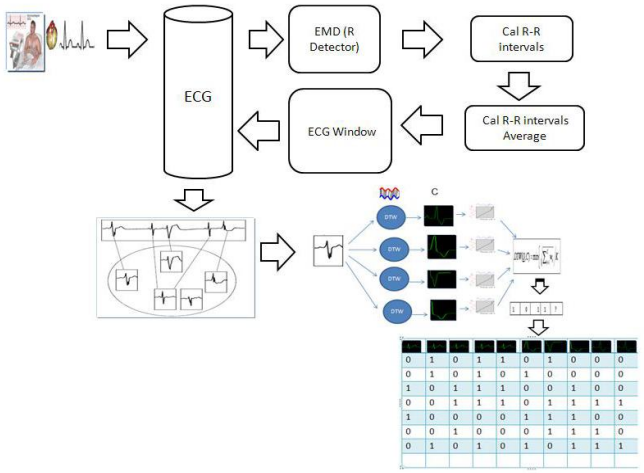


Figure 2.1.a Preprocessing steps in this frame work

The following steps explain the determination of the window size from the heart rate variability. First step is to detect R wave in order to calculate the heart rate variability using empirical mode decomposition (EMD) [17].

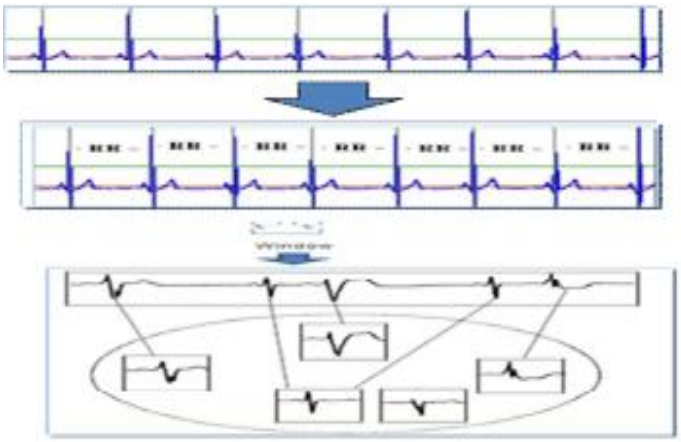


Figure 2.1.b Steps of determining window size.

Detect the R waves to measure the time interval between successive R waves as it is shown in figure 2.1.b. In order to find the window size from heart rate variability, calculate the mean of RR intervals using the following Equations

$$Mean(RR) = \sum_{i=0}^{n} (RR) / N - 1$$

$$L = Mean(RR) / 2$$

$$WindowSize = [L(ms) + Rwave + L(ms)]$$

After calculating the mean of RR- intervals, calculate the window size by dividing the mean of RR intervals into half as shown in equation 2. The window size will be taken for every beat L milliseconds before every R wave and L milliseconds after the R wave as shown in equation 3.After that, determine the ECG templates of shapes. Extracted specific patterns for specific heart arrhythmias specifically premature ventricular contraction (PVCs) and normal patterns or normal beats (N). Every heart cardiac arrhythmia makes specific shape on ECG that changes the amplitude of ECG signal on ECG paper as shown in figure 2.1.c

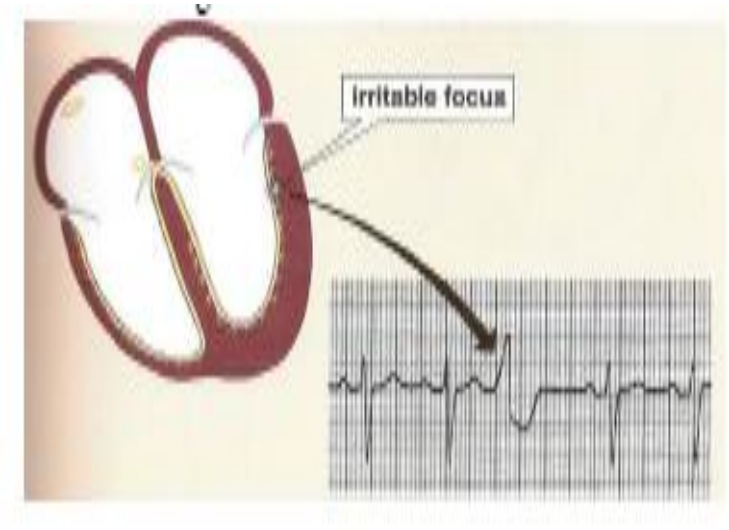


Figure 2.1.c PVC arrhythmia changes the normal shape.

PVC originates suddenly in irritable ventricular automaticity focus and produces again and shapes ventricular on EKG. The PVCs occur early on the cycle easily recognized by great width and enormous amplitude (high and depth), they are usually opposite to the polarity of the normal QRS as shown in figure 2.1.c Extract manually different templates from real ECG signal and from phyionet dataset for PVCs and N beats as shown in figure 2.1.d

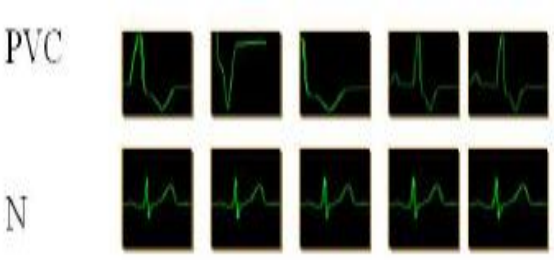


Figure 2.1.d Extraction template shapes of PVC's and N

Beats from ECG signal when the heart beat is road from ECG signal, it is going to be matched with all the shapes that were extracted for PVC and N beats. To find the similarity between the template shapes and heart beat. DTW algorithm is used in which was explained in [18][19] to find the similarity between two signals (time series) that should be computed by aligning significant patterns as shown in figure 2.1.e.

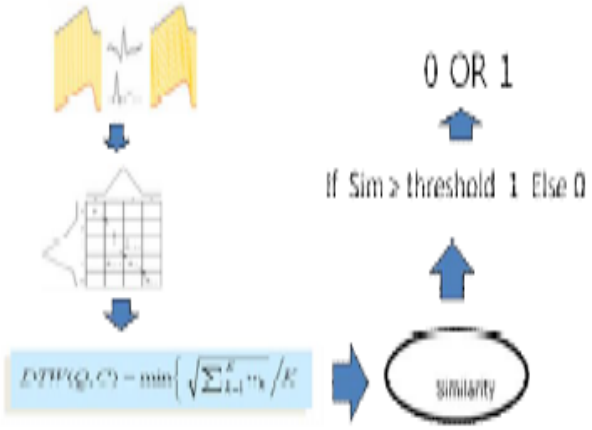


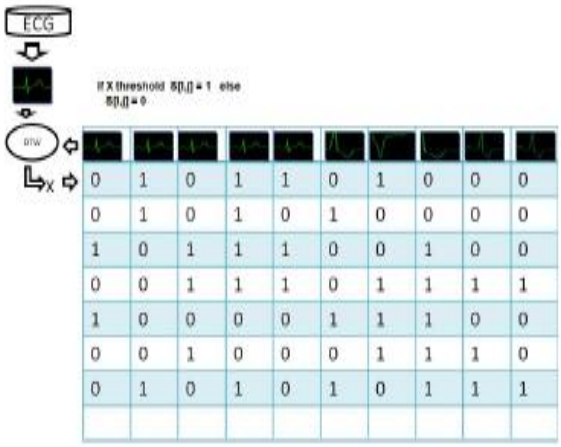
Figure 2.1.e Steps for finding the similarity between two ECG signals

This frame work tries to find the similarities between all the templates and every heart beat. Those similarities equal to the number of ECG templates that were extracted. In order to avoid consuming time in calculating similarities using DTW, use recursive function which gives us the minimum cost path as it shown in equation below

$$\gamma(x, y) = d(qi, ci) + \min(\gamma i\ i-1, j-1), \gamma(i-1, j), \gamma\gamma i\ j-1))$$

Where Q, C are ECG signals. To avoid the fuzziness in the similarities, normalize the similarities into zero or one values based on threshold value which is determined

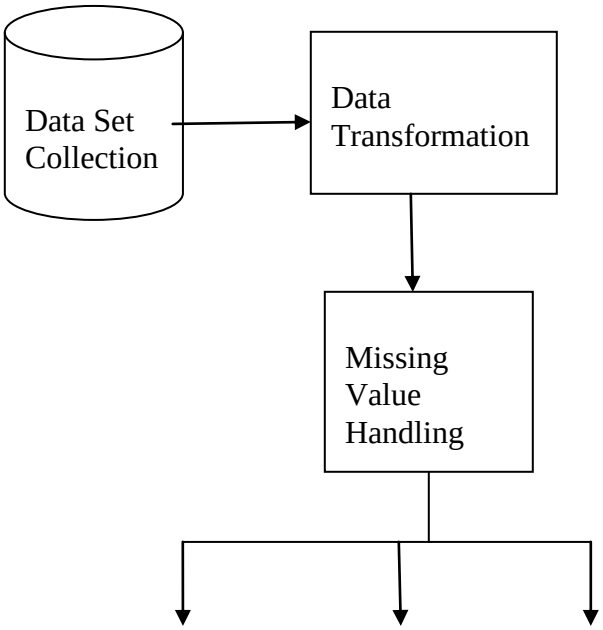
0.06. After normalizing step, put the similarity values into numeric dataset which is considered as input for classification step as shown in figure 2.1.f



0	1	0	1	1	0	1	0	0	0
0	1	0	1	0	1	0	0	0	0
1	0	1	1	1	0	0	1	0	0
0	0	1	1	1	0	1	1	1	1
1	0	0	0	0	1	1	1	0	0
0	0	1	0	0	0	1	1	1	0
0	1	0	1	0	1	0	1	1	1

Figure 2.1.f Output dataset from preprocessing step.

2.2 Architecture of Proposed Methodology



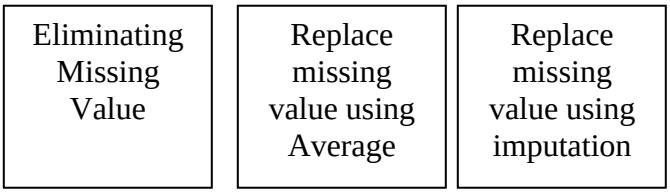


Figure 2.2.a Architecture of proposed Algorithm

2.3 Classification Step

The goal of this step is to apply the classification techniques and generate models .the best model will be selected based on the accuracy, learning time.

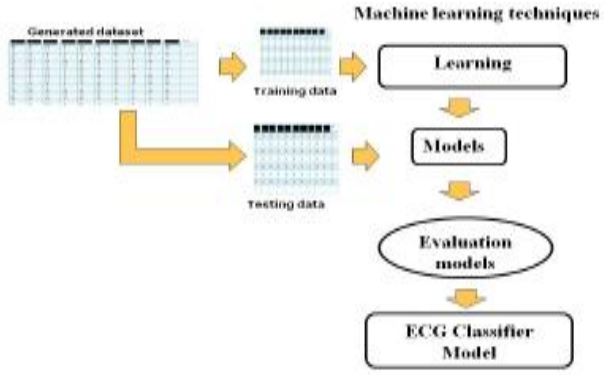


Figure 2.3.a Classification steps.

In this paper, selected four standard machine learning algorithms applied to classify cardiac arrhythmias such like decision tree (DT), bayes naïve (BN) classifier, support vector machine (SVM)and artificial neural network (ANN). Divide the generated dataset into training dataset and testing dataset under different percentages of splitting as shown in figure 2.3.a

2.4METHODOLOGY-CASE/PAIRWISE DELETION METHOD

The first method is the Case/Pairwise Deletion (CD) method which is also known as Complete Case Analysis method (Acuna & Rodriguez, 2009). The CD method in this work adopts the pre-processing strategy where all instances with missing data in at least one of the attributes are deleted during the preprocessing phase. This converts the incomplete data set to a complete one. The complete data set contains only the instances with no missing values in any of the attributes and is input into the mining algorithm.

2.4.1 Algorithm: **Case/pairwise -deletion**. Generate a complete data set from the given experimental data set, D by deleting the records whose attribute contains missing value.
Input: Data set, D.
Output: Data set, D, contains instances with no missing values.
Method:

```
for each case C in D
    for each attribute A of C
        if value of A is null
        delete C
        {endif}
    {end for}
{end for}
```

2.4.2 Algorithm: Replacing Missing Value with *constant numeric value*. Generate a complete data set from the given experimental data set, D by filling the attributes with constant values using zero
Input: Data set, D.
Output: Data set, D, contains instances with no missing values.
Method:

```
for each case C in D
    for each attribute A of C
        if value of A is null
        fill the value of A as Zero
        {endif}
    {end for}
{end for}
```

2.4.3 Algorithm: Replacing Missing Value with attribute mean. Generate a complete data set from the given experimental data set, D by filling the attributes with missing values using mean value

Input: Data set, D.
Output: Data set, D, contains instances with no missing values.
Method:

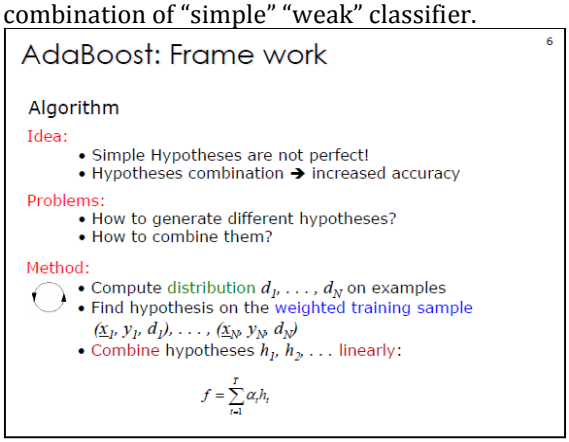
```
    for each selected Attribute A  in D
        Calculate the mean value MV of A
    {end for}
for each selected Attribute A  in D
        for each case C of A
         if the value of A is   null
          fill the value of A as MV
                {endif }
        {end for}
{end for}
```

The proposed substitution method considers that missing values can be substituted by the corresponding attribute value of the most similar complete instance (object) in the dataset.

2.4.4 Algorithm for K-NN

Step 1: determine k
Step 2: calculate the distances based on Euclidean between the missing input record and all the training dataset with complete attribute value.
Step 3: sort the distance and determine k nearest neighbors based on the k_{th} minimum distance.
Step 4: gather the categories of those neighbors.
Step 5: Substitute the missing value by corresponding attribute value of the most similar complete record

The k-nn is boosted using the Adaboost algorithm in order to overcome its weaker classification. AdaBoost is an algorithm for constructing a "strong" classifier as linear combination of "simple" "weak" classifier.



3. EXPERIMENTAL SETUP

The cardiac arrhythmia data analysis and classification study was done using Rapid Miner software environment for machine learning. It is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from some Java code. Rapid Miner contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. Rapid Miner system is open source software issued under the GNU General Public License. In the experiments, the original data set was partitioned into two mutually disjoint sets: a training set and a test set. The training set was used

to train the learning algorithm, and the induced decision rules were tested on the test set.

The settings used in the experiments were as follows. The Genetic Algorithm, C4.5 and Naïve Bayes were used in conjunction with Rapid Miner. Attribute Selection Info Gain Attribute Eval. Attribute Selection Ranker. The cross validation was set to 10 and all other settings were the Rapid Miner program defaults

Results and Impact

The results of the experiment are summarized in Table 3.a, and comparison of the accuracy (or number of correctly classified instances) and learning time (or time taken to build the model) on the dataset between Genetic Algorithm, C4.5 and Naïve Bayes are illustrated in charts 3.a,3.b and 3.1.a. Tables 3.a, 3.b 3.1.a show the trade-off in decreasing learning time and increasing error rate for the three algorithms.

Testing criteria	Naïve bayes		Genetic algorithm		C4.5	
	Classified Instance	Time to build model	Classified Instance	Time to build model	Classified Instance	Time to build model
Training Set itself	51.28	0.74	91.81	0.28	76.55	0.31
% split (50 % train 50%test	59.67	0.57	89.91	0.21	70.8	0.27
% split (70 % train 30%test	58.09	0.44	84.26	0.18	75	0.23
% split (80 % train 20%test	56.04	0.42	87.03	0.12	74.73	0.18

Table 3.a Testing criterion among Genetic Algorithm, C4.5 and Naïve Bayes

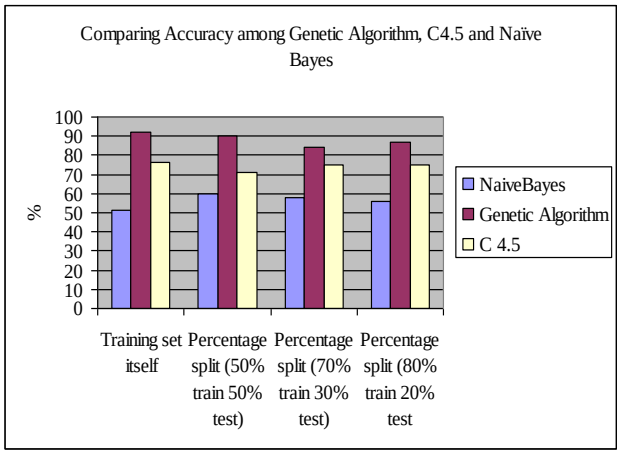


Chart-3.a Comparing *Accuracy* among Genetic Algorithm, C4.5 and Naïve Bayes

As shown in Chart 3.a, the highest accuracy was observed in the case of decision-tree induction algorithm Genetic Algorithm compared with the training data itself.

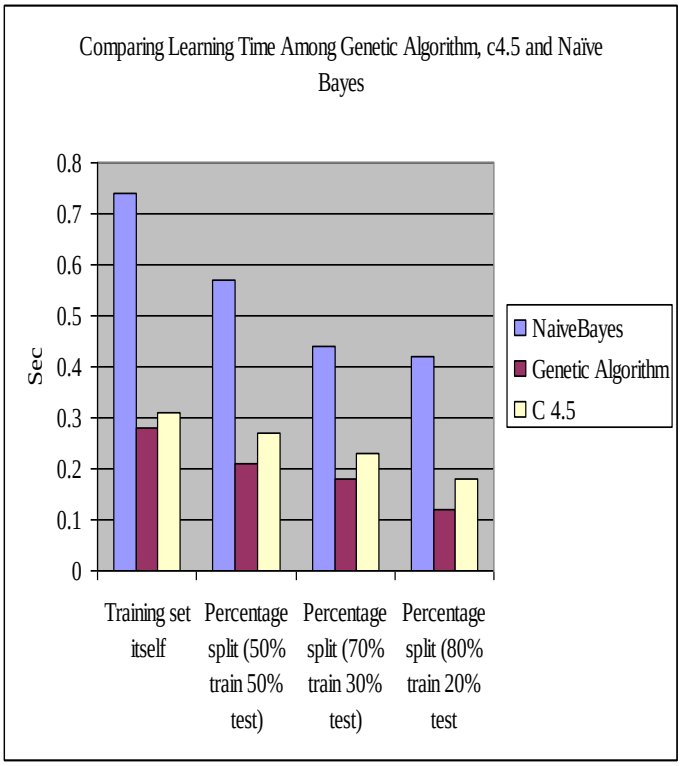


Chart-3.bComparing *Learning Time* Among Genetic Algorithm, c4.5 and Naïve Bayes

Chart 3.b illustrates the learning time comparison of the algorithms. The c4.5 algorithm consumes far more learning time than the other algorithms. The learning time of c4.5 drops drastically at percentage split of 50% and 70%. The learning time of Genetic Algorithm drops at percentage split of 50%. The differences in learning time for Naïve Bayes for different percentage split was found to be not significant

3.1 MISSING VALUE HANDLING

The missing value handling deals with three different approaches they are

- Eliminating Missing Value
- Replace Missing Value with Average
- Replace Missing Value using Imputation Method

The table below shows the performance of the three different classifiers with three different approaches. The result shows that replace missing value using average naïve bayes performs well. The Eliminate missing value and Imputation method using k-nn shows that genetic works fine than the remaining algorithms.

	Accuracy		
	Genetic Algorithm	**C4.5**	**NB**
Replace Missing Value using Average	73.01	76.33	66.59
Eliminate Missing Value	80.23	75.3	63.6
Replace using Imputation Method	90.6	85.3	79.6

Table3.1.aComparing Accuracy among Genetic algorithm, C4.5 and Naïve bayes using 3 methods

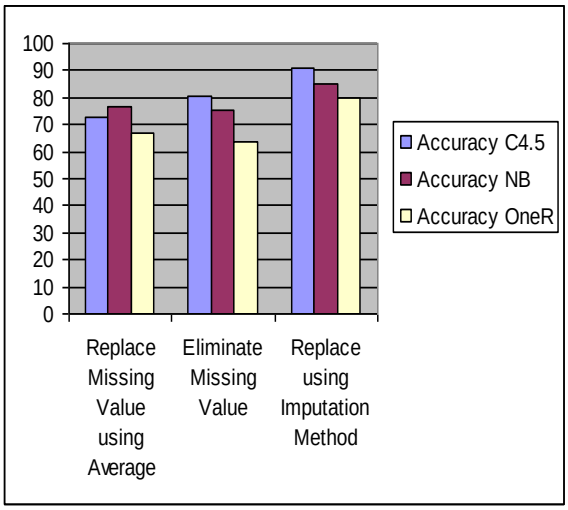


Chart-3.1.a Comparing Accuracy among Genetic algorithm, C4.5 and Naïve bayes using 3 methods

The below table shows the confusion matrix using C4.5 of Imputation Method

C4.5 Replace Missing Value using Average			
	true range1 [-∞ - 1.500]	true range2 [1.500 - ∞]	class precision
pred. range1 [-∞ - 1.500]	208	85	70.99%
pred. range2 [1.500 - ∞]	37	122	76.73%
class recall	84.90%	58.94%	

Table3.1.b Confusion matrix using C4.5 of Imputation Method

The below table shows the confusion matrix using Genetic Algorithm of Imputation Method

GENETIC ALGORITHM Replace Missing Value using Average			
	true range1 [-∞ - 1.500]	true range2 [1.500 - ∞]	class precision
pred. range1 [-∞ - 1.500]	205	67	95.37%
pred. range2 [1.500 - ∞]	40	140	97.78%
class recall	83.67%	67.63%	

Table 3.1.c Confusion matrix using Genetic Algorithm of Imputation Method

The below table shows the confusion matrix using Naïve Bayes Algorithm of Imputation Method

Naïve Bayes Missing Value using Average			
	true range1 [-∞ - 1.500]	true range2 [1.500 - ∞]	class precision
pred. range1 [-∞ - 1.500]	218	124	63.74%
pred. range2 [1.500 - ∞]	27	83	75.45%
class recall	88.98%	40.10%	

Table 3.1.d Confusion matrix using Naïve Bayes Algorithm of Imputation Method

4. CONCLUSION

In the research reported in this paper, three machine learning methods were applied on the task of classifying arrhythmia and the most accurate learning methods was evaluated. Experiments were conducted on the cardiac dataset to diagnose cardiac arrhythmias in a fully automatic manner using machine learning algorithms. The study shows that Genetic Algorithm and have the most stable accuracy rate than Naïve bayes and C4.5 algorithm. The results strongly suggest that machine learning can aid in the diagnosis of cardiac arrhythmias. It is hoped that more interesting results will follow on further exploration of data. Future work includes repeating the experiment with other advanced machine learning algorithms such as support vector machines.

5. REFERENCES

[1] L. Goldberger and E. Goldberger,*Clinical Electrocardiography A Simplified Approach*‖, 5th edition. St. Louis, MO: Mosby, 1994, vol. 1, p. 341.
[2] Fahim Sufi, Ibrahim Khalil, Jiankun Hu ―ECG-Based Authentication‖. Springer 2010
[3] Juan Pablo Martínez, Rute Almeida, Salvador Olmos, Ana Paula Rocha, and Pablo Laguna, ―A Wavelet-Based ECG Delineator: Evaluation on Standard Databases,‖ IEEE Transactions on Biomedical Engineering Vol. 51, No. 4, pp. 570-581, 2004.
[4] Krishna Prasad and J. S. Sahambi, ―Classification of ECG Arrhythmias using Multi-Resolution Analysis and Neural Networks,‖ IEEE Transactions on Biomedical Engineering, vol. 1, pp. 227-231, 2003.
[5] Cuiwei Li, Chongxun Zheng, and Changfeng Tai, ―Detection of ECG Characteristic Points using Wavelet Transforms,‖ IEEE Transactions on Biomedical Engineering, Vol. 42, No. 1, pp. 21-28, 1995.
[6] Saritha, V. Sukanya, and Y. Narasimha Murthy, ―ECG Signal Analysis Using Wavelet Transforms,‖ Bulgarian Journal of Physics, vol. 35, pp. 68-77, 2008
[7] S.Karpagachelvi, Dr.M.Arthanari, M.Sivakumar, ―ECG Feature Extraction Techniques - A survey Approach‖ International Journal of Computer Science and Information Security Vol. 8, No. 1, pp 76-80, 2010.
[8] Sanjay M. Pati1 (1994) ―ECG Analysis - Expert System Approach‖ ME. – Dissertation, V.J.T.I, Bombay – 19
[9] S. C. Saxena, A. Sharma, and S. C. Chaudhary, ―Data compression and feature extraction of ECG signals,‖ International Journal of Systems Science, vol. 28, no. 5, pp. 483-498, 1997.