Report Generation on ECGs Survey Data Analysis Using Threshold Based Inference Engine

Saria Safdar, Shoab Ahmad Khan, and Fahim Arif

Abstract—**Heart diseases and strokes are considered as number one killer as they account for around 35 to 40 per cent of the total disease burden in Pakistan. The ratio of heart patients is increasing day by day, which is an alarming condition for the country. This situation needs a detailed analysis which can show the geographical distribution of heart patients and also the city wise attributes (age, weight, income etc) that are aggregating more in the heart disease. A Threshold Based Inference Engine is designed which infers the knowledge base by generating the association rules on each city. These rules infer the clustered data to extract the city wise more risk increasing attributes, and the common disease in that city. Automated Minnesota code is used for the verification of the collected ECGs. The results show that Threshold based Inference Engine successfully and efficiently generates a detailed report of each city including more diseased people and highlights the attributes increasing the risk factor.**

Index Terms—**Arrhythmias, centroid, ECG, fuzzification, inference engine, membership etc.**

I. Introduction

Cardiac arrhythmia is the leading cause of death worldwide claiming 17.1 [1,] [2], million lives each year, more so than lung, cancer and AIDS combined. In the United States alone, sudden cardiac arrest claims about 450,000 lives each year. Statistics here in Pakistan are even worse. A number of heart patients die every year in Pakistan either due to improper diagnosis/treatment or lack of treatment altogether. Alarmingly, now it starts affecting the younger population in their thirties and forties, the prime of their life. This seems to be very highly alarming ratio in Pakistan especially in the productive year of life. This situation needs attention and a detailed analysis of geographical distribution of heart patients so that proper steps should be taken.

To generate a detailed analysis which can show the geographical distribution of heart patients and the common attributes aggregating the disease, a Threshold Based Inference Engine is designed (TBIS), which takes clustered data as input. A database containing the patient's records along with the ECG's will be generated. Minnesota code is used for the verification of the collected ECGs, all the ECGs are checked according to the conditions specified in the Minnesota code. Statistical analysis in done on the clustered data to extract the statistical based association rules.

These rules are further used to infer the knowledge base. The analysis will highlight geographical areas with maximum disease and the attributes (weight, age, income, drugs etc) contributing more in the heart disease in the particular area. The generated report will provide great benefit for use by Health Organization and international health organization like ministry of health and WHO.

The paper is organized as follows: section 1 describes the introduction, section 2 is literature survey, section 3 is the technique description and section 4 is the conclusion and future work.

II. Literature Survey

Automatic ECG survey data analysis has been an active area of research from the last decade. A wide range of techniques have been used for this purpose which include linear and nonlinear DSP techniques, statistical pattern recognition techniques, Expert Systems, Artificial Neural Networks ,Inference Engines and Fuzzy Logic etc.

According to a World Bank report, which alerts that "Pakistan is facing a health crisis with rising rates of heart disease" [2], Many surveys have been conducted in the country to check the ratio of heart diseases in various cities. A prospective survey conducted for Acute Myocardial Infarction shows that "majority (68%) were males with age group of 52.2 years" [3]. A number of surveys conducted in various cities to check the ratio of Rheumatic Heart Disease. Many studies show "Pakistan among the high risk countries for RHD" [4], [5].Surveys conducted in Lahore and Peshawar shows that, the prevalence of RHD in Lahore is among the highest in the overall ratio of RHD in the world [6], while the ratio in Peshawar has declined as compared to Lahore [7].

Many fuzzy approaches are proposed for ECG diagnosis and classification [8,9] including Fuzzy Adaptive-Resonance TheoryMAP, employed to classify cardiac arrhythmia [9]. A hybrid neurofuzzy system was used for ECG classification of myocardial infarction (MI) [10]. T. M. NAZMY et al, presents an intelligent diagnosis system for (ECG) classification by using hybrid approach of adaptive neuro-fuzzy inference system (ANFIS). The system includes feature extraction of ECGs using Independent Component Analysis (ICA) joined with Power spectrum and the RR interval [11]. ANFIS is commonly used for the classification of ECG signals. To detect premature contraction ANFIS was used in [12], results show that back propagation with the combination of least square convergence is faster than only using back propagation. Sucharita Mitra et al in [13], describes a rule-based rough-set decision system in which an inference engine is used for the detection of the heart disease. Image processing techniques were used for the acquisition of

Manuscript received March 13, 2012; revised April 15, 2012.

S. Safdar and F. Arif are with Military College of Signals (MCS), National University of Sciences and Technology (NUST) Islamabad, Pakistan.

S. A. Khan is with Electrical and Mechanical College (EME), National University of Sciences and Technology (NUST) Islamabad, Pakistan.

the ECGs parameters. Vijilal et al proposed "a hybrid soft computing technique (ANFIS) to estimate the interference and to separate the Electroencephalogram (EEG) signal from its Electrooculogram (EOG)" [14].

III. Threshold Based Inference Engine (TBIE)

Threshold based Inference engine is a form of an engine that tries to derive answers from a knowledge base by inferring the clusters. TBIE consists of a knowledge base and a set of rules, which are derived from the knowledge base; these rules are further used for inference. The output is a detailed report that shows the city wise distribution of the most common disease and also the main attributes that are taking part in aggregating the disease.

A database with details of patient's record along with the city details and disease is generated. The objective of this TBIE is to extract the attributes (city wise) which are aggregating more in the particular heart disease and also the city wise disease and its level. Results show that TBIE successfully extracted the attributes and city wise disease after inferring the knowledge base.

A. Algorithm

The steps that are designed for the algorithm of the proposed Threshold based Inference Engine are presented in Table.I The rest of the section will explain in detail each step, and the outcome of each step.

TABLE I: TBIE Algorithm

Step 1: *Calculate Statistics to Generate Rules*
→Calculate mean and standard Deviation of all attributes in the database.
Step 2: *Calculate Threshold Based on Percentage*
Step 3: *Generate Statistical Rules*
Step 4: *Fuzzification of the input Attributes*
→Apply Trapezoidal Membership Function [15]
Step 5: *Apply Threshold on Membership*
Step 6: *Allegation of Rules*
Step 7:*Calculate Centroid*
Step 8: *Aggregation*
Step 9 : *Report*

B. Flowchart of TBIE

The Flowchart of the proposed technique is shown in figure 1.

Step 1: Shows the input of the inference engine which is the database.

Step 2: Shows the statistical analysis; Mean and standard deviation is calculated for each attribute.

Step 3: Result of step three is the set of detailed rules.

Step 4: Membership is calculated for the selected attributes after applying feature selection.

Step 5: Shows the allegation of rules antecedent to the rules consequent

Step 6: Centroid is calculated for each city.

Step 7: Shows the aggregation of all the centroids.

Step 8: Shows the output which is the detailed report.

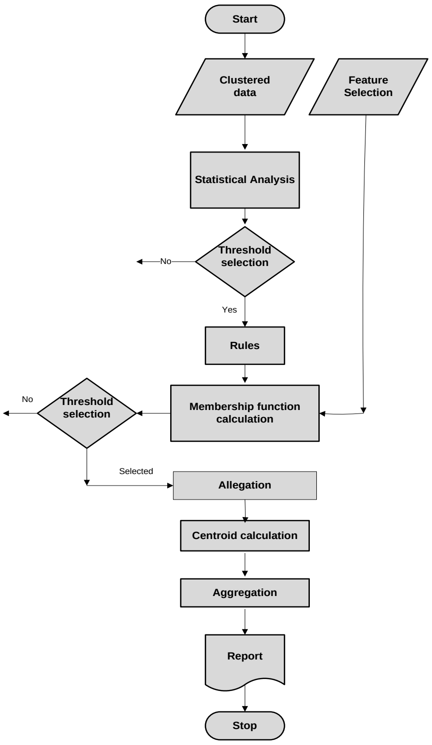


Fig. 1. Flow chart of TBIE.

C. Detail Description of Steps in TBIE

This section describes the details of each step presented in the Threshold Based Inference Engine.

1) Preprocessing

This step involves the preparation of the ECGs database.

2) Minnesota Validation

Minnesota code is a well known medical science paper which is used by the cardiologists for the verification of the ECGs [16]. It has number of clauses which have ranges for the attributes P,QRS, T . These ranges must be satisfied by the selected ECG. This research automated the Minnesota code so that ECGs can be verified and through an automated system and may save time of cardiologists. When the ECG is uploaded it needs verification i.e whether this ECG is accurate to consider; it contains all the parameters or not. If the ECG is verified by the Minnesota code then those ECG parameters, after calculation are saved in the database however if it does not pass the verification criteria it is discarded.

3) Statistical Analysis

A statistical analysis is done on the database .Mean and Standard deviation is calculated for each attribute in the database. This statistical analysis is further used for the rules generation.

a) Statistical Rules Formulation

This section deals with the statistical rules generation. To achieve this task first the threshold is selected. The selected threshold shows the number of diseased people in the province e.g if threshold value is five it will retrieve all the cities in which percentage of diseased people is five or above five. The result is the set of rules, each rule against each city.

Rule for a City:

"*If age (37), and weight (73), and height (5'2''), and education (Fsc), and employment (retired), and income (35k), and per capita income (5000000), and drugs (level never* *used), and BP (100< B <130), and hypertension (permanent), then disease (sinus rhythm), with stage (serious)."*

The rule described above is a very detailed one this rule needs a generalized form with most aggregating attributes only. To achieve this further steps of TBIE are applied and a general rule for each city is extracted.

4) Fuzzification

This section describes the fuzzification process that is applying membership function on each attribute.

a) Trapezoidal Membership Function

A membership function (MF) is defined as "The degree an object belongs to a fuzzy set is denoted by a membership value between 0 and 1. µ*(x)* is called the membership function (or MF) of *x"[15]*. In TBIE trapezoidal membership function is used. "The *trapezoidal* membership function, trapmf, has a flat top and really is just a truncated triangle curve. These straight line membership functions have the advantage of simplicity" [15]. Table II shows the membership calculated for each attribute.

TABLE II: MEMBERSHIP VALUES

City ID	Age	Weight	Education	Employment
331	0.4	0.34	0.3	0.6
Total FI	Drugs	BP		
0.52	0	0.6		

b) Feature Selection

As the database contains so many features some of which are relevant and some of which are not. So it is necessary to highlight those features which are important than other irrelevant features, to achieve this, feature selection is applied on the database. The result of feature selection is a set of features which are used for the membership calculation.

c) Attributes with Maximum Membership

When the membership of all the attributes is calculated there is a need to select most important attributes in each city which are taking part in aggregating the disease. To achieve this task a threshold is set to retain the highest membership holding attributes. Table III shows the selected attributes with there membership values against each city.

TABLE III: ATTRIBUTES WITH MAXIMUM MEMBERSHIP

City ID	BP	Employment	Total FI	PCI	Age
331	0.6	0.6	0.52	0.42	0.4

5) Allegation

A rule has two parts antecedent and the consequent. In this research the antecedent and the consequent are

Antecedent = Age, Weight, BP, Education etc

Consequent = Disease

The consequent part which is the disease has three cases; Serious, Mild, Symptoms.

Like the antecedent part all the rules, statistical analysis and the membership values for each attribute of consequent part is also calculated. The three stages of diseases serious, mild, symptoms for each city is calculated based on their membership values. Table IV shows the result of allegation.

TABLE IV: ALLEGATION

City ID	BP	Employment	Total FI	PCI
331	0.6	0.6	0.52	0.42
Age	Disease	SOD		
0.4	Sinus Rythm	Serious		

6) Province Level Aggregation

To achieve the province level aggregation first city level centroid calculation is done than all centroids form all cities are joined into four sets for four provinces.

a) City Level Centroid Calculation

All the cities with there maximum disease, stage of disease (SOD) and attributes which are most important for that disease are extracted. The work till now is on city level there is a need to know the common attributes on the province level. To achieve this task centroid from each city is calculated by using the formula given below in eq1.

$$C=\frac{\sum_{i=1}^{n} x_{i(index)}\cdot\mu(x_i)}{\sum_{i=1}^{n}\mu(x_i)} \qquad (1)$$

x_i = attribute

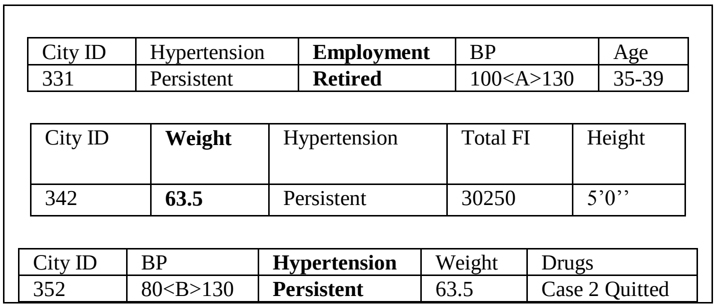
$\mu(x_i)$ = membership of attribute

$\sum_{i=1}^{n}\mu(x_i)$ = sum of membership values

Each attributes (age, weight, education etc) index is multiplied with its membership value. The sum of all the values multiplied with there membership is returned. This sum is divided by the sum of membership values. The result of this formula is a number which indicates the attribute placed on that index in the table.

All the cities with there attribute are shown individually in tables below. When the centroid calculation is done on each city, only one attribute is returned by that city. The Table V below shows the centroid selected from each city.

TABLE V: CENTROID CALCULATION



City ID	Hypertension	**Employment**	BP	Age
331	Persistent	**Retired**	100<A>130	35-39

City ID	**Weight**	Hypertension	Total FI	Height
342	**63.5**	Persistent	30250	5'0''

City ID	BP	**Hypertension**	Weight	Drugs
352	80<B>130	**Persistent**	63.5	Case 2 Quitted

So the centroid calculation is done on each city of each province.

b) Centroid Aggregation

Decisions are based on the testing of all of the generated rules in a TBIE, so the rules must be combined in some manner to make a decision. “Aggregation is the process by which the sets that represent the outputs of each rule are combined into a single set as aggregation only occurs once, just prior to the final step” [17] which is the aggregation of all the provinces. The input of the aggregation process is the list of output attibutes returned by the city level centroid process for each rule. The output of the aggregation process is a set.

After the city level centroid calculation is done there is a need to aggregate all the attributes of all provinces to reach a single output set.

Aggregation= [Employment Weight Hypertension BP Hypertension] + [Age Total Family Income Education Employment] + [Weight Drugs Hypertension BP Education] +[Total Family Income Per Capita Income Drugs Hypertension]

The sets represent the attributes from all four provinces. All the unique attributes from all sets will be extracted but if an attribute like BP has two values than both values will be chosen. The result of aggregation is a single set of attributes showing the most common attributes from all provinces.

Aggregation= [BP (100<A>130), Weight (74), Employment (Retired), Hypertension (Persistent),]

7) *Report Generation*

The section describes the final outcome of the TBIE.

a) Detailed Report Generation

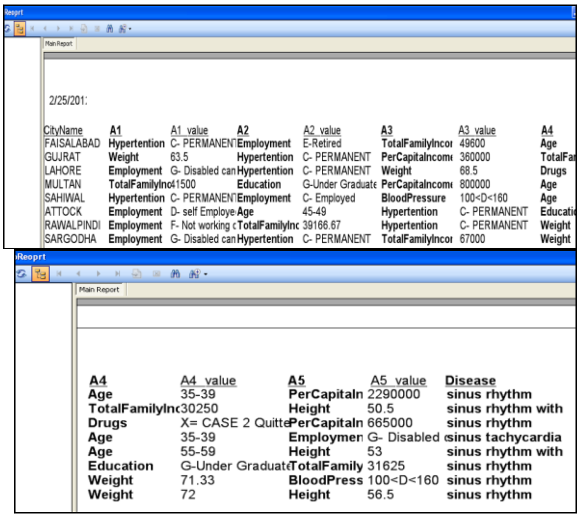
This section describes the desired output of the Threshold based Inference Engine i.e the detailed report. This report captures the city wise distribution of the attributes and the disease in that city. This report is very useful as it clearly shows the overall distribution of the heart disease in Pakistan and the risk factors of the disease.

Crystal reporting is used to generate the detailed report [18].

Fig. 2 shows the detailed report of cities including the common disease and the aggregating factors.

b) Graphical Representation and Discussion

This section describes the graphical representation of the detailed report and discussion on various attributes in various cities.



2/25/201:

CityName	A1	A1 value	A2	A2 value	A3	A3 value	A4
FAISALABAD	Hypertention	C- PERMANENT	Employment	E-Retired	TotalFamilyIncoı	49600	Age
GUJRAT	Weight	63.5	Hypertention	C- PERMANENT	PerCapitalncome	360000	TotalFa
LAHORE	Employment	G- Disabled can	Hypertention	C- PERMANENT	Weight	68.5	Drugs
MULTAN	TotalFamilyInc	41500	Education	G-Under Graduate	PerCapitalncome	800000	Age
SAHIWAL	Hypertention	C- PERMANENT	Employment	C- Employed	BloodPressure	100<D<160	Age
ATTOCK	Employment	D- self Employe	Age	45-49	Hypertention	C- PERMANENT	Educati
RAWALPINDI	Employment	F- Not working c	TotalFamilyInc	39166.67	Hypertention	C- PERMANENT	Weight
SARGODHA	Employment	G- Disabled can	Hypertention	C- PERMANENT	TotalFamilyIncoı	67000	Weight

A4	A4 value	A5	A5 value	Disease
Age	35-39	PerCapitaln	2290000	sinus rhythm
TotalFamilyInc	30250	Height	50.5	sinus rhythm with
Drugs	X= CASE 2 Quitte	PerCapitaln	665000	sinus rhythm
Age	35-39	Employmen	G- Disabled c	sinus tachycardia
Age	55-59	Height	53	sinus rhythm with
Education	G-Under Graduate	TotalFamily	31625	sinus rhythm
Weight	71.33	BloodPress	100<D<160	sinus rhythm
Weight	72	Height	56.5	sinus rhythm

Fig. 2. Detailed report

c) Age

The Fig. 3 shows the graphical representation of age in various cities. It is shown that age is contributing more in heart diseases in Punjab than in any other province. People of Sahiwal and Mansehra of age group above 55 are on more risk than any other age group. More alarming situation is in other cities where age group is below 40 which are considered as the prime of life. Graph shows that young people of Ziarat are on high risk of heart diseases. Other cities having zero value are showing that age is not the important factor in these cities for the heart disease.

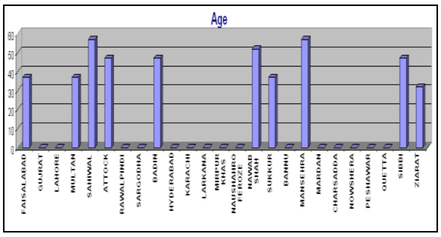


Fig. 3. Age (detailed report)

d) Weight

Fig. 4 shows the distribution of weight in different cities. All the people with weight above 60kg are on high risk for heart disease in many cities as shown in the graph. People of Nawab Shah , Bannu and Mardan having weight above 80kg are involved more in heart diseases than any other weight. In Sindh many cities which are involved in heart diseases, weight above 70kg are on high risk and is the main factor than other provinces. As the graph shows many cities have zero values for the weight which shows that weight is not as much important than other factors in these cities.

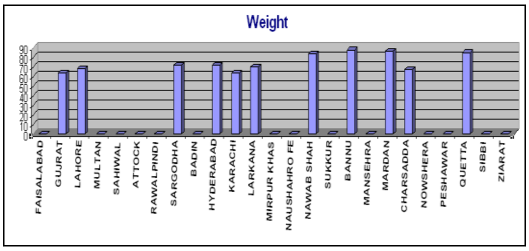


Fig. 4. Weight (detailed report)

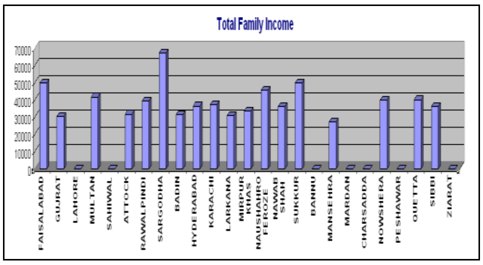


Fig. 5. Total family income (detailed REPORT)

e) Total Family Income

The graph in Fig. 5 shows that income is a very important factor in heart diseases. The ranges in the figure 5 show that both the rich people and the poor people have the heart diseases. Family income above sixty thousand is on same risk as a family with income below thirty thousand .The families with income around thirty thousand and below is on high risk. All the cities in the graph are mainly having family income around thirty or below thirty thousand. Only few cities have zero values for this factor but it can be said for sure that income can be considered as a main factor for the disease.

f) Drugs

Drugs like cigarettes, cigars etc are considered mainly the main cause of heart diseases. But the graph in figure 6 shows that the cities of Baluchistan and NWFP are on more risk for the disease as the people are more involved in drugs intake. On the other hand in Punjab and Sindh drugs are not considered the aggregating factor.

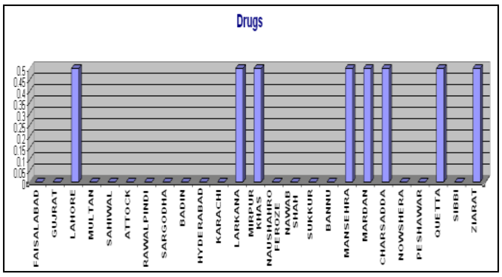


Fig. 7. Drugs (detailed report)

g) Hypertension

Hypertension is the main factor in increasing the heart disease as it is present in all cities shown in Fig. 6 with its maximum value that is hypertension with stage permanent. Multan is the only city where hypertension is not as much important than other factors. So precautions need to be taken for people with hypertension to save there lives from heart diseases.

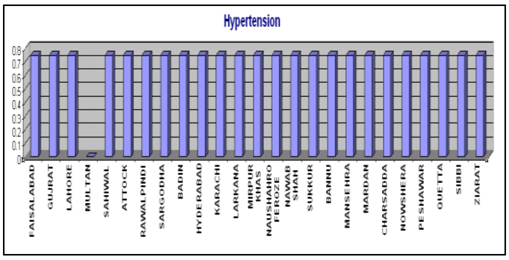


Fig. 7. Hypertension (detailed report)

IV. Conclusion

This research discusses in detail the proposed Inference Engine, flowchart of the Inference Engine and the step wise execution detail of the engine. Results show that Threshold Based Inference Engine successfully extracted the attributes from all the cities and the most common disease. The generated detailed report shows the city wise distribution of the heart patients the aggregating attributes and the disease. The analysis shows that the hypertension has the maximum membership values in all cities which show that it is the major risk factor for the heart diseases. The second factor is the income in all cities, middle class seems to have more heart diseases than the high class people in some cities. But this distinction is not true in some cities where the people with high income and low income are at same risk level for the disease.

The future work which can be done is to make an adaptive Neuro Fuzzy Inference System by preparing the data with the Threshold based Inference Engine and then applying this to the resilient back propagation technique.

Acknowledgment

I would like to thank my parents, my teachers and my friends for their constant support and motivation. I would also like to thank Mr Aihab Khan for their guidance and support.

References

[1] WHO Country Office in Pakistan. [Online]. Available: http://www.emro.who.int/pakistan/programmes_ncd.htm

[2] World Bank report Heart disease, diabetes on the rise in Pakistan – The Express Tribune.mht by Mustafadon Islamabad Pakistan Feb 10,2011

[3] M. H. Jafary, A. Samad, M. Ishaq, S. A. Jawaid, M. Ahmad, and E. A. Vohra, "Profile of Acute Myocardial Infarction (AMI) In Pakistan," *Pakistan Journal Of Medical Sciences Quaterly*, vol. 23, no. 4, 2007

[4] S. S. A., R. M., and H. J. A, "Establishment of comprehensive research and rehabilitation program for persons of various heart diseases," Project, VRA Pak. Karachi: National Institute of Cardiovascular Diseases. pp.8–66, 1973.

[5] S. M. Malik, S. Jaffery, S. Ahmed, and K. Zubeda, "Prevalence of heart disease in school children in Islamabad," *Pak Heart J14*,1981.

[6] M. Sadiq, Department of Paediatric Cardiology, Punjab Institute of Cardiology, Ghaus-ul-Azam (Jail) Road, Lahore, Pakistan; "Prevalence of rheumatic heart disease in school children of urban Lahore," Accepted 9 September 2008 Published Online First 24 October 2008

[7] A. Gul, M. U. Hassan, and M. Hafizullah "Rheumatic Heart Disease in Urban School Children of Peshawar," Department of Cardiology, Postgraduate Medical Institute, Lady Reading Hospital, Peshawar – Pakistan 2009

[8] F. M. Ham and S. Han, "Classification of cardiac arrhythmia using fuzzy ARTMAP," *IEEE Trans. Biomed. Eng*, vol. 43, no. 4, pp. 425–430, 1996.

[9] Y. H. Hu, W. J. Tompkins, J. L. Urrusti, and V. X. Afonso, "Applications of artificial neural networks for ECG signal detection and classification," *J. Electrocardiol.*, vol. 26, pp. 66–73, 1993

[10] P. Bozzola, G. Bortolan, C. Combi, F. Pinciroli, and C. BroHet, "A hybrid neuro-fuzzy system for ECG classification of myocardial infarction," in Proc. Comput. Cardiol., Indianapolis, IN, pp. 241–244, 1996.

[11] T. M. Nazmy, H. El-Messiry, B. Al-Bokhity "Adaptive Neuro-Fuzzy Inference System for Classification of ECG Signals," *Journal of Theoretical and Applied Information Technology* © 2005 - 2009 JATIT. All rights reserved.

[12] S. M. Lam F. Wan, and M. C. Dong, Faculty of Science and Technology, University of Macau Macau China "Automatic Detection of Premature Ventricular Contractions Using Adaptive Neuro-Fuzzy Inference System," 2009

[13] S. Mitra, M. Mitra, and B. B. Chaudhuri, "A Rough-Set-Based Inference Engine for ECG Classification," *IEEE Transactions on Instrumentation and Measurement*, vol. 55, no. 6, 2006.

[14] C. K. S. Vijilal, P. Kanagasabapathy, S. Johnson' and V. Ewards' "Artifacts Removal in EEG Signal using Adaptive Neuro Fuzzy Inference System," *IEEE - ICSCN 2007, MIT Campus, Anna University, Chennai*, India. pp.22-24, pp.589-591,2007.

[15] Membership Function. [Online]. Available: http://www.orsc.edu.cn/~liu/Lecture/USet/Memberfunction/Memberfunction.pdf

[16] The Minnesota Code Classification System= for Electrocardiographic Findings

[17] Fuzzy Inference Systems Tutorial (Fuzzy Logic Toolbox™).mht [Online]. Available: http://www.mathworks/

[18] J. Crick, "White Paper Crystal Reports Server XI Functional Overview" Contributor: Davythe Dicochea, MaryLouise Meckler, Jennifer Meegan, James Thomas