IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 14, Issue 4 (Sep. - Oct. 2013), PP 13-21
www.iosrjournals.org

Data Classification Algorithm Using k-Nearest Neighbour Method Applied to ECG Data

Mrs. A. R. Chitupe, Prof. S. A. Joshi
Department of Computer Engineering, Pune Institute of Computer Technology, Maharashtra, India
Department of Computer Engineering, Pune Institute of Computer Technology, Maharashtra, India

Abstract: *In medical science, the importance of the Electrocardiography is remarkable since heart diseases constitute one of the major causes of mortality in the world. Electrocardiogram (ECG) is the only way for doctors to see the cardiac actions of a particular person. It provides a graphic depiction of the electrical forces generated by the heart and then by analysing this graph doctors can tell about any abnormality present in heart.*

In the paper we focus on the QRS complex detection in electrocardiogram and the idea of further recognition of anomalies in QRS complexes based on some dimensional features of ECG is described. As medical information system is widely used and growing medical databases requires efficient classification method for efficient computer assisted analysis of ECG.

I. Introduction

At present, heart disease became a leading cause of death. Heart disease can be controlled effectively if it is diagnosed at an early stage .But unfortunately; accurate diagnosis of heart disease has never been an easy task. As a matter of fact, many factors can complicate the diagnosis of heart diseases, often causing the delay of a correct diagnosis decision. According to doctors ECG is not the only thing need to be considered to analyse ECG. Some other behavioural features like work pattern, mental stress, exercise of patient should also be considered.

In this paper a new set of parameters is considered to analyse ECG. These parameters consist of some dimensional features of patient's ECG and some behavioural features of patient which would be helpful for efficient analysis of ECG. There can be change in area under normal ECG and abnormal ECG. We have used Simpson's rule to get the area under ECG signal considering X axis as base line and Scanline algorithm to get the same considering Y axis as base line.

An efficient mining strategy need to be used for extracting new knowledge about ECG signals. After collecting all parameters, K nearest neighbour classification method is used to get the result i.e. Doctor's impression about a particular patient.

The rest of this paper is organized as follows. Section 2 explains related work in automated ECG analysis and Section 3 explains programmer's design including mathematical model, dynamic programming and serialization, data independence and data flow architecture and turing machine. Section 4 shows the result. Section 5 concludes the paper.

II. Related Work

Knowledge of the ECG signals for healthy and defective cases is the base for the heart diagnoses [1]. To extract main features from ECG signal for analyzing it, different methods are used as stated below.

Chia-Hung Lin *et al.,* reported that for ECG signals, a two-subnetwork classifier can be used to discriminate normal rhythm from six cardiac rhythm disturbances by combining Morlet wavelets and probability neural network (PNN) [2]. The transformations involve matrix multiplications resulting in worst case complexity of the order O(n3). Alternative to transformation can also be one of the important challenges for effective and efficient image processing and advance software technologies like mining of image and or video data.

P. Sasikala *et al.,* reported that the physiological and geometrical differences of the heart in different individuals display certain uniqueness in their ECG signals [3]. ECG can be used as a Biometric tool for Identification and Verification of Individuals. Geometrical difference in ECG signals are extracted using DWT(Discrete Wavelet Transform). Wavelet transformation involves matrix multiplication and results into increase in complexity. Alternative way to find geometrical difference in ECGs can be one of the objectives for effective image processing.

V. S. Chauhan *et al.,* reported that the feature extraction can be done using a modified definition of slope of the ECG signal[4]. New feature set including other dimensional features of ECG signal can be considered to analyse changes in ECG signal.

A.Dallali *et al.,* used Wavelet Transformation and Artificial Neural Network To analyze the ECG signal [5]. Heart Rate Variability is one of the features considered while classifying ECGs. To consider such new features for ECG classification can be one of objectives.

S. S. Mehta *et al.,* also used Support Vector Machine (SVM) as a classifier for QRS complexes (QRS complex is a name for the combination of three of the graphical deflections seen on a typical ECG) detection in ECG signal and evaluated on the standard Common Standards for Electrocardiography (CSE) database[6]. The proposed method is accurate in many cases but failed to detect the small duration features of ECG signal. Alternative way to detect small duration feature of ECG signal can be considered for effective image processing.

Victor Dan Moga M.D. *et al.,* concluded that Wavelet Transformation is a suitable for analyzing physiological signals like ECG signal although more complicated mathematically than any other technique [7]. Alternate simple way considered for efficient analysis of physiological signals like ECG signal.

C. Saritha *et al.,* reported Wavelet transformation is suitable way to analyze ECG signals but some methodological aspects of wavelet technique require further investigation [8]. Wavelet transformation is a complex task because of involved matrix multiplication resulting into worst case complexity as O(n3). Alternative way for ECG signal analysis need to be considered.

Jiapu Pan et al., developed a real time algorithm by considering slope, amplitude, width information to detect QRS complex from ECG signal [10]. Other features of ECG signal can be used for analysing ECG signal. Finding such new feature set for analysis of ECG signal can be one of the objectives.

III. Design of Proposed Work

The aim to the dissertation is to develop a new dimensional feature set of ECG for better analysis of ECG signal. As per Doctor's opinion ECG signal is not the only thing need to be considered to analyse patient. Some behavioural parameters are also considered before coming to any conclusion. A suitable mining strategy need to be considered. Supervised learning method is used instead of unsupervised learning method because of its better time complexity. The ECG database is trained and classified using K nearest neighbour method for better accuracy.

3.1. Mathematical Model

Let E represents the test ECG

E={i, ar_x, ar_y , A, G, W, H, r | f(x), f(y)}

Where, I represents ECG id which will be unique for each ECG. ar_x and ar_y represents areas under the curve from base line as X axis and Y axis respectively. Let A be a random variable which represents the index of the age of patient having values {0,1,2,3,4} mapped on age set as shown in following Venn Diagram.

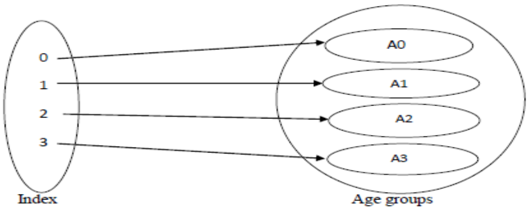

Figure 1: Venn diagram for age group of patient

A0 : 0–10 Years
A1 :10–20 Years
A2 : 21–50 Years
A3: 51- 110 Years

Let G be a random variable having four possible values as 0, 1, 2 and 3 mapped as following groups G0, G1, G2, NN. Also shown in following Venn Diagram.

G0 : Female and Pregnant
G1 : Female and not Pregnant
G2 : Male and not Pregnant
NN : Invalid

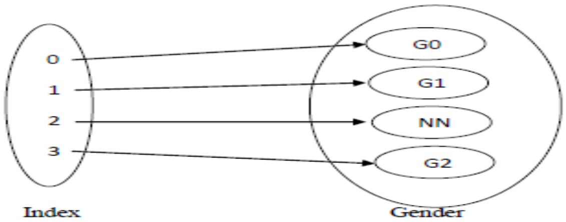

Figure 2: Venn diagram for gender information of patient

Let W be a random variable having eight possible values from 0 to 7 mapped on following groups. This variable records information about physical and mental stress to a patient. The groups are described as follows:
W0 : No Exercise, No Fieldwork, No Mental Stress
W1 : No Exercise, No Fieldwork, Heavy Mental Stress
W2 : No Exercise, Heavy Fieldwork, No Mental Stress
W3 : No Exercise, Heavy Fieldwork, Heavy Mental Stress
W4 : Do Exercise, No Fieldwork, No Mental Stress
W5 : Do Exercise, No Fieldwork, Heavy Mental Stress
W6 : Do Exercise, Heavy Fieldwork, No Mental Stress
W7 : Do Exercise, Heavy Fieldwork, Heavy Mental Stress.

These groups are indexed on values shown in following Venn diagram.

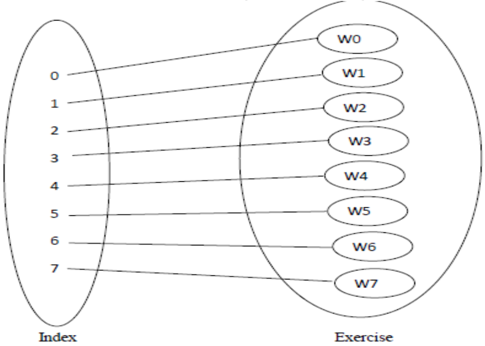

Figure 3: Venn diagram for information about mental and physical stress to patient

Let H be a random variable representing some medical information about patient like if patient is having any hereditary heart disease or respiratory disease. Groups are as follows:
H0 : No hereditary heart disease, No respiration problem
H1 : Invalid
H2 : No hereditary heart disease, Occasionally respiration problem
H3 : No hereditary heart disease, Regular respiration problem
H4 : Hereditary heart disease, No respiration problem
H5 : Invalid
H6 : Hereditary heart disease, Occasionally respiration problem
H7 : Hereditary heart disease, Regular respiration problem

These groups are mapped according to H variable value as shown in figure.

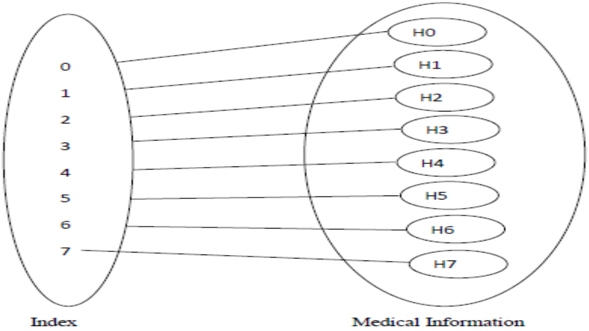

Figure 4: Venn diagram for medical information of patient

3.1.1 New Dimension Features of ECG

Area under ECG signal is considered as a new dimensional feature of ECG to analyse it. Simpson's rule is one of the way to calculate Area under ECG signal considering X axis as base line. Scanline algorithm is used to calculate same considering Y axis as base line.

I. Simpson's Rule

It uses parabolas to approximate each part of the curve as shown in figure. This proves to be very efficient way of calculating area under the curve. Area under the curve using Simpson's rule is having smaller error if compared with area under the curve using Trapezoidal rule. If curve equation can be represented by f(x) then

$$Area=\int_a^b f(x)\ dx$$

By Simpson's Rule,

$$Area\ = \tfrac{1}{3}\,(\Delta X)*(\,Y0 + 4Y1 + 2Y2 + 4Y3 + 2y4 + \dots\dots + 4yn\text{-}1+yn)$$

Where,

n represents the total number of segments(parabolas) in which total area is divided and it must be even. Δx represents the width of each segment. We can also calculate it using a formula

$$\Delta x=\frac{b-a}{n}$$

Y0, Y1,.....Yn represents area of each segment.

Y0= f(a)
Y1= f(a+Δx)
*Y2=f(a+2*Δx)*
Yn-1=f(b-Δx)
Yn=f(b)

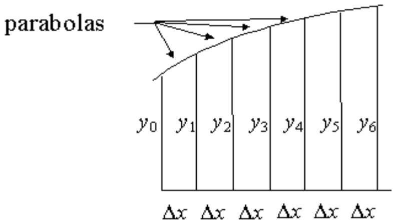

Figure 5: Division of curve in parabolas for Simpson's rule

II. Scanline Algorithm

For each scan line:

1. Find the intersections of the scan line with all edges of the polygon.
2. Sort the intersections by increasing x-coordinate.
3. Fill in all pixels between pairs of intersections.

For each scan-line:

There can be a Problem with corners. Same point need to count twice if it is a corner. To decide whether the intersection point should be considered twice or not following solution can be considered.

1. Make a clockwise or counter-clockwise traversal on edges.
2. Check if y is monotonically increasing or decreasing. If a direction change there is a double intersection, otherwise single intersection.

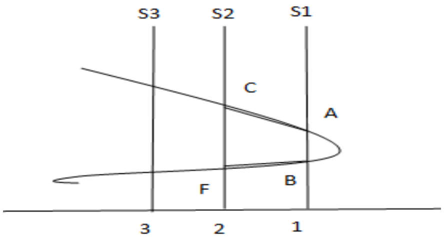

Figure 6: Scanline algorithm

Consider scanline s1 intersects ECG signal at A and B points. Scanline S2 intersects at C and F points. Draw trapezoids as shown in figure 6. S1 line forms two trapezoids, one from each intersection point as AC21 and BF21. Now calculate areas for both the trapezoids.

Let distance between scanlines is d and

Length(A1)=l1	Length(C2)=l2
Length(B1)=j1	Length(F2)=j2

$$\text{Area (AC21)} = \frac{1}{2}\,\text{d (l1+ l2)}$$

$$\text{Area (BF21)} = \frac{1}{2}\,\text{d (j1+ j2)}$$

$$\text{Area (ACFB) = Area (AC21) – Area (BF21)}$$

$$= \frac{1}{2}\,\text{d (l1+ l2- j1- j2)}$$

3.1.2 K-Nearest Neighbour Method for Data Classification

K-nearest neighbour algorithm (KNN) is a classification method based on closest training samples. It is an instance-based learning algorithms that, instead of performing explicit generalization, compare new problem instances with instances seen in training, which have been stored in memory. It is called instance-based because it constructs hypotheses directly from the training instances themselves.

In this method, particular parameter need to be chosen base on which classification will be done. All samples present in training set have their own value. Now compare test sample with all other samples already present in training set. There are various methods present for comparing these values like Hamming distance, Euclidean distance. If the variable is continuous Euclidean distance is suitable distance metric.

Let Xs is a test sample described with parameters as [Xs1,Xs2,Xs3.....,Xsn] and one of the sample from training set is Xt described as [Xt1,Xt2,Xt3.......,Xtn]. The Euclidean distance between these two samples Xs and Xt is defined as

$$ED(Xs,Xt) = \sqrt{(Xs1 - Xt1)^2 + (Xs2 - Xt2)^2 + \cdots .. + (Xsn - Xtn)^2} \quad (1)$$

ED (Xs,Xt) = ED (Xt,Xs)

The Equation (1) uses the Euclidean distance calculated using test samples and samples from training set. Now by comparing all the distance find nearest neighbours that are having minimum Euclidean distance. In K-Nearest neighbour algorithm first K neighbours need to be considered and according to majority neighbours one label is assigned to test sample also.

In figure 7, consider the circle with no colour is a test sample. 1 and 3 belong to same red class, whereas 2,4,5 belong to same yellow class. If the K value is 1 then the test sample will be considered as of red class because the nearest neighbour 1 is of class red. If K value is 2 then there may be a tie. to avoid this confusion generally odd value is assigned to K. if K value is 5 then majority of the neighbours are having yellow class so test sample will be considered as of yellow class.

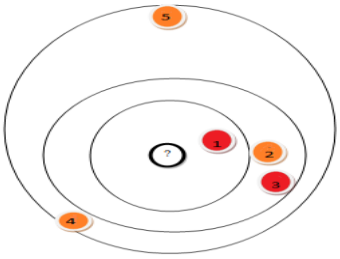

Figure 7 : K-Nearest Neighbour algorithm

3.2. Dynamic Programming and Serialization

Let set s be the set of all the operations that can be done using this system.

S={ A, D, U, V}

Where,

A- Addition of new patient record
D- Deletion of patient record
U- Updating an existing record
V-Viewing all records as per requirement

As all these operations are independent of each other can be done in parallel.

3.3. Data independence and Data Flow architecture

The user will send request for addition of new record with all necessary information of patient and ECG to add record module. Some of in the information required add new entry in database is generated there itself like area under curve for ECG signal, Label and doctor's impression i.e. result. The result is given back to the user and new record is added into database.

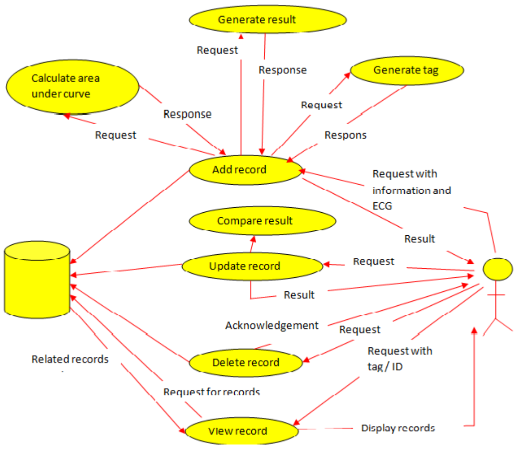

Figure 8: Data flow diagram

If user wants to update existing record then it will send request for updating a record with proper ID of the record and information for updating a record. Then record is updated and new result is generated with new label and area under the curve. This new result is compared with previous result and comparison is displayed to the user.

User can send request for deleting a record with valid ID of the record to be deleted. Record is accessed and deleted from database and acknowledgement is given to the user as a response.

User can send request to view the record or records. Valid ID or tag need to be given to access similar records and then displayed to user as response as shown in figure 8.

3.4 Multiplexer Logic

Consider set P be the set all the information about patient, that need to be considered while classifying it.

$$E=\{i, ar\_x, ar\_y , A, G, W, H \mid f(x), f(y)\}$$

Where, I represents ECG id which will be unique for each ECG. ar_x and ar_y represents areas under the curve from base line as X axis and Y axis respectively. A represents age of patient, G represents gender of patient, W represents the exercise pattern of patient, H is the hereditary medical information. Though the ECG input for each patient is in the same format , all other information of patient and geometrical features of ECG are the main parameters need to consider to classify the given patient record in appropriate class.

K nearest neighbor classification method is used for deciding the class of given record.

3.5 Turing Machine

Turing machine can be adapted to simulate the logic of any computer algorithm, and is particularly useful in explaining the functions any system. Turing machine representation of this dissertation is shown in figure with the states given below.

States:
S1: Accept user credentials for login
S2: Login failed
S3: Login successful
S4: Take choice of action
S5: Addition of new record
S6: If user is doctor
S7: If user is patient
S8: If user is student
S9: Accept patient medical information and ECG image
S10: Validate information
S11: User is already having a record
S12: Invalid information
S13: Valid information
S14: Generate label using KNN
S15: Valid label is generated successfully
S16: Unable to generate label
S17: Calculate area under curve for ECG signal
S18: Add record
S19: Compare result

S20: View result

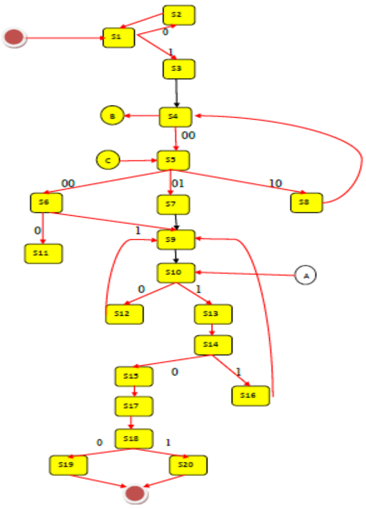

Figure 9: Turing machine diagram for Add module

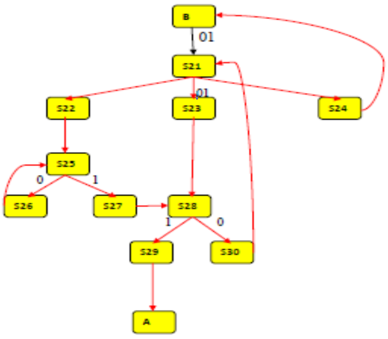

Figure 10: Turing machine diagram for Update module

States:
S4: Take choice of action
S21: Updation of record
S22: if user is doctor
S23: if user is patient
S24: if user is student
S25: Accept ID
S26: Invalid ID
S27: Valid ID
S28: Access Record
S29: Record found
S30: Record not found
S31: Accept information for updation

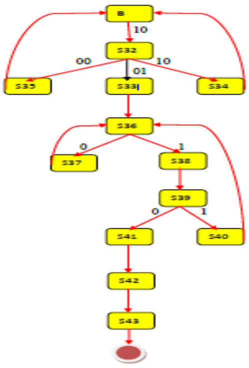

Figure 11 : Turing machine diagram for Delete module

States:
S32: Deletion of record
S33: If user is doctor
S34: If user is patient
S35: If user is student
S36: Accept ID
S37: Invalid ID
S38: Valid ID
S39: Access record
S40: Record not found with specified ID
S41: Record found with specified ID
S42: Delete the record
S43: Send acknowledgement.

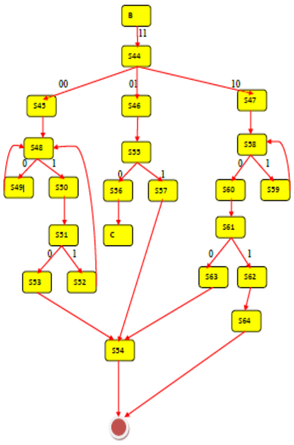

Figure 11: Turing machine diagram for View module

States:
S44: Display record
S45: If user is doctor
S46: If user is patient
S47: If user is student
S48: Accept ID or tag
S49: Invalid ID or tag
S50: Valid ID or tag
S51: Access record
S52: Record not found
S53: Record found
S54: Display record
S55: Access record
S56: Record not found
S57: Record found
S58: Accept tag
S59: Invalid tag
S60: Valid tag
S61: Access records
S62: Record not found
S63: Access record
S64: No records with specified tag are available.

IV. Results and Discussion

In order to analyze the performance of given system classification of sample ECGs need to be done.

Number of correctly classified ECGs
Accuracy=--
Total number of ECGs classified

Accuracy of proposed classification method is totally dependent on the new features used to analyze the ECG.

V. Conclusion

In this work new dimensional features of ECG signal are considered. This would be useful to analyse the ECG and categorize it either in normal or abnormal group. In contrast to the existing way of doing the same thing using regular dimensional features of ECG. A database used to check the result is generated here itself, no other data is used. The input data consist of scanned ECG image which is never used before. This would make the application more user friendly. Anyone who is not friendly with computer can use this application. As ECG images are used as input, feature extraction is simpler than doing same using transformation method which involves high complexity. Use of KNN makes the system more flexible and takes less time for training.

We can enhance the system by introducing some more heart diseases with some more personal and medical information about patient. We can also try different classification methods for better result and lesser time. Cloud computing can be used to make this system available everywhere in the world.

References

[1] Urszula Markowska-Kaczmar *et al*, "Mining of an Electrocardiogram" *in XXI Autumn Meeting of Polish Information Processing Society Conference Proceedings* pp. 169–175, 2005.

[2] Chia-Hung Lin *et al*, "Multiple Cardiac Arrhythmia Recognition Using Adaptive Wavelet Network*", in proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference Shanghai, China, September* pp1-4, 2005.

[3] P. Sasikala, Dr. R.S.D. Wahidabanu, "Robust R Peak and QRS detection in Electrocardiogram Using Wavelet Transform", *in International Journal of Advanced Computer Science and Applications*, Vol. 1, No.6, December 2010.

[4] V.S. Chouhan *et al*, "Delineation of QRS-complex, P and T-wave in 12-lead ECG*", in International Journal of Computer Science and Network Security*, VOL.8 No.4, April 2008.

[5] A. Dallali *et al*, " Integration of HRV, WT and Neural Networks for ECG Arrhythmias Classification", in ARPN Journal of Engineering and Applied Sciences, VOL. 6, NO. 5, May 2011.

[6] S. S. Mehta *et al,* "Comparative Study of QRS Detection in Single Lead and 12-Lead ECG Based on Entropy And Combined Entropy Criteria Using Support Vector Machine", *in Journal of Theoretical and Applied Information Technology,* 2007.

[7] Victor Dan Moga et al, "Wavelets As Methods For ECG Signal Processing", *University of Medicine and Pharmacy Timisoara, Eftimie Murgu Square No. 1, Timisoara Romania*

[8] C. Saritha, V. Sukanya, Y. Narasimha Murthy, " ECG Signal Analysis Using Wavelet Transforms", *Anantapur, Andhrapradesh, India.*

[9] K.V.L.Narayana *et al*, "Noise removal using adaptive noise cancelling, analysis of ECG using MATLAB", *in International Journal of Engineering Science and Technology,* Vol. 3 No. 4, Apr 2011.

[10] Jiapu Pan, Willis J. Tompkins, "A Real Time QRS Detection Algorithm", *in IEEE Transactions on Biomedical Engineering,* Vol. 3 No. 3, March 1983.

[11] Chia-Hung Lin *et al, "* Fractal Features for Cardiac Arrhythmias Recognition Using Neural Network Based Classifier", *Proceedings of the 2009 IEEE International Conference on Networking, Sensing and Control, Okayama, Japan,* March 2009

[12] Manish Sarkar, Tze-Yun Leong, " Application of K-Nearest Neighbour Algorithm on Brest Cancer Diagnosis Problem*", The National University of Singapore, Singapore.*

[13] Farid Melgani and Yakoub Bazi, *"*Classification of Electrocardiogram Signals With Support Vector Machines and Particle Swarm Optimization", *IEEE Transactions on Information Technology in Biomedicine*, vol. 12, no. 5, september 2008

[14] Fahim Sufi and Ibrahim Khalil, "Diagnosis of Cardiovascular Abnormalities From Compressed ECG: A Data Mining-Based Approach", *IEEE Transactions on Information Technology in Biomedicine,* vol. 15, no. 1, january 2011

[15] Hui Wang,"Nearest Neighbors by Neighborhood Counting", *IEEE Transaction Pattern Analysis And Machine Intelligence,* Vol. 28, No. 6, JUNE 2006

[16] Mahfuzah Mustafa *et al; "* Comparison between KNN and ANN Classification in Brain Balancing Application via Spectrogram Image", *Journal of Computer Science & Computational Mathematics,* Vol. 2, Issue 4, April 2012 .

[17] Mohamed I. Owis *et al,*"Study of Features Based on Nonlinear Dynamical Modeling in ECG Arrhythmia Detection and Classification", *IEEE Transactions on Biomedical Engineering,* vol. 49, no. 7, july 2002