Classification of ECG Data for Predictive Analysis to Assist in Medical Decisions.

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

Abstract
In recent years due to physical and mental stress in the working environments the cases of medical diagnosis using ECG are increasing up-bounds. The critical decisions in diagnosis referring to the normal ECG or indicative dysfunctions of the heart results into overlapped data values causing ambiguities. This research paper performs analytical processing and related mining to classify normal and abnormalities of the ECG. The ECG is a graphical representation generated due to polarities of the weak electrical signals generated in certain defined timely manner. With reference to time an ECG is used to measure the rate and regularity of heartbeats, as well as some special behavior of the patient. ECG can be used to investigate heart abnormalities. With increased number of patients and reported diseases, it is becoming mandatory of maintaining medical databases and effective classification method for mining the effective relation between causes. This paper investigates the results of KNN (K-Nearest Neighbor) algorithm to find relation between geometric parameters like area and behavioral parameters of ECG especially in pregnancy cases.

Keywords
ECG, QRS Complex, Data mining, KNN.

1. Introduction

The functional diagram is shown in Figure 1. Possible input will be from either existing datasets or the dataset generated from scanned ECGs. Medical information of patient is also accepted as an input to generate bit pattern. After pre-processing the ECG signal using Matlab functions, Geometrical parameter values are calculated. For this Scanline algorithm and Simpson's rule are used. All these geometrical and behavioural parameters are given to KNN algorithm to classify given sample. According to classification, result is displayed either as Normal or Abnormal ECG.

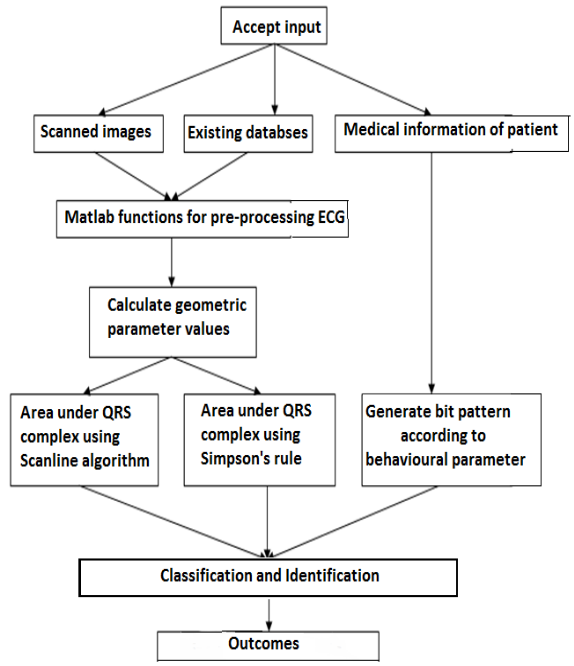

Figure 1: Functional Diagram

According to the medical definition, one of the most important information about ECG signal is QRS complex [3]. Data mining is a process of analyzing and establishing correlation or pattern among different fields of database. It allows user to analyze the data from variance perspective, categorize it and summarize useful relationships. The patterns, associations, or relationships among all this data can provide useful information. To analyze ECG some geometric and behavioural parameters are considered [1]. The outcome analysis of KNN algorithm is presented and compared with the existing results of different methods.

Table 1: Other proposed methods for detecting R-peaks

Manuscript received October 5, 2015
Manuscript revised October 20, 2015

	Parameters					
Methods	Cases considered	True peaks (TP)	missed peaks (FN)	False peaks (FP)	Detection error De=(FP+FN)/(TP+FN)	Sensitivity S=TP/(TP+FN)
DoM	48	115971	166	58	0.19	99.8
WT	48	116025	112	65	0.18	99.9
Method suggested by F. Chiarugi *et al (ref4)*	48	115871	266	210	0.4	99.7
method suggested by Mohamed Elgendi et *al(ref5)*	20	43343	1224	37	34.37	97.5
Method suggested by JIAPU PAN *et al (ref6)*	48	115860	277	507	0.68	99.7

Methods in above table are using Discrete Wavelet Transform for ECG analysis. As the transformation involves matrix multiplication resulting in high worst case complexity, an alternative method is used here.

Data required for this is used from DAISY dataset And PHYSIONET dataset, available freely for research purpose. Data from actual visits and random sampling of scanned images of ECG signals is also considered.

Table 2: Total number of samples considered

Database Name	No. of signals considered
MITBIH Database	35
DAISY Databse	6
visits & random sampling	38
Total	79

The rest of this paper is organized as follows. Section 3 explains need of data mining in analyzing medical data. related work in automated ECG analysis and . Section 4 explains Parameters for analyzing ECG. In the same section, geometric parameters and methods to calculate it are explained. Section 5 shows the result tables and Section 6 concludes the paper.

3. Need of data mining in analysing medical data

Classification is one of the data mining tasks and new emerging technology, which is well suited for the analysis of data [2].The main difficulties while analysing any medical data are as follows

I. In future, database can be very large.

II. There are some exceptional cases that creates confusion even for Doctors, like an abnormal ECGs and Pregnant normal woman ECGs have similar geometric parameter values. It may create confusion and results in incorrect analysis of ECG. Table 3 shows the potential side effects of drugs used for heart problems, if taken in pregnancy.

Table 3: Drug used for heart problems and their potential side effects in pregnancy [18]

Drug	Potential Side Effects
Adenosine	None reported
Beta blockers	Fetal bradycardia, low birth weight, hypoglycemia, respiratory depression, prolonged labor
Digoxin	Low birth weight, prematurity
Diuretics	Reduced uteroplacental perfusion
Lidocaine	Neonatal CNS depression
Low-molecular-weight heparin	Hemorrhage, unclear effects on maternal bone mineral density
Nitrates	Fetal distress with maternal hypotension
Procainamide	None reported
Unfractionated heparin	Maternal osteoporosis, hemorrhage, thrombocytopenia, thrombosis,
Warfarin	Warfarin embryopathy, fetal CNS abnormalities, hemorrhage

Adverse Drug reaction (ADR) i.e. harm directly caused by a drug at normal doses is the third most common error in medical field [15].If automated system for analyzing ECG is considered, ECG in pregnancy may be analyzed as abnormal ECG and one of these drugs may be suggested. To avoid such confusion these special cases need to be handled carefully.

III. Poor mathematical categorization of the data so very few efforts have taken to generate and analyze the mathematical formulation.

4. Parameters for analyzing ECG

The performance of an automated ECG analysis systems depends heavily on the reliable detection of QRS complex [2]. After detecting QRS complex, geometrical parameters like area and behavioural parameters like age, gender are used to analyze the ECG [1]. Following methods are used get information about area.

4.1 Area under QRS complex

QRS complex is a name for the combination of three of the graphical deflections seen on a typical ECG. QRS complex is an irregular curve. Following integration methods are used to calculate area under QRS complex considering it as an irregular curve.

4.1.1 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. By Simpson's Rule, area of irregular curve is given as follows

$$\text{Area} = \frac{1}{3}(\Delta x) * (y_0 + 4y_1 + 2y_2 + 4y_3 + 2y_4 + + \cdots \ldots + 4y_{n-1} + y_n) \quad (1)$$

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.

$$\Delta x = \left(\frac{b-a}{n}\right) \qquad (2)$$

$y_0, y_1, \ldots \ldots, y_n$ represents area of each segment.

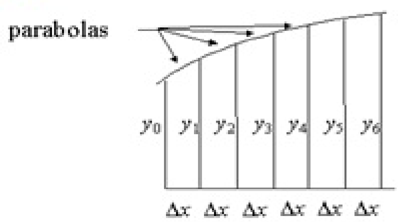

Figure 2: Division of curve in parabolas for Simpson's rule

Figure 2 shows the flowchart for Simpson's Rule used to calculate Area under QRS complex. Where,
N- Total Number of Segments
dx- Width of each segment
Y[1]……Y[N]- Length of each segment

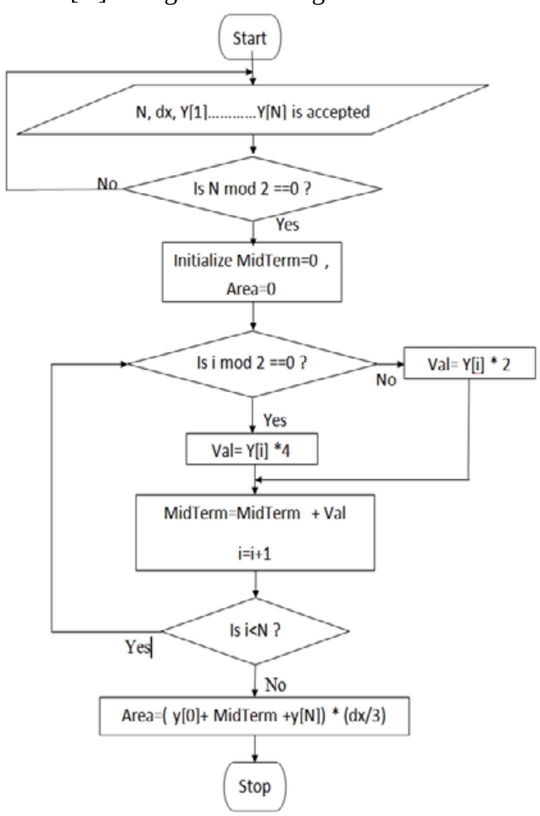

Figure 3: Flowchart for Simpson's Rule

4.1.2 Trapezoidal Rule

It uses Trapezoids to approximate each part of the curve. By Trapezoidal Rule, area can be calculated as

$$\text{Area}=\frac{1}{2}(\Delta x) * (y_0 + 2y_1 + 2y_2 + 2y_3 + \cdots \ldots + + 2y_{n-1} + y_n) \qquad (3)$$

Where, n represents the total number of segments (trapezoids) in which total area is divided and it must be even. Δx represents the width of each segment. Where,

$$\Delta x = \left(\frac{b-a}{n}\right) \qquad (4)$$

$y_0, y_1, \ldots \ldots, y_n$ represents area of each segment.

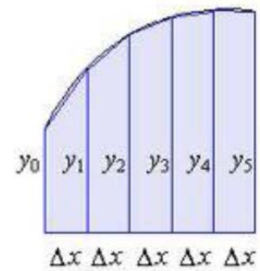

Figure 4: Division of curve in trapezoids for Trapezoidal rule

Figure 4 shows the flowchart for Trapezoidal Rule used to calculate Area under QRS complex. Where,
N- Total Number of Segments
dx- Width of each segment
Y[1]……Y[N]- Length of each segment

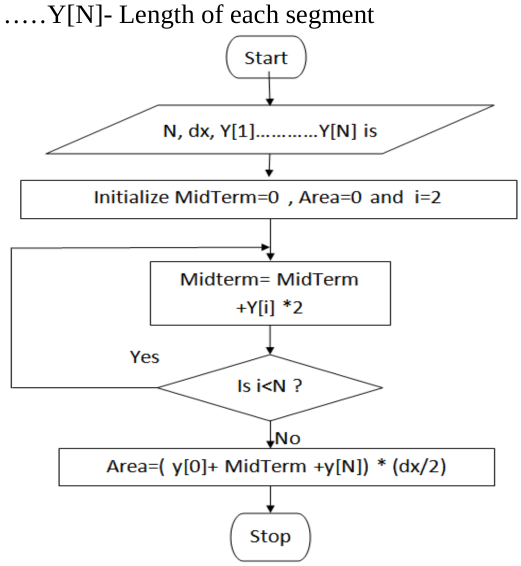

Figure 5: Flowchart for Trapezoidal Rule

4.2 Scanline algorithm

This is a graphical method used to calculate area under QRS complex. After detecting Q,R and S points from the ECG signal, an area is calculated using Scanline algorithm. Simpson's method and Trapezoidal method used to calculate approximate area, so there is an error present in area, whereas Scanline algorithm gives area with minimal error.

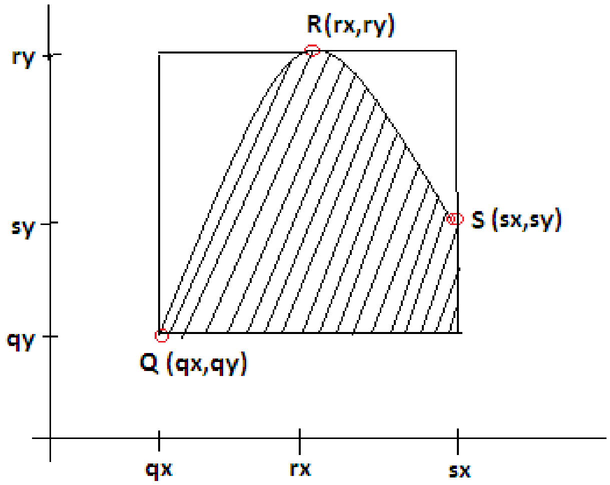

Figure 6: Scanline algorithm for area under QRS complex

After detecting Q,R and S points , it is bounded with the square as shown in figure 6 with the help of minimum x, y coordinates and maximum x , y coordinates. The scanline is then applied to calculate area under one QRS complex present in signal, as shown in yellow colour. Area of all QRS complexes present in one signal are calculated and added.

5. Results

Table 4 shows the addition of areas under QRS complexes present in some samples from different databases. samples used are from PHYSIONET and DAISY ECG database [16][17] . Following table also has

-Number of true R peaks detected in a sample

-Number of false peaks detected in a sample

-Number of missed R peaks in a sample.

Table 4.a: Results of samples from dataset containing Scanned ECGs

Type of database	samples	Area under curve using scanline	Area under curve	Area under curve	True pea	Missed peak	False peaks
Scanned ECGs	100	8459	7732	7722.7	2	0	0
	101	14261	13928	13923	2	0	0
	102	12996	9585	9537	3	1	0
	103	16087	15459	15446	2	0	0
	104	9146	8645	8489	2	0	0
	105	24853	24309	24306	2	0	0
	106	7079	6426	6415	2	0	0
	108	3307	9975	9913	2	0	0
	109	18754	17105	16853	2	1	0

Table 4.b: Results of samples from dataset containing Scanned ECGs

Type of database	samples	Area under curve using scanline	Area under curve	Area under curve	True pea	Missed peak	False peaks
Scanned ECGs	111	14287	12747	12698	4	0	0
	112	10118	9379	9372	2	0	0
	113	15524	14395	14373	2	0	0
	114	234	7986	7931	0	1	2
	121	8886	8488	8204	3	0	1
	122	6790	6358	6264	3	0	0
	123	6894	6066	6046	2	0	0
	124	3593	2771	2724	1	1	0

Table 4.c: Results of samples from dataset containing Scanned ECGs

Type of database	samples	Area under curve using scanline	Area under curve	Area under curve	True pea	Missed peak	False peaks
Scanned ECGs	200	2178	1984	1904	1	2	0
	201	5595	5373	5108	2	0	1
	202	5870	3755	2347	1	0	2
	203	4432	3557	3490	2	0	0
	208	21459	16999	16906	3	0	0
	209	9942	8635	8639	2	0	2
	210	7798	7544	7342	3	0	0

Table 4.d: Results of samples from dataset containing Scanned ECGs

Type of database	samples	Area under curve using scanline	Area under curve	Area under curve	True pea	Missed peak	False peaks
Scanned ECGs	16265	27388	21567	21521	3	0	1
	16272	8894	8275	8250	2	0	0
	16273	35499	34094	34075	3	0	0
	16420	17793	16621	16560	2	0	1
	16483	7209	5770	5723	2	1	0
	16786	24710	23777	23768	2	0	0
	16795	13940	12591	12578	2	0	0
	17052	11585	10887	10879	2	0	0
	17453	16346	15292	15279	2	0	0
	18177	14655	23292	23210	3	0	2
	18184	10543	9920	9909	2	0	0
	f1	5702	5014	5007	2	0	0
	p1	86887	64293	63658	2	0	1
	p2	93028	76552	76283	4	1	1
	f2	3964	3550	3539	1	1	0
	f3	8354	7582	7577	2	0	0

Table 4.e: Results of samples from DAISY dataset

Type of database	samples	Area under curve using scanline	Area under curve	Area under curve	True pea	Missed peak	False peaks
DAISY	p3	33520	21674	20427	13	0	0
	p4	29205	21924	21412	13	0	0
	p5	53532	38089	37135	13	0	0
	p6	40547	27519	26585	13	0	0
	p7	25606	17924	17306	13	0	0
	p8	25985	17767	17069	12	0	0

Figure 6 shows the graphical representation of area under QRS complexes of different samples.

Three methods are used to calculate it as Scanline algorithm, Trapezoidal rule, Simpson's Rule.

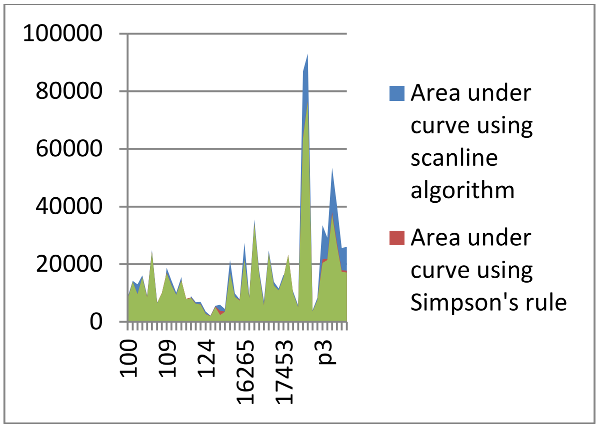

Figure 6: Graphical representation of area under QRS complex using different methods

Table 5 shows the possible range of areas for ECG from different categories. Abnormal ECG and pregnant normal ECG have some common range for areas. This may lead to a situation where pregnant normal ECG may be detected as an abnormal ECG and suggest some drugs accordingly.

Table 5: Opportunity to identify signal as normal or abnormal based upon area under QRS complex.

Type of ECG	Simpson's Method	Trapezoidal Method	Scanline Algorithm
Normal	8275 - 21567	8250 - 21521	8894 - 27388
Pregnant (Normal)	17763 - 76552	17069 - 76283	25985 - 93028
Abnormal	< 8275 and >21567	<8250 and >21521	<8894 and >27388

Following figure shows the common range of QRS area which is normal when considered as in case of pregnancy and same is abnormal when considered as in normal case. K-Nearest Neighbor method can be used to classify such data successfully [1].

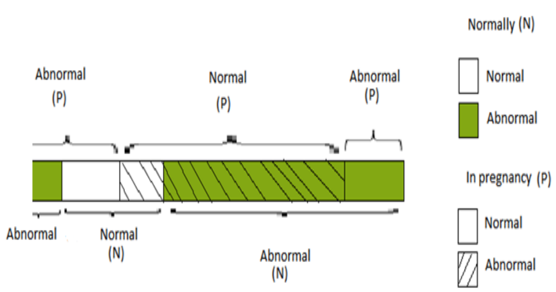

Figure 7: Graphical representation of table 5 (Simpson's method).

Two samples from scanned ECG dataset are of pregnant women as follows

samples	Area under curve using scanline algorithm	Area under curve using Simpson's rule	Area under curve using Trapezoidal rule
p1	86887	64293	63658
p2	93028	76552	76283
p3	33520	21674	20427
p4	29205	21924	21412
p5	53532	38089	37135
p6	40547	27519	26585
p7	25606	17924	17306
p8	25985	17767	17069

Figure 8: pregnant women samples

First two samples are scanned ECGs and other are from DAISY dataset. These samples are successfully classified as normal ECG using KNN algorithm [1]. If the patient is considered as non pregnant then same ECG sample is classified as abnormal because of the difference between bit pattern of medical information of both the patient. Patient ptest1 and p1 are having bit pattern according to following information.

Medical information	**Patients**	
Name	ptest1	p1
Respiration Problem	No	NO
Gender	Male	Female- Pregnant
Hereditary Problem	No	NO
Mental Stress	No	NO
Field Work	No	NO
Exercise	No	NO
Age group	21-50	21-50

Figure 9: Medical information for patients

Though the geometrical parameters are having same values, difference in medical information of patient may results in different result.

Whereas the suggested method using KNN algorithm classify this sample as abnormal if bit pattern suggests it is a general patient but it will classify it as abnormal if bit pattern for the patient provides pregnancy information

6. Conclusion

To use an automated systems in any field like medical, cases which creates confusion need to be handled very carefully. Normal ECG for pregnant women may be detected as abnormal ECG by automated system. The use of certain medications during pregnancy increases the risk of birth defects and other adverse birth outcomes. Efforts need to be taken to prevent such confusion in diagnosis in special cases discussed above in order to decrease the risks of adverse birth outcomes and birth defects.

References

[1] A.R. Chitupe et al, “Data Classification Algorithm Using K-Nearest Neighbour Method Applied to ECG Data”, in International Journal of Advanced Research in Computer Science and Electronics Engineering (IJARCSEE),September2013.

[2] Urszula Markowska et al, " Mining of an electrocardiogram", in XXI Autumn Meeting of Polish Information Processing Society Conference Proceedings pp. 169–175, 2005.

[3] Yun-Chi Yeh et al, “QRS complexes detection for ECG signal: The Difference Operation Method”, in Journal of computer methods and programs in biomedicine 91, 2008.

[4] F Chiarugi et al, “Adaptive Threshold QRS Detector with Best Channel Selection Based on a Noise Rating System”, in Journal of Computers in Cardiology pp157−160,2007.

[5] Mohamed Elgendi et al, “ Improved QRS Detection Algorithm using Dynamic Thresholds”, in International Journal of Hybrid Information Technology Vol. 2, No. 1, January, 2009

[6] Jiapu pan et al, “A Real-Time QRS Detection Algorithm”, in IEEE Transactions on Biomedical Engineering Vol. 32, No. 3, March 1983.

[7] M.Sabarimalai Manikandan et al, “A novel method for detecting R peaks in electrocardiogram (ECG) signal”, in Journal of Biomedical Signal Processing and Control, March 2011.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] B.Madasamy et al, "General Framework for Biomedical Knowledge With Data Mining Techniques", in International Journal of Computer Trends and Technology, Vol. 4 No. 5, May 2013

[14] "Medication Error-Risk Prevention in Infusion Therapy", study by B. Braun Medical Pvt. Ltd.

[15] ECG database from PHYSIONET, "http://physionet.org/physiobank/database/".

[16] DAISY ECG database. " http://homes.esat.kuleuven.be/~smc/daisy/daisydata.html".

[17] Anjli Maroo et al, " Pregnancy and Heart Disease", in disease Management Project , January 2009