International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 17 (2015) pp 37525-37530

An Exhaustive CHAID based Authentication Approach for Remote Health Monitoring

Meenakshi Nawal[1]
Assistant Professor, Department of Computer Science,
Banasthali University, Jaipur (Rajasthan)
Meenakshi.nawal.02@gmail.com

Dr. Mahesh Bundele[2]**,**
Coordinator research Department of Computer Science and Engineering
Poornima University Jaipur (Rajasthan)
maheshbundele@gmail.com

Dr.G.N.Purohit[3]
Dean, AIM & ACT Department of Computer science and Mathematics
Banasthali University Jaipur (Rajasthan)
gn_purohitjaipur@yahoo.co.in

Abstract
In today's life due to increasing population and large number of diseased people remote health monitoring has become an indispensable aspect. Authentication of the patient and the corresponding data plays a significant role in transferring the patient's data for analyzing and subsequent treatment advice while monitoring the patient remotely. In most cases it is required to remotely monitor patient's electrocardiogram signal that can also serve the purpose of human identification. This paper presents analysis of 12-lead Electrocardiogram (ECG) signal using statistical classifier Exhaustive Chi-square Automatic Interaction Detector (Exh.CHAID) for authentication in remote patient monitoring systems. The 12 lead ECG signal has been processed to obtain 3 frank leads Vx, Vy, Vz and then Principal Component Analysis (PCA)has been used to preprocess the frank leads. Wavelet and statistical features were extracted and analyzed through Exh. CHAID for classification of ECGs. The performance analysis of the classifier was carried out by varying number of subjects under consideration. The analysis for patient identification could lead to better accuracy for lower number of subjects, but decreases as the number of subjects are increased. The maximum value of Identification Accuracy found was 100 % for 5 subjects, 72 % for 10 subjects, and likewise it decreases to 49.92% for 25 subjects. This work showed that statistical classifiers can also be used for authentication in remote health monitoring systems.

Keywords: ECG, PCA, Wavelet, CHAID, Authentication, Remote Health Monitoring.

Introduction

The objective of this work is to design a simple patient identification system to be used in remote patient monitoring system for authentication. Most of the remote patient monitoring systems uses physiological parameters including ECG for monitoring, critical analysis and advice / treatment. Before the storage and analysis of received physiological or health data the monitoring system needs to make sure that the data coming from the right person under treatment. As the patients could handover credentials to someone else or may be unable to use their credentials, there is a need for captured data based authentication instead of username password etc. Moreover, one time authentication using credentials or trait based biometrics, face, fingerprints do not cover the entire monitoring period and may lead to unauthorized post authentication use. Remote health monitoring plays a crucial role for patients with high risks and chronic disorders. Furthermore high cost of hospital treatment, the necessity of home care assistance, less availability of expert doctors and high quality instruments also demand auto patient monitoring. Recent studies have shown that some human body parameters unveil unique patterns that can be used to discriminate individuals. This paper deals with a simple patient identification system using ECG as identification and authentication parameter. The human ECG offers several benefits as a biometric as it is universal, continuous and difficult to counterfeit.

Basics of ECG:

The human ECG reflects the specific pattern of electrical activity of the heart throughout the cardiac cycle, and can be seen a changes in potential difference. The proposed system uses ECG based biometrics scheme as an input physiological parameter to monitor the patients. The ECG signal from different individuals not only confirms to a fundamental morphology but also exhibits several personalized traits such as retaining timings of the various peaks, beat geometry and responses to stress and activity. There are various factors that affect the ECG including age, body, weight and cardiac abnormalities. Fig. 1. illustrates structure of ECG signal.

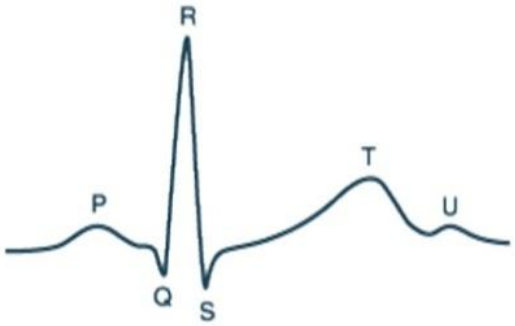


Fig.1. Basic shape of an ECG

The named data networking is a architecture of network recently proposed for the internet. In NDN, data are addressed by names instead of locations of nodes. Named Data networking has very smart benefits for wireless sensor network. First by sending out new interest packets a mobile receiver can update the routing states in intermediate nodes continuously. Thus the data can flow through the same reverse paths traversed by the packets. This also guarantees that the last energy information of sensor nodes can reach the sensors in time. Second to scale to larger network size we divide the network into groups, we formed the clusters. The energy information is gathered in aggregated forms.
Thus we ensure that the data packets are bounded to an group rather than particular node. Thus the data can be addressed by the area name.

Related Work

Many feature extraction and classification techniques are used by researchers to identify individuals through ECG. Biel et al (1999) [1] stated that ECG can be used to identify individuals. SIMENS equipment was used to extract temporal and amplitude features. They have reduced dimensionality using correlation matrix. Multivariate analysis has been used to classify 20 subjects and 100% accurate identification could be achieved for selected subjects. The research did not include the ECG with physical activity like jogging, stressing, walking etc. Shen et al (2002) [2] used template matching method and Decision Based Neural Network (DBNN)for patient identification through one lead ECG. Correlation coefficients were obtained and the template matching was used. The results were compared with DBNN for the possible candidates selected. From template matching method 95% correct identification was achieved and 80%correct identification could be achieved for DBNN. Combining both methods the accuracy percentage could reach 100%. Further the limitation has been that the physical activity aware ECG data was not considered. Israel et al (2005) [3] applied the lambda method for features selection and Linear Discriminate Analysis(LDA) has been used to classify the impact of 7 mental stress for 29 subjects the identification rate was achieved up to 98% but again this study also did not examine the impact of physical activity. Wang et al(2008) [4] proposed an approach that used Discrete Cosine Transform (DCT) along with Auto Correlation analysis (AC). AC\DCT method captured the recurring but non periodic characteristic of ECG signal by computing the coefficients of auto correlation. Identification accuracy of 96%could be achieved from classification of 13 subjects with Linear Discriminate Analysis(LDA). Chan et al [5] and Chiu et al (2008) [6] used a Discrete Wavelet Transform for feature extraction. Identification accuracy was 89% and 100% with 50 and 35 subjects. David et al. (2009) [7] included activity stimulate ECG variations by extracting a set of features that characterize different physical activities along with ECG. The performance of two classifiers K-Nearest Neighbor and Baysian Network (BN) was analyzed. From classifying 17 subjects the identification accuracy was obtained as 88%. Can ye et al (2010) [8] used Independent Component Analysis (ICA) and Wavelet Transform to extracts features. Support Vector Machine (SVM) was used to classify the data of 47 subjects and the accuracy could be found was 99.6%. M.M.T Abdelraheem et al (2012) [10] used two algorithms, first equal distance descriptor and second Fourier Descriptor Coefficient of the main loop of Vector Cardiogram (VCG). Performance of the methods was 99.45% and 95% correct identification. In this paper a simple patient identification based on statistical method has be presented.

Architecture of Proposed System

Fig. 2 shows an architectural view of remote health monitoring system wherein the patient's 12 lead ECG is recorded and transmitted via wireless router to distant Hospital Data Management server through internet. On server side it depicts the process of authentication proposed.
The detailed working of patient identification system proposed is shown in Fig. 3. uses 12 lead ECG input for processing. It clearly demonstrates the preprocessing of 12 lead ECG received at the Hospital Database Management server that includes conversion of 12 dimensional ECG data into 3 dimensional and later to 1 dimensional and its PCA done. The output Pxy of Preprocessing unit is given to Feature extraction unit where is processed through Daubechies wavelet of the order 3 and level varying from 3 to 15. Each time one level of decomposition is selected. The statistical features Mean, Max, Min, Mode, Var and Std. deviation were extracted from approximate and detail coefficients. These features are the outputs of Feature extraction stage those are fed to Exh. CHAID classifier.

A. Data Collection

Some of the researchers used either self-collected ECG data or MIT-BIH database for their experimentation. In this paper ECG data is collected from a private pathological lab for 25 subjects that were recorded online under various conditions and states of patients as per their record. The 12-lead ECG recording consist of three limb leads I, II, and III, three augmented leads avL, avR and avF, and six precordial leads V1, V2, V3, V4, V5, V6. ECG for 25 subjects for was used for analyzing the performance of Exh. CHAID classifier.

B. Preprocessing of ECG Signal

At first 12 lead ECG data has been used to convert to 3 lead data. Further 3 lead data has been processed through PCA and

then to 1 dimensional data vector. The following steps demonstrates the process of conversion.

Step (i): Conversion of 12 lead ECG to 3 lead data

Synthesis of three orthogonal frank leads Vx, Vy, Vz by known mathematical transformations involving the high resolution recording of 12 standard ECG leads is as follows;

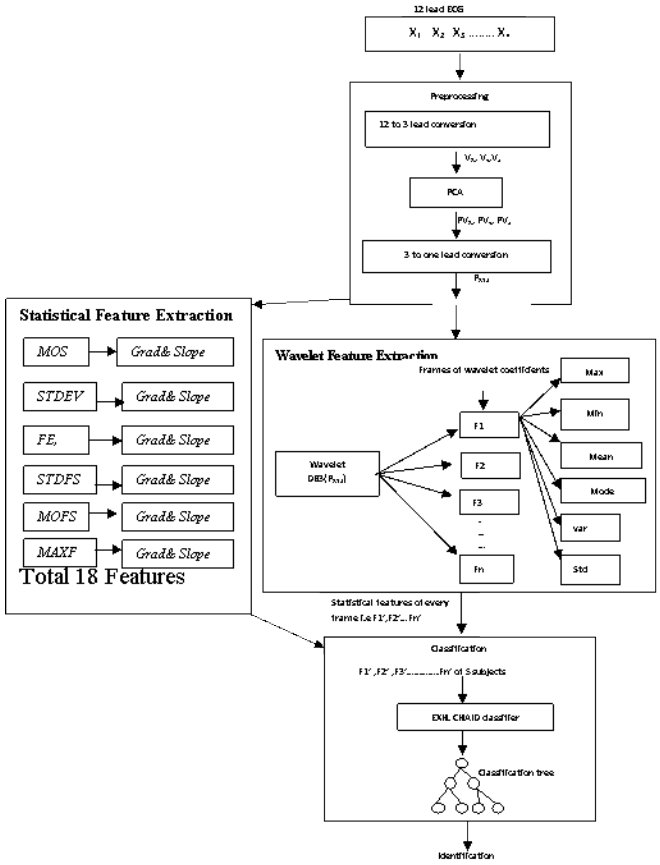


Fig.3. Architectural Diagram of Patient Identification System

Vx = 0.4*II-0.8*(II+III)/3+0.2*V5+0.5*V6+0.1*V4
Vy = 0.3*III+0.8*II+0.5*(II+III)/3-0.2*V5-0.3*V6
Vz=-0.1*III-0.2*II+0.4*(II+III)/3-0.3*V1-0.1*V2-0.1*V3-0.2*V4-0.1*V5+0.4*V6

The variables are described in previous subsection.

Step (ii): Principal Component Analysis

Applied PCA on above three frank leads (Vx, Vy, Vz) to convert correlated variables to uncorrelated variables. These modified three vectors were denoted by PVx, PVy and PVz.

Step (iii): Conversion of 3 lead data to 1 lead data

The modified 3 frank leads values PV_X, PV_Y, PV_Z, are then converted into 1 dimensional data vector by calculating spatial magnitude (SM) as follows:

$$P_{XYZ} = \sqrt{PVx^2 + PVy^2 + PVz^2}$$

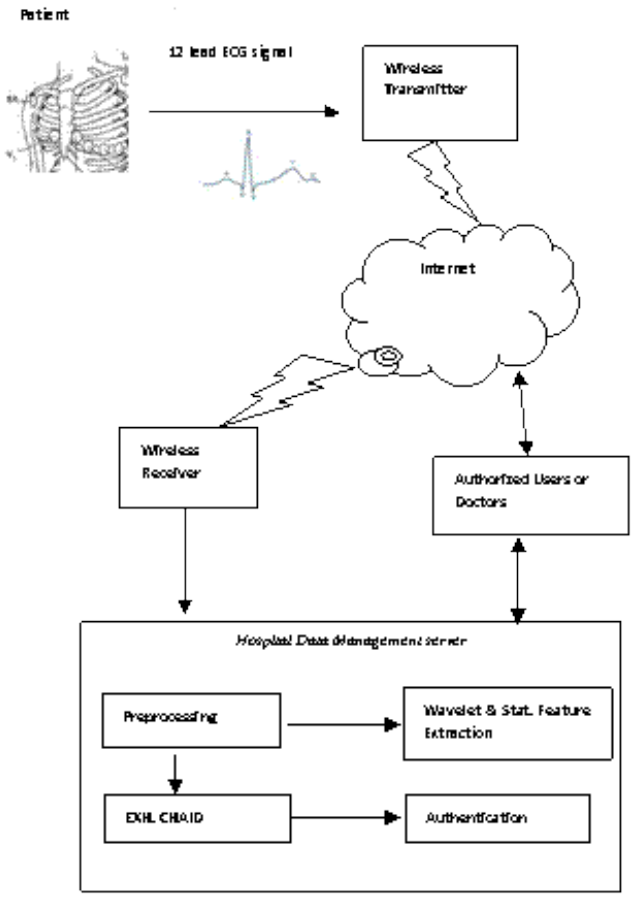


Fig.2. Architectural Diagram of Remote Health Monitoring System

C. Feature Extraction

To analyze the uniqueness of ECG signals recorded wavelet features were extracted using Daubechie's mother wavelet of the order 3 and decomposition levels from 6 to 15 for 25 subjects. While applying wavelet feature extraction method, 1 dimensional data vector was divided into 119 frames each of 500 sample values. These decompositions fetched approximate and detail coefficients depending upon the level of decomposition. Wavelets are obtained from a single prototype wavelet Ψ(t) called mother wavelet by dilations and shifting:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi(\frac{t-b}{a})$$

Where a is the scaling parameter and b is the shifting parameter. The 1 dimensional ECG data was divided into small frames and processed through Daubechies mother wavelets. The decomposed signal was reconstructed and the detail coefficients and approximate coefficients obtained through decomposition and for reconstructed waves were used to determine statistical parameters such as :

- Maximum (Max)
- Minimum (Min)
- Standard deviation (Std)

- Mean (Mean)
- Variance (Var)
- Mode (Mode)

For example if the signals are treated with Daubechies order 3 (DB3) with Level 6 decomposition, then while decomposition - approximate coefficient CA6, and detail coefficients CD1-CD6 were obtained. Similarly while reconstructing components A6, and D1-D6 were obtained. Max, Min, STD, Mean, VAR and Mode were calculated for these 14 feature vectors leading to 84 values for each frame. To study the relationships between various feature vectors so found, scatter plots were observed for selection of classifier. Figure 4 shows relationship between CD2MIN and CA7MODE in terms of scatter plot for 5 subjects. This plot shows that there is overlapping of feature spaces of five subjects.

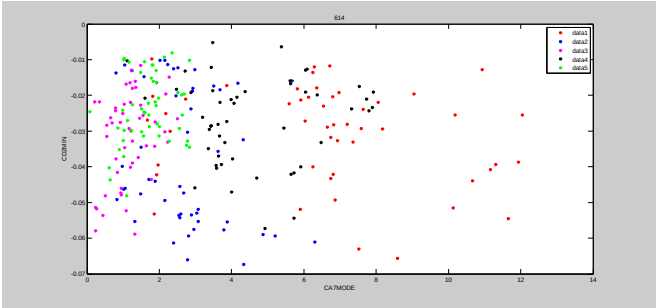


Fig.4. Scatter Plot for CD2Min Vs CA7Mode for Five Subjects

In addition to above wavelet features 18 statistical features were also derived for each frame. Maximum of Signal (MOS), Standard Deviation (SD) of Signal (STDEVS), Frame Energy (FE), Maximum of Frequency Spectrum (MAXF), Standard Deviation (SD) of Frequency Spectrum (STDFS), and Mean of Frequency Spectrum(MOFS), gradient of these parameters and their slopes. The details of estimation of gradient and slope are illustrated ahead.

- *Gradient*: The gradients of above six features were computed over each frame assuming a_1, a_2, a_3... a_n as sample values for n number of frames corresponding to the *MOS*. The gradients were defined as Δa_1, Δa_2, Δa_3... $\Delta\ a_n$ such that $\Delta a1 = (a_1-a_2)/a_1$, $\Delta a2 = (a_2-a_3)/a_2$ *and so on...* $\Delta\ a_{(n-1)} = (a_{(n-1)}-a_n)/a_{(n-1)}$ implied the indicator of relative change in the feature with reference to its initial value.
- *Slope* **:** The slopes of the these features were obtained as, $S_{a12} = (a_2-a_1)/(t_2-t_1)$, $S_{a23} = (a_3-a_2)/(t_3-t_2)$... where, S_{a12}represents the slope of MOS for its values for frame 1 and 2 (a_1 and a_2) and t_1 is the initial time of frame 1 and t_2 is the final time of frame 1. Similarly, t_2 is the initial time of frame 2 and $t_3$is the final time of frame 2, and so on. As the duration of each frame is $2\ s$, $t_2-t_1=t_3-t_2=...=2\ s$.

Therefore for Level 6 decomposition and other statistical features for one subject could fetch 119x66 feature matrix that was processed in combination of 5, 10, 15, and 25 subjects to test the performance of Exh. CHAID classifier. The input feature matrix for various analysis carried out is shown in Table I.

TABLE.1. Datasets of Varying Level of Decomposition and Number of Subjects used

No of subject		5	10	15	20
Level 6	Fea Mat	595x102	1190x102	1785x102	2975x102
	Obs	595	1190	1785	2975
Level 7	Fea Mat	595x114	1190x114	1785x114	2975x114
	Obs	595	1190	1785	2975
Level 8	Fea Mat	595x126	1190x126	1785x126	2975x126
	Obs	595	1190	1785	2975
Level 9	Fea Mat	595x138	1190x138	1785x138	2975x138
	Obs	595	1190	1785	2975
Level 10	Fea Mat	595x150	1190x150	1785x150	2975x150
	Obs	595	1190	1785	2975
Level 11	Fea Mat	595x162	1190x162	1785x162	2975x162
	Obs	595	1190	1785	2975
Level 15	Fea Mat	595x210	1190x210	1785x210	2975x210
	Obs	595	1190	1785	2975

(Fea Mat-Feature Matrix, Obs-Observations)

D. Exhaustive Chi-Square Automatic Interaction Detector (CHAID) as a Classifier

CHAID is a classification technique or an algorithm to study the relationship between dependent variable and a series of predictor variables. It uses merging, splitting and stopping processes. A tree is grown by repeatedly using these three steps on each node starting from the root node. Splitting and stopping steps in Exhaustive CHAID algorithm are the same as those in CHAID. Merging step uses an exhaustive search procedure to merge any similar pair until only single pair remains. Following are the algorithmic steps defined for merging, splitting and stopping processes used in Exh. CHAID algorithm.

i. Step 1: Merging

This step uses a exhaustive search procedure to merge similar pairs until a single pair remains. The merging steps are as follows:

1. If the independent variable X has one category then set p=1.
2. Set index=0. Calculate the p value which is based on the set of categories and p (index) = p (0).
3. Else calculate the p value and find the pair's of categories of X which have the largest p value. (The similar P value shows the most similar pair's of categories of X).
4. Merge the pair's that give the largest p value.
5. Update index = index + 1.
6. Repeat 3 to 6 until two categories remain then find the set of categories such that p (index) is smallest.
7. The adjusted p value is computed by applying Bonferroni adjustments.

ii. Step 2: Splitting

This step selects which predictor to be used to best split the node and it is dependent upon the p Value which is calculated in merging step.

1. Select the predictor that has smallest p Value.
2. If this p value is less than or equal to a user specified level then split the node using this predictor. Else do not split the node and consider it as a terminal node

iii. Step 3: Stopping

1. If a node becomes pure; that is, all cases in a node have identical values of the dependent variable, the node will not be split.
2. If all cases in a node have identical values for each predictor, the node will not be split.
3. If the current tree depth reaches the user specified maximum tree depth limit value, the tree growing process will stop.
4. If the size of a node is less than the user-specified minimum node size value, the node will not be split.
5. If the split of a node results in a child node whose node size is less than the user- specified minimum child node size value, child nodes that have too few cases as compared with this minimum, will merge with the most similar child node as measured by the largest of the p-values. However, if the resulting number of child nodes is 1, the node will not be split. If in a node all the values of dependent variable are identical then the node will not be split.

Results and Discussion

The parameters for Exh-CHAID were selected as:
Measure: Likelihood
Maximum tree depth: 10
Minimum parent size: 2 / Minimum son size: 1
Number of intervals: 10

TABLE.2. Percentage Identification Accuracy for Varying Decomposition Level and no. of Subjects

from \ to	1	2	3	4	5	6	7	8	9	10	Total	% correct
1	86	20	3	0	0	4	2	2	1	1	119	72.27%
2	1	111	0	1	0	0	4	0	1	1	119	93.28%
3	0	7	99	0	5	1	1	2	3	1	119	83.19%
4	2	30	0	73	3	0	7	0	2	2	119	61.34%
5	3	18	1	1	67	3	5	15	6	0	119	56.30%
6	3	3	4	1	2	97	1	6	2	0	119	81.51%
7	0	5	3	0	3	2	94	2	7	3	119	78.99%
8	0	15	0	0	10	3	2	86	3	0	119	72.27%
9	1	12	2	2	0	3	2	11	86	0	119	72.27%
10	4	26	2	4	9	1	13	2	0	58	119	48.74%
Total	**100**	**247**	**114**	**82**	**99**	**114**	**131**	**126**	**111**	**66**	**1190**	**72.02%**

The proposed identification system has been analyzed for 5, 10, 15, 25 subjects with Daubechieswavelet of order 3 and level variation of 6, 7, 8, 9, 10, 11 and 15. The detection accuracy is varying with number of subjects as well as level of decomposition as shown in Table II. It can be seen that the best identification accuracy of 100 % was obtained at level 15 of decomposition corresponding to 5 subjects. For each level of decomposition it can be seen that the accuracy decreases with increase in number of subjects. The lowest level of decomposition under test level 6 could provide maximum of 76.64 % accuracy.

The output of each of the experimentation was analyzed using confusion matrix of the classifier. Table III shows the confusion matrix for 5 subjects. Looking at the table it is clear that for each subject the detection accuracy is 100 %.

TABLE.3. Confusion Matrix with 5 subjects at Decomposition Level 15

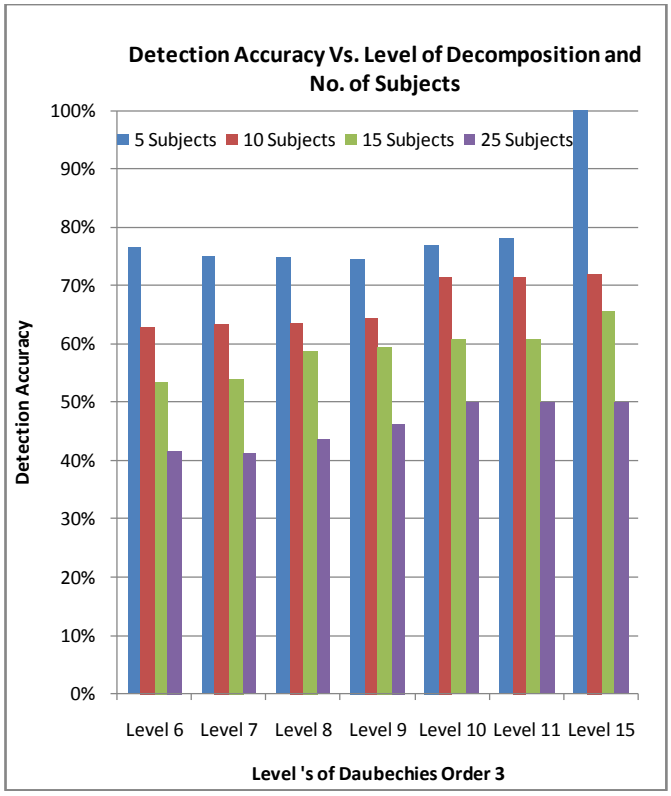
S.No	Noof sub	Level 6	Level 7	Level 8	Level 9	Level 10	Level 11	Level 15
1	**5**	76.64%	74.96%	74.79%	74.62%	76.89%	77.98%	100%
2	**10**	63.03%	63.45%	63.70%	64.54%	71.43%	71.51%	72.02%
3	**15**	53.53%	53.87%	58.73%	59.47%	60.76%	60.71%	65.49%
4	**25**	41.75%	41.28%	43.72%	46.25%	49.95%	50.05%	49.92%

From Table IV confusion matrix for 10 subjects at decomposition level 15, it can be seen that the best accuracy could be found for subject 2 with 93.28 % value, while the worst for subject 10 with 48.74 % value. It can further be seen from confusion matrix that total subject 2 detections has been highest value of 247, indicating that the classifier is most inclined towards subject 2 class. The minimum number of correct detections is 66 for subject 10, indicating that the classifier is least inclined towards class 10.

TABLE.4. Confusion Matrix with 10 Subjects at Decomposition Level 15

From to	1	2	3	4	5	Total	% correct
1	119	0	0	0	0	119	100.00%
2	0	119	0	0	0	119	100.00%
3	0	0	119	0	0	119	100.00%
4	0	0	0	119	0	119	100.00%
5	0	0	0	0	119	119	100.00%
Total	**119**	**119**	**119**	**119**	**119**	**595**	**100.00%**

The Fig.5 below shows bar chart for various patient identification accuracies obtained for all experimental variations at a glance.



Conclusion

In the proposed system of remote patient authentication with detailed discussion and analysis of patient identification system using 12 lead ECG, showed that the proposed methodology of Exh. CHAID with wavelet features in combination with statistical features provide 100% detection accuracy if examined for less number subjects at tree depth selected as 10. In future work robustness of the system needs to be tested by changing the feature sets and the tree depth.

References

[1] L. Biel, O. Pettersson, L. Philipson, and P. Wide, ”ECG analysis: a new approach in human identification, ” Proceedings of the 16th Instrumentation and Measurement Technology Conference, IEEE, volume1, pp.808-12, 2001.

[2] T. W. Shen, W. J. Tompkins, and Y. H. Hu, ” One-lead ECG for identity verification”:Proceedings of the 24th Annual Conference on Engineering in Medicine and Biology and the Annual Fall Meeting of the Biomedical Engineering Society, “ Volume 57, issue 2, pp 62-3, 2002.

[3] Steven A. Israel, Jhon M.I, Andrew C., Mark, D.W, ”ECG to identify individuals, ” Virtual reality medical center, USA, Pattern Recognization, Volume 38, issue1, pp 133-142, 2005.

[4] Y. Wang, F. Agrafioti, D. Hatzinakos, and K. N. Plataniotis, ”Analysis of human electrocardiogram for biometric recognition, ” EURASIP Journal on Advances in Signal Processing, 2008.

[5] D. C. Chan, M. M. Hamdy, A. Badre, and V. Badee., ”Wavelet distance measure for person identification using electrocardiograms”, Transactions on Instrumentation and Measurement, IEEE, Volume 57, issue 2, pp. 248-253, 2008.

[6] C. Chiu, C. Chuang, and C. Hsu, “A novel personal identity verification approach using a discrete wavelet transform of the ECG signal, ” Proceedings of the International Conference on Multimedia and Ubiquitous Engineering, issue 4, Volume 6, pp. 201-206, 2008.

[7] Janani S., Minho S., Tanzeem C., David K., ”Activity-aware ECG-based patient authentication for remote health monitoring, ” International Conference on Mobile Systems, ACM, Nov.2009

[9] Can Y., Miguel C., B.V.K Vijaya k, “Investigation of human identification using two lead electrogram signals, ” Proceedings of the 4th International Symposium on Applied Sciences in Biomedical and Communication Technologies, IEEE, Volume 50, pp.1-8, 2010.

[10] M. M. T. Abdelraheem, HanySelim, Tarik Kamal Abdelhamid, ”Human Identification Using the Main Loop of the Vectorcardiogram, ” American Journal of Signal Processing, pp.23-29, 2012.