Feature Normalization for Enhancing Early Detection of Cardiac Disorders

Swati Negi*, C. Santhosh Kumar*, A. Anand Kumar$
*Machine Intelligence Research Lab, Department of Electronics and Communication Engineering,
Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, Amrita University, India - 641112
Email: swatinegi04@gmail.com, cs_kumar@cb.amrita.edu
$Department of Neurology, Amrita Institute of Medical Science, Elamakkara, Kochi, Kerala - 682041

***Abstract*—Early detection of cardiac disorders can help save many lives. Time and frequency domain statistical features derived from RR interval series of electrocardiogram (ECG) signals with a support vector machine (SVM) backend classifier can be used for distinguishing congestive heart failure (CHF) and sudden cardiac death (SCD) patients from the normal sinus rhythm (NSR) patients. We empirically found that ninety minutes of duration gave the optimal classification results after exploring with different heart rate variability (HRV) time durations.**

We obtained a classification accuracy of 92.85% for our baseline system using linear SVM kernel. In this work, the input statistical features consists of patient independent and patient specific variations. The patient specific variations were considered as noise in the input feature vector, while patient independent variations as informative.

In this work, we experimented with two approaches. The first approach used was principal component analysis (PCA) to obtain dimensionality reduced features with maximum information stored. We obtained a performance improvement of 0.65% absolute over the baseline system. In the second approach, covariance normalization (CVN) was used to remove/minimize the effect of patient specific variations. The overall system performance was improved by 1.96% absolute over the baseline system.

I. Introduction

Cardiac disorders such as, congestive heart failure (CHF) and sudden cardiac death (SCD), are taking an exponential growth in almost all the countries. Myocardial infarction [1] is a condition which impairs the heart's ability to circulate blood eventually leading to heart failure. SCD condition arises when the heart suddenly stops functioning. In such condition, the blood flow to the brain and other vital organs is restricted. Such problems demand accurate methods for the early detection of cardiac disorders. Heart disease turns out to be the leading causes of mortality in the world, accounting for 17.3 million deaths per year, which is expected to grow more than 23.6 million by 2030 [2]. Cardiac transplantation is an adopted treatment for the end-stage heart failure patients when medical treatment and less drastic surgery have failed [3]. Hence, early stage diagnosis play a critical role in preserving and maintaining heart health.

Electrocardiography (ECG) is a mainstream diagnostic technique used for recording the electrical activity of heart. Heart rate variability (HRV) is a physiological phenomenon which occurs mainly due to variation of beat-to-beat alterations in heart rate, which can be obtained from ECG. It is the most promising non-invasive markers representing the variation of RR intervals over time and reflecting the activity of the autonomic nervous system (ANS) [4], [5], [6]. HRV gives the necessary information about the patient's heart condition. A person with lower heart rate (HR) shows higher HRV, in turn indicating high power levels in the higher frequency components of HRV and vice-versa.

Estimating the consequences of a disease at an early stage is one of the most challenging task. In order to solve such problems related to diagnosis of disorders, latent information from the available datasets can be extracted. For the past a few years knowledge discovery in databases (KDD) including machine learning algorithms , has gained its popularity as one among the best research tool for the medical researchers. The best application of such algorithms gives a solution which is used more often for making use of patterns and relationships among large scale of variables. It is also used to foretell the final outcome of a disease using the information kept within the datasets [7]. It is more than the data analysis which includes classification, clustering, and association rule discovery. This work is focused on the early stage detection of cardiac disorders by segmenting HRV data into shorter durations and developing the machine learning algorithms along with a SVM backend classifier. SVM is a supervised learning model and its performance can be further improved by using data generated from different machine learning algorithms. Due to its higher generalization ability, SVM can classify unseen data accurately [8].

In previous studies [9], [10], [11], [12], [13], ambulatory analysis has been done to monitor the condition of cardiac patients. However, in practical situations, patients feel uncomfortable when analysis is performed for longer durations. Hence, we focus on analyzing the condition of the patients through shorter duration HRV segments using support vector machine (SVM) approach. Early stage diagnosis of cardiac disorders [14] requires prediction of their progression which indicates the susceptibility of patients. In this study, we have developed linear SVM model to classify CHF and SCD patients from NSR patients for the optimal duration of ninety minutes.

In this work, we used principal component analysis (PCA) [15], [16], [17], [18] and used it with SVM backend classifier (PCA-SVM). PCA projects the data in the direction of maximal variance. PCA-SVM system turns out to be very effective when the small dataset is used [19]. In order to improve the overall performance of the baseline system, covariance normalization (CVN) [20] technique was exploited, considering the patient-specific variations in the input feature vector as noise. It was seen that CVN-SVM system helped enhance the performance of the early stage detection system when compared to the baseline system and PCA-SVM system.

II. System Description

In practical situations, preemptive approach is necessary for the early stage detection of cardiac disorders. Hence, in this work, we segmented long-term duration of HRV data upto fifteen minutes duration and empirically ninety minutes is found to be the optimal time duration. Segmentation of data into smaller duration helps in increasing the number of examples per patient and to capture the variations of different time intervals. The details of the baseline system is explained in Fig. 1. The sequence of steps followed are: (i) RR interval series has been obtained for three different classes such as CHF, SCD and NSR patients from physionet-ECG database [21]. (ii) HRV data has been segmented into shorter durations. (iii) HRV segments that were noisy and could distort the analysis were cut out of the sample and discarded. (iv) Linear analysis of HRV signal has been done to get the selected time-domain and frequency-domain HRV features respectively which are discriminative in nature. (v) Time and frequency domain statistical features extracted from RR interval series has been provided as an input to linear SVM backend classifier. (vi) PCA used as a dimensionality reduction technique which is important due to cost of calculation and accuracy in classification. PCA-SVM system used in this work is described in Fig. 2 (vii) CVN as a feature normalization technique has been applied to further reduce the effect of patient-specific factors by considering them as noise. The description of CVN-SVM system is provided in Fig. 3.

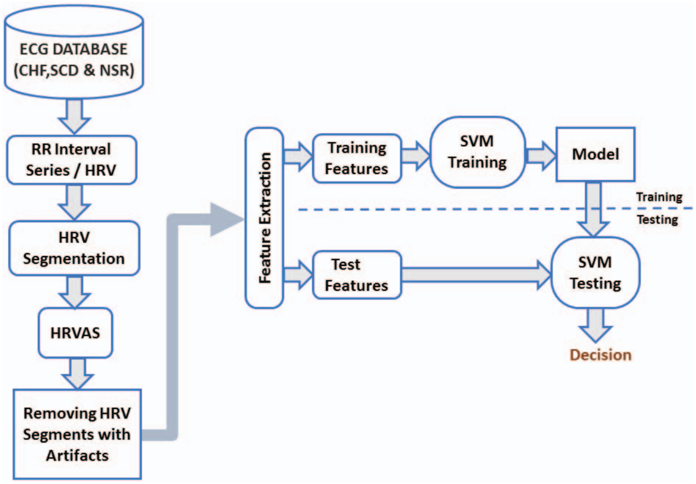

Fig. 1. Baseline System

A. Pre-processing of HRV signal

Two crucial observations has been made on optimal baseline linear SVM. We observed that there are artifacts in the data segments leading to poor classification of CHF, SCD and NSR patients. Artifacts can be removed either by : (i) Ectopic intervals detection and removal/replacement of the same by mean, median or cubic spline method in the particular segment. (ii) By discarding the entire segment containing artifacts. The baseline system performance has been improved by using the latter approach to remove artifacts.

B. Heart rate variability(HRV) analysis

Time-domain and frequency-domain statistical features has been extracted (https://github.com/jramshur/HRVAS) from inter-beat interval (IBI) or RR interval series and power spectrum analysis has been performed.

1) Linear analysis: Linear analysis of HRV signal includes extraction of time-domain and frequency-domain statistical features as:

- Time-domain statistical features

1) Mean = Mean of IBI series.
2) Median = Median of IBI series.
3) SDNN = Standard Deviation of NN intervals.
4) SDANN = Computes the mean IBI of each segment and then returns the Standard Deviation of all means as:

$$SDANN = \sqrt{\frac{1}{M-1}\sum_{i=1}^{M}[meanIBI(i) - meanIBI(i)^2]} \tag{1}$$

5) NNx = Number of successive differences that are greater than 50 msec.
6) pNNx = Percentage of total intervals that successively differs by more than 50 msec.
7) RMSSD = Root Mean Square of the successive differences of the IBI series.
8) SDNNi = Computes by finding the Standard Deviation of each IBI segment and then returning the mean value of standard deviations. (i=1minute)

$$SDNN_i = \frac{1}{M}\sum_{i=1}^{M} SDNN(i) \tag{2}$$

9) meanHR = Mean value of heart rates.
10) sdHR = Average standard deviation value of heart rates.
11) HRVTi = HRV Triangular Index (i=1 minute).

$$HRV_{t_i} = \frac{N_{IBI}}{Y} \tag{3}$$

12) TINN = Triangular interpolation of IBI histogram

- Frequency-domain statistical features (ms^2)

1) VLF Power = Spectral flux of entire NN intervals in 0.003-0.04 Hz.
2) LF Power = Spectral flux of entire NN intervals in 0.04-0.15 Hz.
3) HF Power = Spectral flux of entire NN intervals in 0.15-0.4 Hz.

4) Total Power = Gross spectral flux of entire NN intervals upto 0.4 Hz.

5) LF/HF Power = Ratio of low to high frequency spectral flux.

2) Power spectrum analysis: Heart rate variability (HRV) spectrum analysis has been done using Lomb-Scargle periodogram (LSP). LSP is acknowledged to be the best method when the input signal is highly non-stationary. As HRV is a highly time-varying and a non-stationary signal, hence, LSP satisfies the requirements of this work. Quantifying the fluctuations in heart rate (HR) within the IBI time series can be done by calculating the power spectrum density (PSD). The PSD presents spectral power density of a time series as a function of frequency. Therefore, PSD estimates can give information about the amount of power in which certain frequencies contribute to a time series. Total power is computed by integrating the PSD between the certain band frequency limits. The LSP of a non-uniformly sampled, real-valued data sequence X of length N for arbitrary times t_n is defined by :

$$p_{LS}(f) = \frac{1}{2\sigma^2} \frac{[\sum_{n=1}^{N}(X(t_n - \bar{X}))cos(2\pi f(t_n - \tau))]^2}{\sum_{n=1}^{N}\cos^2(2\pi f(t_n - \tau))} + \frac{[\sum_{n=1}^{N}(X(t_n - \bar{X}))sin(2\pi f(t_n - \tau))]^2}{\sum_{n=1}^{N}\sin^2(2\pi f(t_n - \tau))} \tag{4}$$

where $\bar{X}$ and $\sigma 2$ are mean and variance of the time series, and

$\tau \equiv tan^{-1}((\sum_{n=1}^{N} sin(4*\pi*f*t_n))/(\sum_{n=1}^{N} cos(4*\pi*f*t_n)))$, τ is a frequency dependent time delay, defined to make the periodogram insensitive to time shift. In this work, analysis of PSD estimates shows that there are higher power levels in the very low frequency components of SCD and CHF patients, indicating higher HR and lower HRV. PSD estimates of NSR patients has shown higher power levels in high frequency range, giving lower HR and hence higher HRV.

C. Dimensionality reduction technique

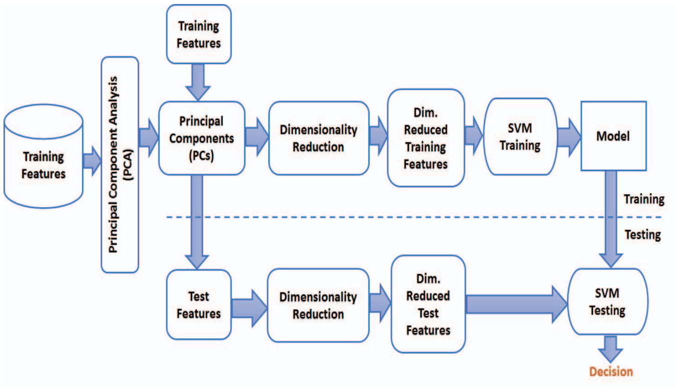

Fig. 2. PCA-SVM system

1) Principal component analysis (PCA): In PCA, data is projected towards the directions which are highly variable [16], [17], [18]. According to the decrease in the order of variability, the principal components (PCs) are computed as the basis vectors. Setting the threshold for a percentage of entire data variability, helps in selecting the number of PCs. PCs computation for the data includes estimation of covariance matrix (CVM). Subsequently, eigenvalue decomposition (EVD) and then sorting of eigenvectors (evecs) in the decreasing order of eigenvalues (evals). Finally, the data is projected into the new basis with the help of PCs defined by taking the scalar product of the original signals and the sorted evecs. When large multivariate datasets are analyzed, it is desirable to reduce their dimensionality. PCA thus helps in solving the eigenvalue problem as:

$$\lambda_j ev_j = V ev_j, j = 1, ..., n. \tag{5}$$

where j = one of the eigenvalues of CVM. ev_j = corresponding evec. Based on the estimated ev_j, the components of r_t are then calculated as the orthogonal transformations of y_t as :

$$r_t(j) = ev_j^T y_t, j = 1, ..., n. \tag{6}$$

The count of PCs in r_t can be reduced by using the first several evecs which are sorted according to the decreasing order of the evals. So PCA has the dimensional reduction characteristics.

PCA-SVM system methodology is illustrated in Fig. 2. While performing PCA , the covariance matrix is first computed from the training features. Evals and evecs are then obtained. In the decreasing sequence of evals, the sorting of evecs matrix has been done. Lastly, training features are projected into the newly formed basis determined by PCs. Projection is done by computing the scalar product of the original signals (training and test features) and the ordered evecs. Dimensionality reduced training and test feature vectors are obtained by reducing the dimensions of projected matrix. SVM backend classifier has been trained using this dimensionality reduced set of feature vectors.

D. Feature normalization technique

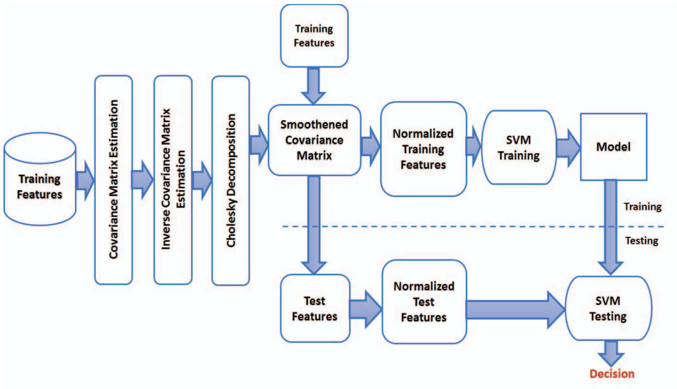

Fig. 3. CVN-SVM system

1) Covariance Normalization (CVN) : There is an intrinsic relationship between different features. Therefore, using covariance matrix (CVM) for normalization of the features would take such relationship into consideration, and would therefore be more accurate [20].

Usually the estimates obtained after covariance are noisy. Therefore, in order to compensate, smoothing of the covariance matrix helps in improving the performance as:

$$S = \lambda * R + (1 - \lambda) * I \tag{7}$$

where, S= Smoothened covariance matrix. λ = Smoothing factor. $(0 < \lambda < 1)$ R = Covariance matrix. I = Identity matrix.

The methodology for the feature normalization (CVN-SVM) system is shown in Fig. 3. CVN compensate the effect of noise due to patient-specific factors. CVM [22], [20] was obtained from the training features and then the inverse covariance matrix (ICVM) was computed. Cholesky decomposition [20] requires the matrix to be positive definite. Further, Cholesky decomposition was performed on the ICVM which gives an upper triangular matrix and the more significant features. Smoothing factor (λ) was varied in the range of 0 to 1 to find the smoothened covariance matrix (SCVM). In order to obtain a normalized set of feature vectors, SCVM was then multiplied with original training and testing features.

III. Experiments and Results

A. Subjects

In this work, subjects include three classes of data such as CHF, SCD and NSR patients retrieved from physionet-ECG database [21]. ECG database details include: (i) The BIDMC CHF Database : Long-term ECGs (about 20 hours each) from 15 subjects (11 men, aged 22 to 71, and 4 women, aged 54 to 63) with severe CHF(NYHA class 3-4). The ECG signal sampled at 250 samples per second with 12-bit resolution over a range of 10 millivolts using ambulatory ECG recorders with a typical recording bandwidth of approx. 0.1 Hz to 40 Hz. (ii) SCD holter database : This is a collection of 23 complete holter recordings with different signal durations who experienced sudden cardiac death during the recordings which includes 18 patients with underlying sinus rhythm (4 with intermittent pacing), 1 who was continuously paced, and 4 with atrial fibrillation. All patients had a sustained ventricular tachyarrhythmia, and most had an actual cardiac arrest. (iii) The MIT-BIH NSR database: It includes 18 (5 men, aged 26 to 45, and 13 women, aged 20 to 50) long-term ECG recordings.

B. HRV data segmentation

HRV data is segmented into shorter time-durations. Classification accuracy for different HRV time-durations is shown in Table I and Table II. Ninety minutes duration is found to gave the best classification results as 85.38% (146/171) and considered as a baseline system. After removal of segments containing artifacts and varying cost/regularization factor, classifier accuracy has further improved to 92.85% (143/154). Total 154 number of test examples among which 143 were correctly classified.

TABLE I
Classification accuracy(%) for different heart rate variability (HRV) time-durations

S.No.	Kernel	24 Hours	120 minutes	90 minutes
1.	**Linear**	66.66	79.80	**85.38**
2.	Polynomial	66.66	66.34	54.97
3.	RBF	41.66	39.42	39.76
4.	Sigmoid	41.66	39.42	39.76

TABLE II
Classification accuracy(%) for different heart rate variability (HRV) time-durations

S.No.	Kernel	60 minutes	30 minutes	15 minutes
1.	**Linear**	84.72	64.26	75.52
2.	Polynomial	62.03	55.68	51.74
3.	RBF	39.35	39.90	43.30
4.	Sigmoid	39.35	39.44	39.44

C. Performance of model based on dimensionality reduction

Reducing the number of dimensions to 16 and 15 gave the highest classification accuracy as 92.20% and 90.90% respectively as shown in Table III. Further, variation in cost/regularization factor (c=0.3) for the 15th dimension improved the overall classification accuracy to 93.50% (144/154).

TABLE III
Classification accuracy(%) for linear SVM backend classifier by varying different dimensions

S.No.	Dimension	Accuracy(%)	S.No.	Dimension	Accuracy(%)
1.	17	88.96	10.	8	77.27
2.	**16**	**92.20**	11.	7	76.62
3.	**15**	**90.90**	12.	6	71.42
4.	14	74.02	13.	5	69.48
5.	13	68.18	14.	4	69.48
6.	12	84.41	15.	3	57.14
7.	11	50.64	16.	2	66.23
8.	10	68.18			
9.	9	64.93			

D. Performance of model based on feature normalization

Feature normalization helped in the removal of patient-specific factors by considering them as noise and patient-independent factors were considered as significant features. Overall classification accuracy has significantly increased by 1.95% absolute by varying smoothening factor (λ) and cost/regularization factor (c=0.2) as shown in Table IV. Hence, covariance normalization outperforms the baseline and PCA-SVM system.

TABLE IV
Overall classification accuracy(%) for linear SVM backend classifier after feature normalization

S.No.	Lambda (λ)	Accuracy(%)
1.	0.0024	93.50
2.	0.000195	94.15
3.	0.000342	**94.80**

IV. Conclusion

Cardiac disorder is one of the challenging problems in the medical field as it affects both heart and circulatory system. Usually these studies are carried out over a period of twenty four hours in an ambulatory environment. A longer duration analysis often discourages the patient from taking up such an analysis, and also makes them unsuitable for emergency situations.

In this work, we explored the effectiveness of heart rate variability (HRV) over a short period of ninety minutes for the early detection of cardiac disorders. Our backend classifier for the baseline system is a linear support vector machine (SVM). In an effort to improve the classification accuracy, we first experimented using principal component analysis (PCA) which is a dimensionality reduction technique. We obtained a performance improvement of 0.65% absolute through PCA. Secondly, covariance normalization (CVN) of the features was explored in an effort to minimize the effect of patient dependent variations in the input feature vectors. It was seen that CVN has helped improve the performance of the system by 1.96% absolute over the baseline system.

The proposed system can be integrated into a mobile phone for the early detection of cardiac disorders in emergency situations making it more affordable to a common man. For the early stage diagnosis of heart diseases, health care workers, nurses and technicians who provide care for cardiology patients will find the proposed system very useful.

Acknowledgment

The authors would like to thank Aparna S Raghunath, Sreekumar K. T., Arun Das K, Kuruvachan K. George, R Gopinath, Rajevencelta and Arun Kumar S of Machine Intelligence Research laboratory (MIRL), Amrita Vishwa Vidyapeetham University, Ettimadai, Coimbatore for their support through out this work.

References

[1] A. J. Moss, W. Zareba, W. J. Hall, H. Klein, D. J. Wilber, D. S. Cannom, J. P. Daubert, S. L. Higgins, M. W. Brown, and M. L. Andrews, "Prophylactic implantation of a defibrillator in patients with myocardial infarction and reduced ejection fraction," *New England Journal of Medicine*, vol. 346, no. 12, pp. 877–883, 2002.

[2] D. Mozaffarian, E. J. Benjamin, A. S. Go, D. K. Arnett, M. J. Blaha, M. Cushman, S. R. Das, S. de Ferranti, J.-P. Després, H. J. Fullerton *et al.*, "Heart disease and stroke statistics2016 update a report from the american heart association," *Circulation*, pp. CIR–0 000 000 000 000 350, 2015.

[3] M. Metra, P. Ponikowski, K. Dickstein, J. J. McMurray, A. Gavazzi, C.-H. Bergh, A. G. Fraser, T. Jaarsma, A. Pitsis, P. Mohacsi *et al.*, "Advanced chronic heart failure: a position statement from the study group on advanced heart failure of the heart failure association of the european society of cardiology," *European journal of heart failure*, vol. 9, no. 6-7, pp. 684–694, 2007.

[4] M. S. Manikandan and K. P Soman, "A novel method for detecting r-peaks in electrocardiogram (ecg) signal," *Biomedical Signal Processing and Control*, vol. 7, no. 2, pp. 118–128, 2012.

[5] H. ChuDuc, K. NguyenPhan, and D. NguyenViet, "A review of heart rate variability and its applications," *APCBEE Procedia*, vol. 7, pp. 80–85, 2013.

[6] R. Metelka *et al.*, "Heart rate variability-current diagnosis of the cardiac autonomic neuropathy. a review," *Biomedical Papers*, vol. 158, no. 3, pp. 327–338, 2014.

[7] J. Qiu, Q. Wu, G. Ding, Y. Xu, and S. Feng, "A survey of machine learning for big data processing," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, pp. 1–16, 2016.

[8] K. P Soman, R. Loganathan, and V. Ajay, *Machine learning with SVM and other kernel methods*. PHI Learning Pvt. Ltd., 2009.

[9] R. M. Carney, J. A. Blumenthal, P. K. Stein, L. Watkins, D. Catellier, L. F. Berkman, S. M. Czajkowski, C. OConnor, P. H. Stone, and K. E. Freedland, "Depression, heart rate variability, and acute myocardial infarction," *Circulation*, vol. 104, no. 17, pp. 2024–2028, 2001.

[10] P. Zimetbaum and A. Goldman, "Ambulatory arrhythmia monitoring choosing the right device," *Circulation*, vol. 122, no. 16, pp. 1629–1636, 2010.

[11] L.-M. Liao, S. S. Al-Zaiti, and M. G. Carey, "Depression and heart rate variability in firefighters," *SAGE open medicine*, vol. 2, p. 2050312114545530, 2014.

[12] T. A. McDonagh, R. S. Gardner, M. Lainscak, O. W. Nielsen, J. Parissis, G. Filippatos, and S. D. Anker, "Heart failure association of the european society of cardiology specialist heart failure curriculum," *European journal of heart failure*, vol. 16, no. 2, pp. 151–162, 2014.

[13] F. Shahbazi and B. M. Asl, "Generalized discriminant analysis for congestive heart failure risk assessment based on long-term heart rate variability," *Computer methods and programs in biomedicine*, vol. 122, no. 2, pp. 191–198, 2015.

[14] E. S. Ketchum and W. C. Levy, "Establishing prognosis in heart failure: a multimarker approach," *Progress in cardiovascular diseases*, vol. 54, no. 2, pp. 86–96, 2011.

[15] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2002.

[16] N. Kambhatla and T. K. Leen, "Dimension reduction by local principal component analysis," *Neural Computation*, vol. 9, no. 7, pp. 1493–1516, 1997.

[17] B. Nie, J. Du, H. Liu, G. Xu, Z. Wang, Y. He, and B. Li, "Crowds' classification using hierarchical cluster, rough sets, principal component analysis and its combination," in *Computer Science-Technology and Applications, 2009. IFCSTA'09. International Forum on*, vol. 1. IEEE, 2009, pp. 287–290.

[18] R. Kottaimalai, M. P. Rajasekaran, V. Selvam, and B. Kannapiran, "Eeg signal classification using principal component analysis with neural network in brain computer interface applications," in *Emerging Trends in Computing, Communication and Nanotechnology (ICE-CCN), 2013 International Conference on*. IEEE, 2013, pp. 227–231.

[19] A. M. Martínez and A. C. Kak, "Pca versus lda," *IEEE transactions on pattern analysis and machine intelligence*, vol. 23, no. 2, pp. 228–233, 2001.

[20] C. S. Kumar, K. I Ramachandran, and A. Kumar, "Vital sign normalisation for improving performance of multi-parameter patient monitors," *Electronics Letters*, vol. 51, no. 25, pp. 2089–2090, 2015.

[21] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000 (June 13), circulation Electronic Pages: http://circ.ahajournals.org/cgi/content/full/101/23/e215 PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.

[22] R. Rojas, "The secret life of the covariance matrix," *Freie Universität Berlin [online], URL: http://www. inf. fu-berlin. de/inst/ag-ki/rojas_home/documents/tutorials/secretcovariance. pdf [cited 1 October 2012]*, 2009.