Classification of Individual and Combined Finger Movements Using Electromyogram (EMG) Signals

(Prof. AMITAVA CHATTERJEE)

- *Soumyadeep Basak (001810801132)*
- *Souvik Mandal (001810801135)*
- *Mrinmoy Barman(001810801155)*
- *Prashant Giri(001810801161)*
- *Saurabh Shit(001810801165)*

Toward improved control of prosthetic fingers using surfac lectromyogram (EMG) signals

ftnmi *N. Khushaba , Sarath Kodagoda, Caen Takrvri,* Comini *Dissonayoke*

DESCRIPTION:

A fundamental component of many modern prostheses is the myoelectric control system, which uses the electromyogram (EMG) signals from an individual's muscles to control the prosthesls movements. Though this research rigorously focus on the myoelectric control of arm and gross hand movements, more dexterous individual and combined fingers control has not received the same attention. The main contribution of this paper ts an invest 8ation into accurately discriminating between individual and combined fingers movements using surface EMG signals, so that different finger poses of a prosthetic hand can be controlled in response. For this purpose, two EMG electrodes located on the human forearm are utilized to collect the IMG data from eight panicipants. Various feature sets are extracted and potrayed in a

manner that ensures maximum separation between the finger movements and then fed to two different classifiers. The second contribution is the use of a Bayesian data fusion post-processing approach to maximize the probability of correct classification of the Et 'IG data belonging to different movements. Practical results and significant statistical tests prove the feasibility of the proposed approach with an average

CldSSification accuracy of =90% across different subjects proving the significance of the proposed fusion scheme in finger movement Classification.

OUR WORK:

Now, for detecting individual and combined finger movements ,we have obtained the raw dat from these 2 channels and then classified it into 10 classes using 2 steps

1. Dimenslonality Reduction using PCA,LDA.

2. Classlflcat\on uslngANN,SVld.

pErlNlTlop o_FJz_nNS;

Principal Component Analysis (PCA):

h Is a statistical procedure that uses an orthogonal transformation to convert a set observations of possibly correlated varlables Into a set of values of uncorreIatedvariab called principal components.

Basically, PCA is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

Reducing the number of variables of a data set naturally comes at the expense of accuracy, but the trick in dimensionality reduction is to trade a little accuracy forsimplicity. Because smaller data sets are easier to explore and visualize and make analyzing data much easier and faster for machine learning algorithms without extraneous variables to process. So to sum up, the idea of PCA is simple — reduce the number of variables of a data set, while preserving as much information as possible.

The whole process of obtaining principle components from a raw data-set can be simplified in six parts :-

- Take the whole data-set consisting of d+1 dimensions and ignore the labels such that our new data-set becomes d dlmenslonal,

- Compute the mean for every dimension of the whole data-set.

» **Compute the covariance matrix of** the **whole** data-set **using** the **formula:**

$$\text{cov(X,Y)=}\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{X})(y_i-\overline{Y})$$

It is nothing but basically a table that summaries the correlations between all the possible pairs of variables.

- Compute eigenvectors and the corresponding elgen values.

« Sort the eigenvectors by decreasing eigen values and choose k eigenvectors with the largest
eigenvalues to form a d x k dimensional matrix W.

- Use this d x k eigenvector matrix to transform the samples onto the new subspace.

However it is necessary to perform standardization prior to PCA because if there are large differences between the range of initial variabIes,those variables with larger ranges will dominate over those with small ranges which might lead to biased results. So transforming the data to comparable scales can prevent this problem. Mathematically this can be done by subtracting the mean and dividing by the standard deviation for each value of each variable.

Sometimes variables are highly correlated in such a way that they contain redundant information. So to identify these correlations we compute the covariance matrix.

Now,the covariance matrix iS a d «d symmetric matrix (where *d* is the number of dimensions) that has as entries the co-variances associated with all possible pairs of the initial variables.

Since the covariance of a variable w'ch itself is its variance (Cov(a,a)=Var(a)), in the main diagonal (Top left to bottom rlght) we actually have the variances of each initial variable. And since the covariance is commutative (Cov(a,b)=Co v(b,a)), the entries of the covariance matrix are symmetric with respect to the main diagonal, which means that the upper and the lower triangular portions are equal.

What do the co-variances that we have as entries of the matrix tell us about the correlations between the variables?

It’s actually the sign of the covariance that matters :

- If positive then : the two varlables increase or decrease together (correlated)
- **If negative then** : **One increases when the other decreases (Inversely correlated)**

So, the Idea Is If we are having 10-dimensional data, then it gives us 10 principal components, but the information **withln the initial variables Into the first components, then maximum remaining information in the second component and so on.** But the new vafiables being mixtures of initial variables are combined in such a way that they are uncorrelated.

2. Linear Discriminant Anal is (LDA)

Linear Discriminant Analysis is a dimenslonallty reduction technique used as a preprocessing step in Machine Learning and pattern **classification appllcatlons.**

The main *goal* of dimensionality reduction techniques is to reduce the dimensions by removing the redundant and dependent features by transforming the features from higher dimensional space to a space with lower dimensions.

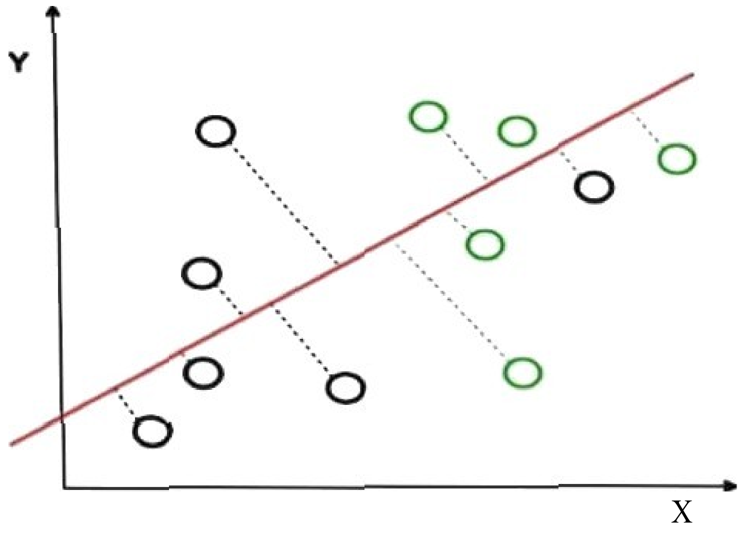
Linear Discriminant Analysis is a supervised classification technique which takes *labels into* consideration.

Suppose we have two sets of data points belonging to two different classes that we want to classify. As shown in the given 2D graph, when the data points are plotted on the 2D plane, there’s no straight line that can separate the two classes of the data points completely. Hence, in this case, LDA (Linear Discriminant Analysis) is used which reduces the 2D graph into a 1D graph in order to maximize the separability between the two classes.

Here, Linear Discriminant Ana\ysis uses both the axes {X and Y) to create a new axis and projects data onto a new axis in a way to maximize the separation of the two categories and hence, reducing the 2D graph into a 1D graph.

Two criteria are used by LDA to create a new axis:

1. **Maximize the distance between means of the two classes.**
2. **MInImize the varlatlon w”‹thin each class.**



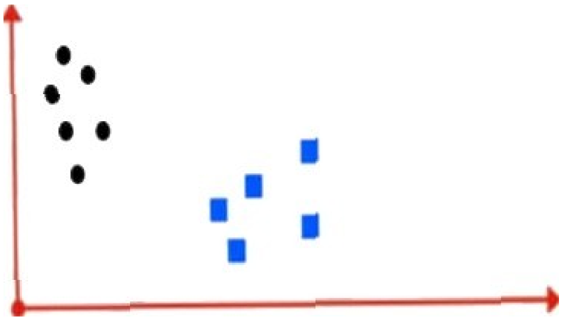
In the above graph, it can be seen that a new axis (in red) is generated and plotted in the 2D graph such that it maximizes the distance between the means of the two classes and minimizes the variation within each class. In simple terms, this newly generated axis increases the separation between the data points of the two classes. After generating this new axis using the above-mentioned criteria, all the data points of the classes are plotted on this new axis and are shown in the figure given below.

But Linear Discriminant Ana\psis fails when the mean of the distributions are shared, as it becomes impossible for LDA to find a new axis that makes both the classes.

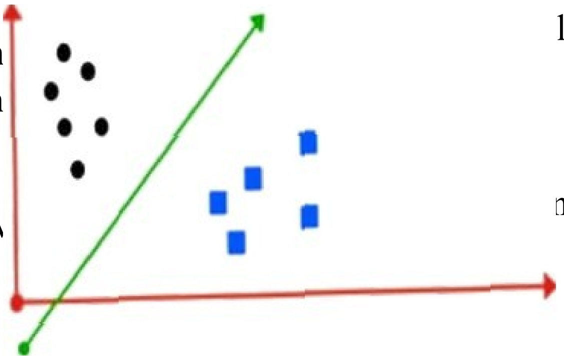
y. Support Vector Machine(SVM):

A Support Vector Machine (SVM) **is a discriminative classified** formally **defined** by a separatinġ hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optlmal hyperplane which categorizes new examples. In two dimenslonal space **this hyperplane is a line dividing a plane in** two **parts where In** each class lay in either side.

However let us understand this in simple terms. Suppose we are given plot of two label classes on graph . Can you decide a separating line for the classes?



You might have come up with
So mething similar to following graph drawn below, It fairly
s ep a rates the two classes. Any point that is left of line falls into black circle class and on right falls
classes. That's what SVM does. It finds out a line/ hyper-
into blue square class. Se aration Oo classes). I wrote multidimensional space.
plane (in multidimensional space that separate outs



lassifica tion tasks. BMt,
H support vector machin
It
owever used for classificatio n
is widely
so basically
nsional space that distinctly
the objective of S\
classifies the data points.

To separate the two classes of data points, there are many possible hyper-planes that could be chosen. Main objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence.

Hyperplanes are decision boundaries that help classify the data points. Data points falling on either side of the hyperplane can be attributed to different clas5es. Also, the dimension of the hyperplane depends upon the number of features. If the number of input features is 2, then the

hyperplane is just a line. If the number of input features is 3, then the hyperplane becomes a two-dimensional plane. It becomes difficult to imagine when the number of features exceeds 3.

Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. Using these support vectors, we maximize the margin of the classifier. Deleting the support vectors will change the position of the hyperplane. These are the points that help us build our SVM.

4. Artificial Neural Network:

ANN learning is robust to errors in the training data and has been successfully applied for learning real-valued, discrete-valued, and vector-valued functions containing problems such as interpreting visual scenes, speech recognition, and learning robot control strategies. The study of artiflCidl neural networks (ANNs) has been inspired in part by the observation that biological learning systems are built of very complex webs of interconnected neurons in brains. The human brain contains a densely interconnected network of approximately 10^11-1O^12 neurons, each connected neuron, on average connected, to I0^4-10^5 others neurons. So on an average human brain take approximate 10"-1 to make surprisingly complex decisions. ANN systems is motivated to capture this kind of highly parallel computation based on distributed representations. Generally, ANNs are built out of a densely interconnected set of simple units, where each unit takes a number of real-

$$Summed\ Input = \sum_i w_i I_i$$

valued inputs and

Produces a single real-valued output. But ANNs are less motivated by biological neural systems, there are many complexities to biological neural systems that are not modeled by ANNs.

Advantage of Using Artificial Neural Networks:

- *Problem* in ANNs can have instances that are represented by many attribute-value pairs.
- ANNs used for problems having the target function output may be discrete-valued, real-
valued, or a vector ol several real- *or* discrete-valued attributes.
- ANN learning methods are quite robust to noise in the training data. The training examples may *contain* errors, which do not affect the final output.
- It is used generally used where the fast evaluation of the learned target function may be required.
- ANNs can bear long training times depending on factors such as the number of weights in the network, the number of training examples considered, and the settings of various learning algorithm parameters.

Single-layer Neural Networks (Perceptrons):

Input is multi-dimensional (i.e. input can be a vector):
input x = (11, 12, .., In)

Input nodes (or units) are connected (typically fully) to a node (or multiple nodes) in the next layer. A node in the next layer takes a weighted sum of all its inputs:

The rule:
The output node has a "threshold" t.
Rule:
If summed input >t, then it "fires"
(output y - 1).
Else (summed input < t) It doesn't fire
(output y= 0).

$$if \sum_i \qquad \geq t$$

then y = 1

else $i/ $\sum_i w_i I_i$

$< t)$

thru y = 0

Boolean Functions and Perceptrons

Limitations oj'Perceptrons:

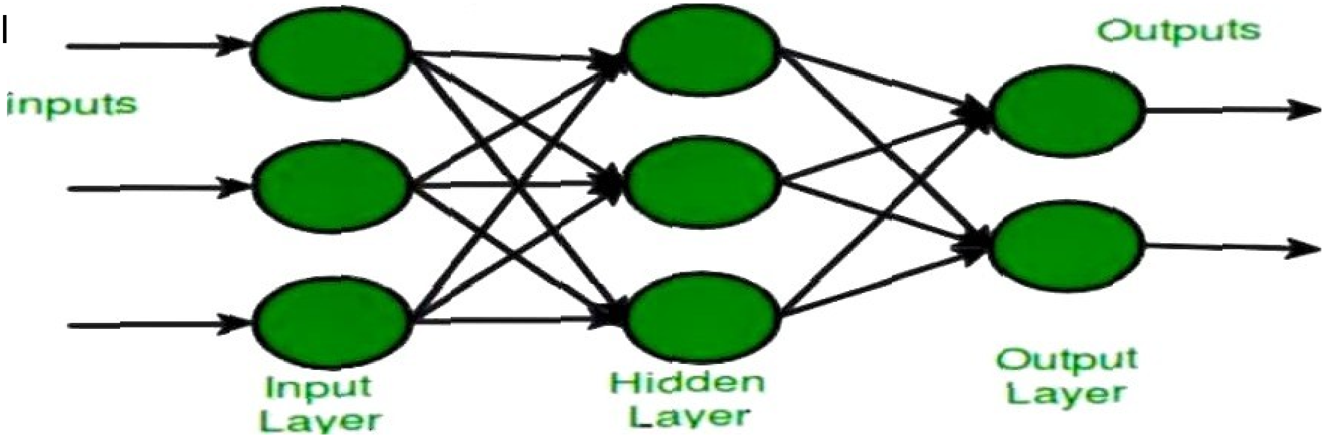
(i) The output values of a perceptron can take on only one of two values (0 er 1) due to the hard-limit transfer function.

(ii) Perceptrons can only classify linearly separable sets of vectors. If a straight line or a plane can be drawn to separate the input vectors into their correct categories, the input vectors are linearly separable. If the vectors are not linearly separable, learning will never reach a point where dll vectors are classified properly

The Boolean function XOR is not linearly separable (Its positive and negative instances cannot be separated by a line or hyperplane). Hence a single layer perceptron can never compute the XOR function. This is a bi8 drawback which once resulted in the stagnation of the field of neural networks. But this has been solved by multi-layer.

Multi-layer Neural Networks:

A Multi-Layer Perceptron (MLP) or Multi-Layer Neural Network contains one or more hidden layers (apart from one input and one output layer). While a single layer perceptron can only learn linear functions, a multi-layer perceptron can also learn non - I



This neuron takes as input x1,x2,....,x3 (and a +1 bias term), and outputs f(summed inputs+bias),

where with f(.) called the activation function. The main function of Bias is to provide everY node

a trainable constant value (in addition to the normal inputs that the node receves) Every activation function tor non-linearity) takes a sin8!e number and performs a certNn fxed

mathematical operation on it. aF£? se veral activation functions you may encounter in There practice:

Sigmoid: takes real-valued input and squash es it to range between 0 and 1.

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

ronh: takes real-valued input and squashes it to the range {-1, 1].

$$\tanh(x) = 2P(2z) - 1$$

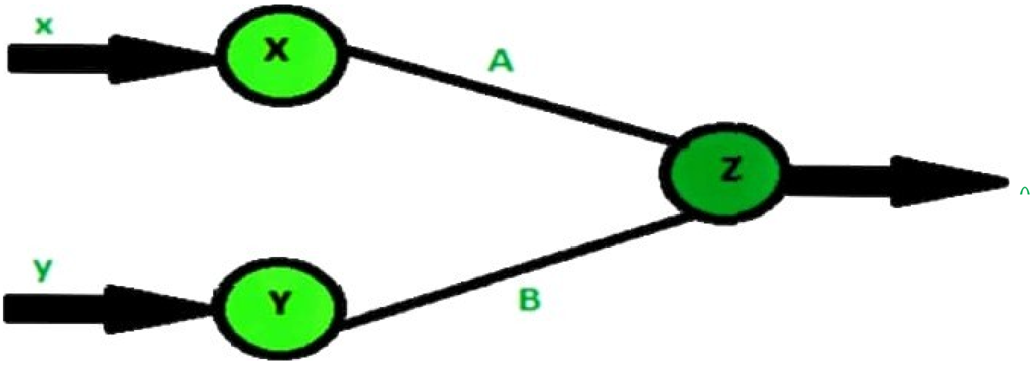
g 1u: ReLu stands for Rectified LimEI r Units. it takes rE'EIl- valued input and thresholds it IO 0

g '
(re places negative values to 0
).

/(g) max(0, z)

Chaq •xeñ stivs pf A fi$ial heural hletworh

•It is neurally implemented mathematical model



Arehilec trrc a Sim[aIe Artificial Net

•It contains huge number of interconnected processing elements called neurons to do all operations

•Information stored in the neurons are basically the weighted linkage of neurons

•The input si8nals arrive at the processing elements through connections and connecting weights.

•It has the ability to learn , recall and generalize from the given data by suitable assignment and adjustment of weights.

eThe collective behavior of the neurons describes its computational power, and no single neuron

carries specific information

How simple neuron works ?

Let there are two neurons X and Y which is transmitting signal to another neuron Z . Then X and y are input neurons for transmitting signals and Z is output neuron for receiving signal The input neurons are connected to theoutput neuron , over a interconnection links (A and B) as shown in figure

For above neuron architecture , the net input has to be calculated in the way .

$$I= xA + yB$$

where x and y are the activations of the input neurons X and Y . The output z of the output neuron Z can be obtained by applying activations over the net input .

$$O=f(I)$$

Output=Function(net input calculated)

The function to be applied over the net input is called activation function . There are various activation function possible for this.

CLASSIFICATION OF EMG SIGNALS USING ANN and SVM CLASSIFIERS FOR 10 CLASSES OF HAND MOVEMENTS USING 2 CHANNELS ELECTRODES .

So, we basically are having raw data of 10 persons and each of them having 10 different classes of individual and combined fingers movements including the flexion of each of the individual fingers, i.e., Thumb(T), Index (I), Middle (M), Ring (R), Little (L) and the pinching of combined Thumb-Index (T—I), Thumt›—Middle (T—M), 7humARing (T-R), ThumI>-Little (T-L), and finally the hand close (HC).

Now, each 10 classes have 6 files which contains 20,£ I EMG data within 2 EMG channels each. So,we are left with a huge data collection of dimension (10 x 10 x 6 x 2fXAKI x 2) where **(20,QX?x2)** can be featured as row and column of a matrix respectively.

Our task is to increase the classification accuracy where ten classes of individual and combined finger movements are to be recognized.

Our First Approach:

What we did is at each time instants 40,000 features(20,fX i feature sets for 2 EMG channels each) were obtained and using PCA -SVM and **PCA-ANN** combination technique we tried to suppress the data huge data limit to 500 features (Dimension reduction method) as it requires large memory to store all training patterns. But in this method we observe that as a large chunk of available data is squeezed into a small packet range of data, it might happen that we might lose some important or crucial principal components without our notice due to which we might end up with biased form of results. As we did the implementation part in python so the maximum data compression limit was up to 500 and after the training and tested part was completed, we found out the classlflcation accuracy to be lying between 0.50-0.60t5D'/t to 60%) which is indeed a un satisfactory result.

So, we had to discard this approach and need to adopt for any other approach where we can effectively reduce number of extracted patterns without compromising the classification accuracy.

Our Second Approach.'

Here 7 feature sets are extracted from the preprocessed raw EMG signals. So, 7 features are available for 2 channels each which results into 14 features. Now , we calculate first order moments to seventh order moments for each channels. As a result we get total 28 features(2 channels x 7 features - 14 features + 7 + 7 moments).After training and testing part was done

using the available offline raw data ,we ended up with classification efficiency lying between 0.85
-0.96 (85 % to 96%) which is quiet good and a satisfactory result than previous method.

Now , here the question arises why such 7 features are important for EMG analysis ?

Features in the time domain are more commonly used for EMG pattern recognition. This is because they are easy, and quick to calculate as they do not require any transformation. Time domain features are computed based upon the input signals amplitude. The resultant values give a measure of the waveform amplitude, frequency, and duration with some limitations.

The fidelity of an EMG signal is influenced by two main concerns.

- Signal to noise ratio - the ratio of the energy in the EMG signal to the energy in the noise signal
- Distortion of the signal - the relative contribution of any frequency component in the EMG signal should not be altered.

The amplitude of the EMG signal is stochastic (random) with a Gaussian distribution which ranges from 0 to 10 mV (peak to peak). Two parameters are commonly used to measure the amplitude, the root-mean-square(RMS) value and the mean absolute (MA) value.

Root Mean Square Value

The RMS represents the square root of the average power of the EMG signal for a given period of time. It is known as a time domain variable because the amplitude of the signal is measured as a function of time.

$$x_{\mathrm{rms}} = \sqrt{\frac{1}{T_2 - T_1}\int_{T_1}^{T_2} \wedge' [f((t))]^2 dt}$$

dt

Mean Absolute Value

The MA value is the computer calculated equivalent of the average rectified value (ARV). The MA value is known as a time domain variable because it is measured as a function of time. It represents the area under the EMG signal once it has been **rectified,** meaning that all of the negative voltage values have been made positke. The MA value is used as a measure of the amplitude of the EMG signal like the root mean square (RMS).

However the RMS is often preferred over the MA value because it provides a measure of the power of the EMG signal while the MA value does not.

Integrated Absolute value

Integrated EMG (IEMG) is calculated as the summation of the absolute values of the EMG signal amplitude. Generally, IEMG is used as an index to detect the muscle activity that used to oncoming the control command of assistive control device. It is related to the EMG signal sequence firing point, which can be expressed as

MAV -— / n ., IxiI where N denotes the length of the signal and xc represents the EMG signal in a segment.

Autoregressive feature

Autoregressive (AR) model described each sample of EMG signal as a linear combination of previous samples plus a white noise error term. AR coefficients are used as features in EMG pattern recognition. The model is basically of the following form:

$$-\sum_{i=1} a_i \, x_{n-i}$$

where Xn a sample of the model signal, a is AR coefficients, w is white noise or error sequence, and p is the order of AR model.

Zero Crossing

Zero crossing (ZC) is the number of times that the amplitude value of EMG signal crosses the zero y-axis. In EMG feature, the threshold condition is used to abstain from the background noise. This feature provides an approximate estimation of frequency domain properties. It can be formulated as

$$ZC = \sum_{n=1}^{N-1} [sgn(x_n * \cap x_{n+1}) \qquad |x_n - x_{n+1}| \geq threshold$$

Slo e Si han e

Slope Sign Change (SSC) is similar to ZC. It is another method to represent the frequency information of EMG signal. The number of changes between positive and negative slope among three consecutive segments are performed with the threshold function for avoiding the interference in EMG signal. The calculation is defined as

SSC -)J (z_n — =.—›)(=. — =.»)]

Waveform Lenpth

Waveform length (WL) is the cumulative length of the waveform over the time segment. WL is related to the waveform amplitude, frequency and time. It is given by

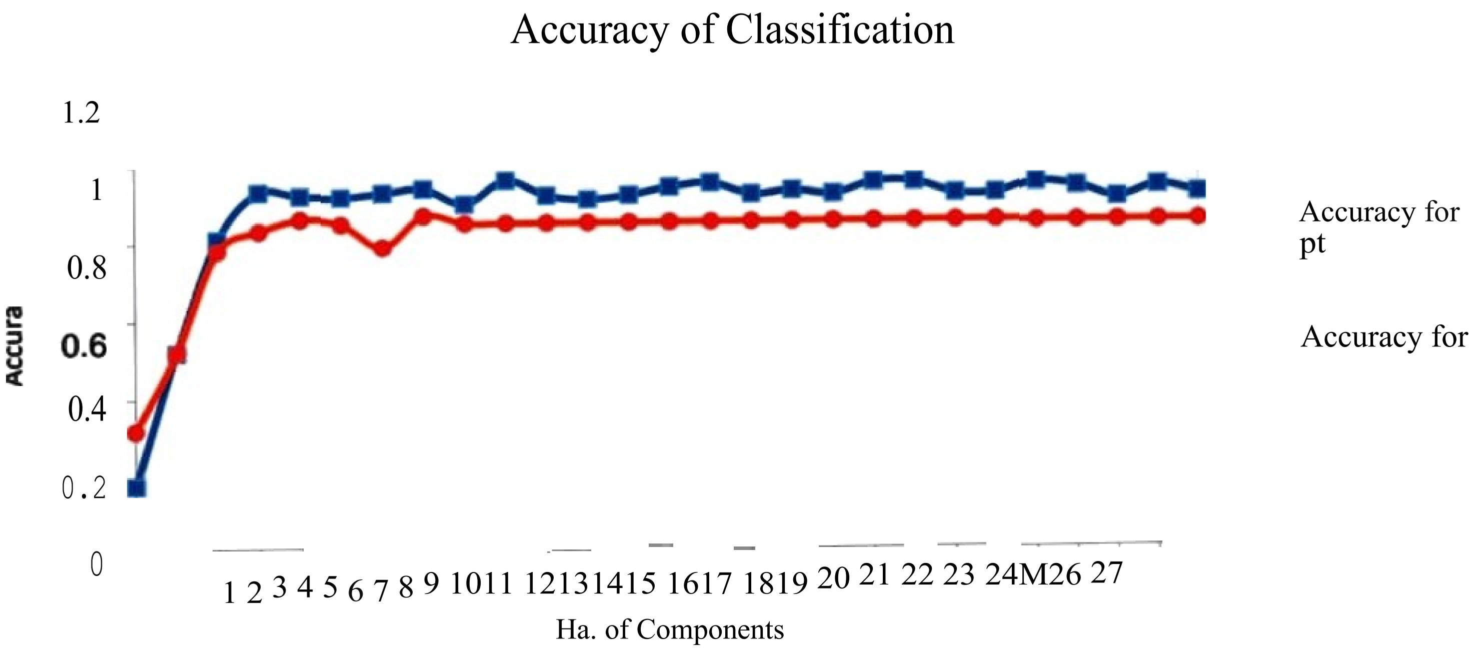
WL=/ xI i I*<+s — *nl

Moments

Other than the above mentioned features we have calculated the moments also. The nth Moment is given by the equation given below,

Mx = S[(I - X)^]

Here we have calculated 7 moments of 1st to 7th order for every channel.



t1] Negi, S., Kumar, Y. and Mishra, V.M., 2016, S0ptombsr. Feature extraction and classification for EMG signals using linear dlscrimlnant analysis. In 2016 2nd international Conference on Advances in Computing, Communication, &amp; Automation (iCACCA)(Fall) (pp. 1-6). IEEE.

{2) Dunteman, G.H., 1959. Principal components analysis (No. 69). Sage.

{3] Mehrotra, K., Mohan, C.K. and Ranka, S., 1997. Elements of artificial neural networks. MIT press.

[4] Jaachims, T., 1996. Making large-scale SVM learning practical (No. 1998, 28). Technical report, SFB 475: Komplexitâtsreduktion in Multivariaten Datenstrukturen, Universitât Dortmund

[5] Zurada, J.M., 1992. Introduction to artificial neural systems (Vol. 8). St. Paul: West publishing company.

}6) Ye, J., Janardan, R. and M, Q., 2Ns. Two-dimensional linear discriminant analysis. In Advances in neural information processing systems (pp. 1569-1576).

[7] M, B.C., 1993. A new computation of geometric moments. Pattern Recognition, 26(1), pp.109-113.

1.Material and methods

1.1. Subjects and data recording

We used the publicly available data described in Andrzejak et al. (2001). The complete data set[1] consists of five sets (denoted A–E) each containing 100 single-channel EEG segments. Sets A and B consisted of segments taken from surface EEG recordings that were carried out on five healthy volunteers using a standardized electrode placement scheme. Volunteers were relaxed in an awake state with eyes open (A) and eyes closed (B), respectively. Sets C, D, and E originated from EEG archive of presurgical diagnosis. EEGs from five patients were selected, all of whom had achieved complete seizure control after resection of one of the hippocampal formations, which was therefore correctly diagnosed to be the epileptogenic zone. Segments in set D were recorded from within the epileptogenic zone, and those in set C from the hippocampal formation of the opposite hemisphere of the brain. While sets C and D contained only activity measured during seizure free intervals, set E only contained seizure activity. All EEG signals were recorded with the same 128-channel amplifier system, using an average common reference. The data were digitized at 173.61 samples per second using 12 bit resolution. Band-pass filter settings were 0.53–40 Hz (12 dB/oct). In this study, we used two dataset (A and E) of the complete dataset as in Subasi (2007). Typical EEGs are given in Fig. 1.

1.2. Analysis using discrete wavelet transform

A signal is said to be stationary if it does not change much over time. Fourier transform can be applied to the stationary signals. However, like EEG, plenty of signals may contain non-stationary or transitory characteristics. Thus it is not ideal to directly apply Fourier transform to such signals. In such a situation time–frequency methods such as wavelet transform must be used. In wavelet analysis, a variety of different probing functions may be used. This concept leads to the defining equation for the continuous wavelet transform (CWT):

[1] EEG time series are available under (http://www.meb.unibonn.de/epileptologie/science/physik/eegdata.html).

*W*ð*a*; *b*Þ ¼ Z 1 *x*ð*t*Þ ——1ffi *w*.*t* — = *b*Σ*dt*

ð1Þ

where *b* acts to translate the function across *x*(*t*), and the variable *a* acts to vary the time scale of the probing function, *w*. If *a* is greater than one, the wavelet function, *w*, is stretched along the time axis, and if it is less than one (but still positive) it contacts the function. While the probing function *w* could be any of a number of different functions, it always takes on an oscillatory form, hence the term ‘‘wavelet.’’ The * indicates the operation of complex conjugation,

and the normalizing factor 1 ensures that the energy is the same

j *a* j

for all values of *a*. In applications that require bilateral transformations, it would be preferred a transform that produces the minimum number of coefficients required to recover accurately the original signal. The *discrete wavelet transform* (DWT) achieves this parsi- mony by restricting the variation in translation and scale, usually to powers of 2. For most signal and image processing applications, DWT-based analysis is best described in terms of filter banks. The use of a group of filters to divide up a signal into various spectral components is termed *sub-band coding*. This procedure is known as multi-resolution decomposition of a signal *x*[*n*]. Each stage of this scheme consists of two digital filters and two down-samplers by 2. The first filter, *h*[ ] is the discrete mother wavelet, high-pass in nature, and the second, *g*[ ] is its mirror version, low-pass in nature. The down-sampled outputs of first high-pass and low-pass filters provide the detail, D1 and the approximation, A1, respectively (Adeli et al., 2003; Marchant, 2003; Semmlow, 2004).

Selection of appropriate wavelet and the number of levels of decomposition is very important in analysis of signals using DWT. The number of levels of decomposition is chosen based on the dominant frequency components of the signal. The levels are chosen such that those parts of the signal that correlate well with the frequencies required for classification of the signal are retained in the wavelet coefficients. Since the EEG signals do not have any useful frequency components above 30 Hz, the number of levels was chosen to be 5. Thus the signal is decomposed into the details D1–D5 and one final approximation, A5. The ranges of various frequency bands are shown in Table 1. Figs. 2 and 3 show five different levels of approximation and details of an EEG signal taken from an unhealthy and healthy subject, respectively. These approximation and detail records are reconstructed from the Daubechies 4 (DB4) wavelet filter (Adeli et al., 2003).

The extracted wavelet coefficients provide a compact representation that shows the energy distribution of the EEG signal in time and frequency. Table 1 presents frequencies corresponding to different levels of decomposition for Daubechies order 4 wavelet with a sampling frequency of 173.6 Hz. In order to further decrease the dimensionality of the extracted feature vectors, statistics over the set of the wavelet coefficients was used (Kandaswamy, Kumar, Ramanathan, Jayaraman, & Malmurugan, 2004). The following statistical features were used to represent the time–frequency distribution of the EEG signals:

(1) Mean of the absolute values of the coefficients in each sub- band.
(2) Average power of the wavelet coefficients in each sub-band.
(3) Standard deviation of the coefficients in each sub-band.
(4) Ratio of the absolute mean values of adjacent sub-bands.

Features 1 and 2 represent the frequency distribution of the signal and the features 3 and 4 the amount of changes in frequency distribution. These feature vectors, calculated for the frequency bands A5 and D3–D5, were used for classification of the EEG signals (Kandaswamy et al., 2004).

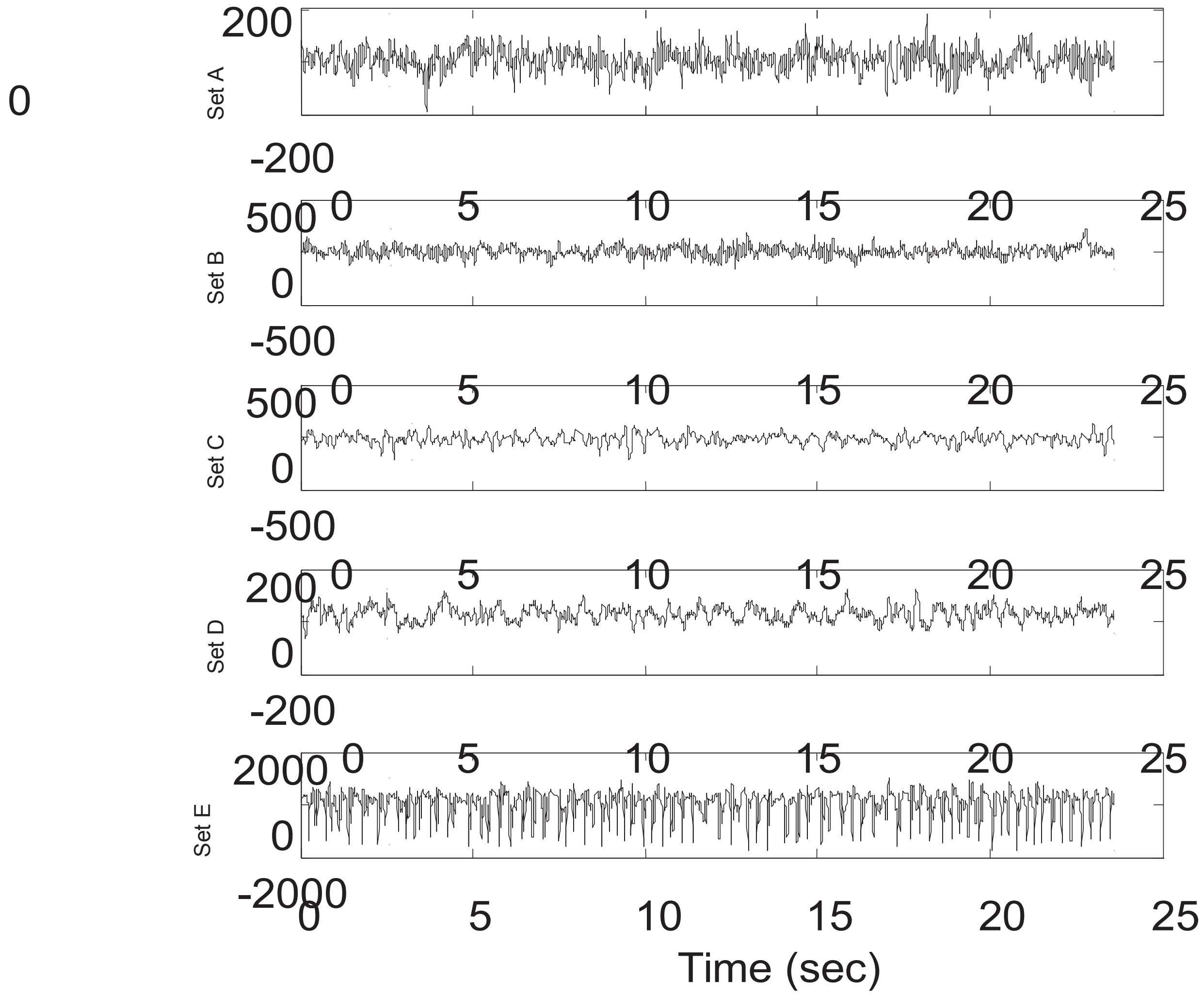


Fig. 1. Examples of five different sets of EEG signals taken from different subjects.

Table 1
Frequencies corresponding to differ- ent levels of decomposition for Daubechies 4 filter wavelet with a sampling frequency of 173.6 Hz.

Decomposed signal	Frequency range (Hz)
D_1	43.4–86.8
D_2	21.7–43.4
D_3	10.8–21.7
D_4	5.4–10.8
D_5	2.7–5.4
A_5	0–2.7

1.3. Feature extraction methods

2.3.1. Principal component analysis (PCA)

Principal component analysis (PCA) is a well-established method for feature extraction and dimensionality reduction. In PCA, we seek to represent the d-dimensional data in a lower-dimensional space. This will reduce the degrees of freedom; reduce the space and time complexities. The objective is to represent data in a space that best expresses the variation in a sum-squared error sense. This technique is mostly useful for segmenting signals from multiple sources. It facilitates significantly if we know how many independent components exist ahead of time, as with standard clustering methods. The basic approach in principal components is theoretically rather simple. First, the d-dimensional mean vector l and $d \times d$ covariance matrix R are computed for the full data set. Next, the eigenvectors and eigenvalues are computed, and sorted according to decreasing eigenvalue. Call these eigenvectors e_1 with eigenvalue k_1, e_2 with eigenvalue k_2, and so on. Sub-sequently, the largest k such eigenvectors are chosen. In practice, this is done by looking at a spectrum of eigenvectors. Often there will be dimension implying an inherent dimensionality of the subspace governing the "signal." The other dimensions are noise. Form a $k \times k$ matrix A whose columns consist of the k eigenvectors. Preprocess data according to:

x^{0} ¼ A^{t}ðx − lÞ ð2Þ

It can be shown that this representation minimizes a squared error criterion. Details are given in Cao et al. (2003), Duda, Hart, and Strok (2001).

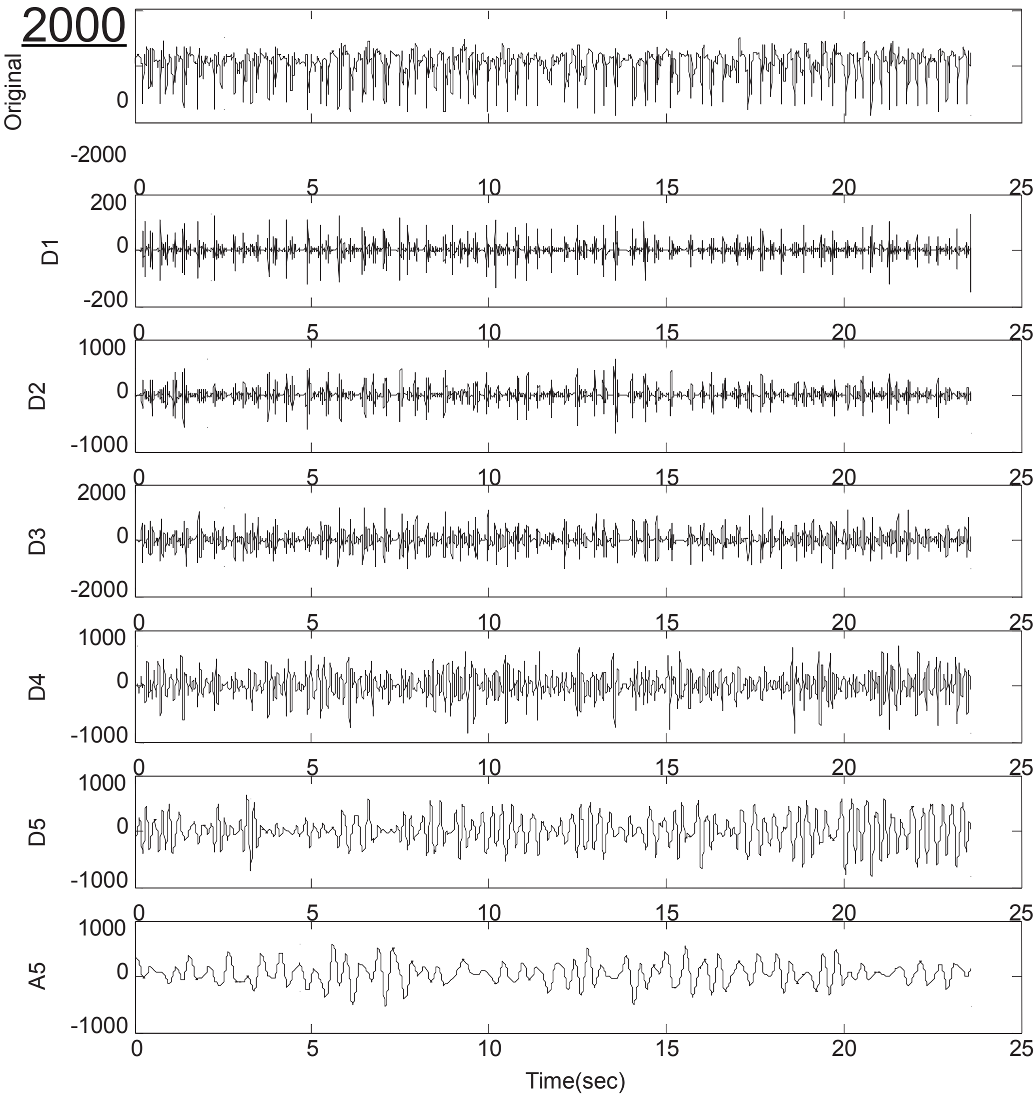


Fig. 2. Approximate and detailed coefficients of EEG signal taken from unhealthy subject (epileptic patient).

2.3.2. discriminant analysis (LDA)

The aim of LDA is to create a new variable that is a combination of the original predictors. This is accomplished by maximizing the differences between the predefined groups, with respect to the new variable. The goal is to combine the predictor scores in such a way that, a single new composite variable, the discriminant score, is formed. This can be viewed as an excessive data dimen- sion reduction technique that compresses the *p*-dimensional pre- dictors into a one-dimensional line. At the end of the process it is hoped that each class will have a normal distribution of discriminant scores but with the largest possible difference in mean scores for the classes. In reality, the degree of overlap between the discriminant score distributions can be used as a measure of the success of the technique. Discriminant scores are calculated by a discriminant function which has the form:

D ¼ w_1Z_1 þ w_2Z_2 þ w_3Z_3 þ· · · þ w_pZ_p

ð6Þ

As a result a discriminant score is a weighted linear combination of the predictors. The weights are estimated to maximize the differences between class mean discriminant scores. Generally, those predictors which have large dissimilarities between class means will have larger weights, at the same time weights will be small when class means are similar (Fielding, 2007).

1.4. Support vector machines (SVMs)

Support vector machines (SVMs) are build on developments in computational learning theory. Because of their accuracy and ability to deal with a large number of predictors, they have more attention in biomedical applications. The majority of the previous classifiers separate classes using hyperplanes that split the classes, using a flat plane, within the predictor space. SVMs broaden the

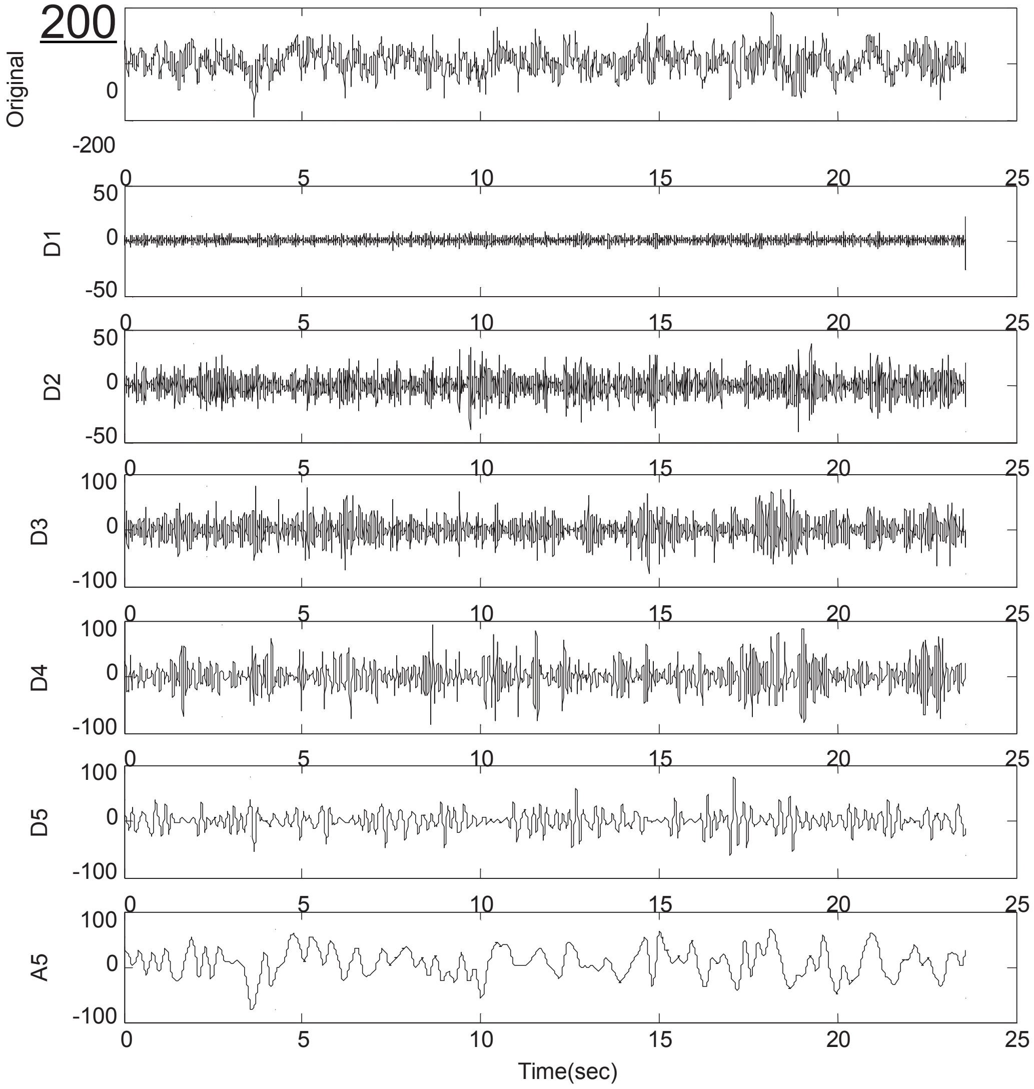


Fig. 3. Approximate and detailed coefficients of EEG signal taken from a healthy subject.

concept of hyperplane separation to data that cannot be separated linearly, by mapping the predictors onto a new, higher-dimensional space in which they can be separated linearly.

The method's name derives from the support vectors, which are lists of the predictor values taken from cases that lie closest to the decision boundary separating the classes. It is practical to assume that these cases have the greatest impact on the location of the decision boundary. In fact, if they were removed they could have large effects on its location. Computationally, finding the best location for the decision plane is an optimization problem that makes uses of a kernel function to build linear boundaries through nonlin- ear transformations, or mappings, of the predictors. The intelligent component of the algorithm is that it locates a hyperplane in the predictor space which is stated in terms of the input vectors and dot products in the feature space. The dot product can then be used to find the distances between the vectors in this higher-dimen- sional space. A SVM locates the hyperplane that divides the support vectors without ever representing the space explicitly. As an alternative a kernel function is used that plays the role of the dot product in the feature space. The two classes can only be separated absolutely by a complex curve in the original space of the predic- tor. The best linear separator cannot totally separate the two clas- ses. On the other hand, if the original predictor values can be projected into a more suitable feature space, it is possible to sepa- rate completely the classes with a linear decision boundary. As a result, the problem becomes one of finding the suitable transfor- mation. The kernel function, which is central to the SVM approach, is also one of the main problems, especially with respect to the selection of its parameter values. It is also crucial to select the magnitude of the penalty for violating the soft margin between the classes. This means that successful construction of a SVM necessi- tates some decisions that should be informed by the data to be classified (Abe, 2005; Burbidge, Trotter, Buxton, & Holden, 1998; Burges, 1998; Duda et al., 2001; Fielding, 2007).

The basic support vector classifier is very similar to the percep- tron. Both are linear classifiers, assuming separable data. In percep- tron learning, the iterative procedure is stopped when all samples in the training set are classified correctly. For linearly separable data, this means that the found perceptron is one solution arbi- trarily selected from an (in principle) infinite set of solutions. In contrast, the support vector classifier chooses one particular solu- tion: the classifier which separates the classes with maximal mar- gin. The margin is defined as the width of the largest ‘tube’ not containing samples that can be drawn around the decision boundary (see Fig. 4). It can be proven that this particular solution has the highest generalization ability.

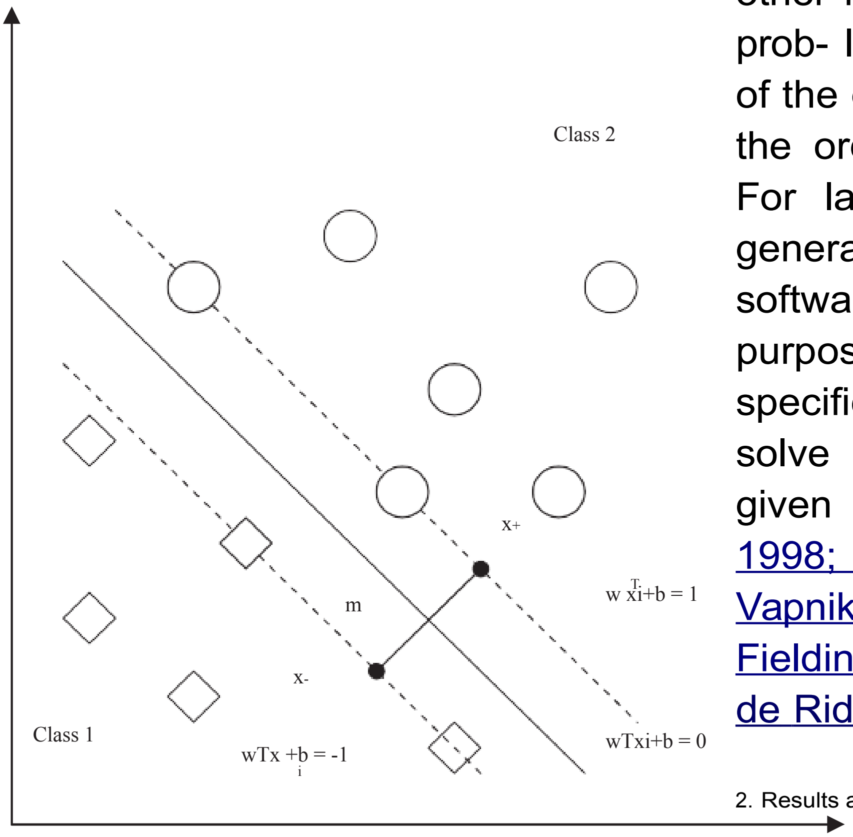


Fig. 4. The linear support vector classifier.

The support vector classifier has many advantages. A unique global optimum for its parameters can be found using standard optimization software. Nonlinear boundaries can be used without much extra computational effort. Moreover, its performance is very competitive with other methods. A drawback is that the problem complexity is not of the order of the dimension of the samples, but of the order of the number of samples. For large sample sizes $N_S > 1000$ general quadratic programming software will often fail and special-purpose optimizers using problem-specific speedups have to be used to solve the optimization. Details are given (Abe, 2005; Burbidge et al., 1998; Burges, 1998; Cortes & Vapnik, 1995; Duda et al., 2001; Fielding, 2007; Van der Heijden, Duin, de Ridder, & Tax, 2004; Vapnik, 1995).

2. Results and discussion

In this study, we used EEG signals of normal and epileptic patients in order to perform a comparison between the PCA, ICA and LDA by using SVM. EEG recordings were divided into sub-band frequencies such as *a*, *b*, *d* and *h* by using DWT. Then a set of statistical features was extracted from the wavelet sub-band frequencies *d* (1–4 Hz), *h* (4–8 Hz), *a* (8–13 Hz) and *b* (13–30 Hz). After normalization, the EEG signals were decomposed using wavelet transform and the statistical features were extracted from the sub-bands.

Then dimension of these features are reduced by using ICA, PCA and LDA. A classification system based on SVM was implemented using these data as inputs.

The objective of the modelling phase in this application was to develop classifiers that are able to identify any input combination as belonging to either one of the two classes: normal or epileptic. For developing neural network classifiers, 800 examples were randomly taken from the 1600 examples and used for training the neural networks, and the remaining 800 examples were kept aside and used for testing the developed models. The class distribution of the samples in the training and test data set is summarized in Table 2.

Additionally, because the problem involves classification into two classes, sensitivity and specificity were used as a performance measure. In order to analyze the output data obtained from the application, sensitivity (true positive ratio) and specificity (true negative ratio) are calculated by using confusion matrix. The sensitivity value (true positive, same positive result as the diagnosis of expert neurologists) was calculated by dividing the total of diagnosis numbers to total diagnosis numbers that are stated by the expert neurologists. Sensitivity, also called the true positive ratio, is calculated by the formula:

$$\text{Sensitivity} = \text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\% \qquad (7)$$

On the other hand, specificity value (true negative, same diagnosis as the expert neurologists) is calculated by dividing the total of diagnosis numbers to total diagnosis numbers that are stated by the expert neurologists. Specificity, also called the true negative ratio, is calculated by the formula:

$$\text{Specificity} = \text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100\% \qquad (8)$$

2.1. Experimental results

Epileptic seizure detection in EEG can be thought as a sort of pattern recognition concept. It consists of data acquisition, signal processing, feature extraction, feature reduction and seizure detection. A novel EEG signal classification method is proposed, which is based on DWT, the dimension reduction (based on ICA, PCA and LDA) and SVM classification. The procedure of the proposed system can be summarized as follows:

Step 1: The features calculated with statistical features parameter from time–frequency domain using DWT.

Step 2: We extract the features using ICA, PCA and LDA algorithm to reduce the dimensionality. This step is performed to remove the irrelevant features which are redundant and even degrade the performance of the classifier.

Step 3: The classification process for epileptic seizure detection is carried out using SVM-based classification.

The procedure was repeated on EEG recordings of all subjects (healthy and epileptic patients). In this work, the radial basis function (RBF) kernel is used as the kernel function of SVMs. There are two parameters related with this kernel: r and c. The upper bound r for penalty term and kernel parameter c plays a critical role in performance of SVMs. Hence, inappropriate selection of parameters r and c, may cause over-fitting or under-fitting problem. Therefore, we should find optimal r and c so that the classifier can accurately classify the data input. In this work, we use 10-fold cross-validation to investigate the appropriate kernel parameter r, and c. Principally, all the pairs of (r, c) for RBF kernel are tried and the one with the best cross-validation accuracy is selected. After the selection of optimal kernel parameters r, and c, the whole training data was trained once more to construct the final classifier.

In this work, the training process carried out using RBF kernel to PCA + SVM, ICA + SVM, and LDA + SVM. After training, we used three different feature extraction methods and get the test results which are shown in Table 3. By using PCA, ICA and LDA features are extracted from original feature sets. In addition, the number of support vectors (SVs) decreased due to feature extraction. As seen in Table 3, the classification rate with LDA feature extraction is highest (100%) and ICA came second (99.5%). The PCA had lowest correct classification percentage (98.75%) compared to LDA and

Table 3
The values of statistical parameters of the ICA, PCA and LDA models for EEG signal classification.

Feature extraction method	Accuracy	Specificity	Sensitivity
PCA (%)	98.75	98.5	99.00
ICA (%)	99.5	99	100
LDA (%)	100	100	100

ICA counterparts. Also the simulation shows that SVM by feature extraction using PCA, ICA, or LDA can always perform better than that without feature extraction (98%). The excellent of LDA is also shown by the number of SVs which is reduced and smaller than PCA and ICA. In these circumstances, classification process using ICA feature extraction needs fewer numbers of SVs than PCA feature extraction. This fact can be explained that ICA finds the components not only uncorrelated but independent. Independent components are more valuable for classification rather than uncorrelated components. However, according to training time, the classification process using LDA feature extraction and SVMs is relatively longer than PCA and ICA feature extraction. Furthermore, it is obvious that kernel parameter selection is crucial to get good performance. Besides, the use of appropriate kernel parameter will overcome the problems of under-fitting and over-fitting so the best classification process is yielded.

2.2. Discussion

Although the previous works have shown good performance on the EEG signal classification, there still remain some problems to be solved. First, the number of available EEG patterns for the classifier training is not much more, which shows us that the generalization ability of a classifier dominates the accuracy of online EEG classification. On the other hand, the classifiers used in the previous works, for instance, the ANNs did not minimize the generalization error bound for unseen EEG patterns. In this work, SVM is implemented to overcome this limitation. Second, the systems in previous works sent all the extracted features into the classifiers directly. But, due to a great deviation in EEG pattern distribution there exist mixed distribution between classes in general. As a result, if a feature transformation mechanism that can minimize the within-class scatter and maximize the between-class scatter is set into the system, it can be anticipated that the size of between-class overlap region can be significantly reduced and the classification performance can be significantly improved. In order to achieve this, the PCA, ICA, and LDA algorithms are used in proposed structure.

Based on the results of the present study and experience in the EEG signal classification problem, we would like to emphasize the following:

1. The high classification accuracy of the SVM classifier gives insights into the features used for defining the EEG signals. The conclusion drawn in the applications demonstrated that the DWT coefficients are the features, which well represent the EEG signals, and by the usage of these features a good distinction between classes can be obtained.
2. Support vector machines (SVMs) are based on preprocessing the data to represent patterns in a high dimension—typically much higher than the original feature space. With an appropriate non- linear mapping to a sufficiently high dimension, data from two categories can always be separated by a hyperplane. As a result, while the original features bring sufficient information for good classification, mapping to a higher-dimensional feature space make available better discriminatory evidence that are absent in the original feature space. The problem of training an SVM is to select the nonlinear functions that map the input to a higher-dimensional space. Often this choice will be informed by the designer's knowledge of the problem domain. In the absence of such information, we might choose to use polynomi- als, Gaussians or other basis functions. The dimensionality of the mapped space can be arbitrarily high (though in practice it may be limited by computational resources). For training the SVMs we chose Radial Basis Function (RBF) and tried to find an appropriate kernel parameters r, and c. The optimal r, and c values can only be ascertained after trying out different values. In addi- tion, the choice of c parameter in the SVM is crucial in order to have a suitably trained SVM. The SVM has to be trained for different kernel parameters until to get the best result (Cortes & Vap- nik, 1995; Ubeyli, 2008; Vapnik, 1995).
3. Subasi (2007) evaluated the diagnostic accuracy of the Mixture of Expert (ME) model and ANN on the same EEG data sets (A and E) (Andrzejak et al., 2001) and the total classification accuracy of the ME model was 94.5% and ANN was 93.2%. Thus, the accuracy rates of the SVM with the ICA, PCA and LDA for this application were found to be significantly higher than that of the ANN and ME model presented in the previous study (Subasi, 2007).
4. Nigam and Graupe (2004) used the same EEG data sets (A and E) by using different feature extraction with ANN and the total classification accuracy of their model was 97.2%. The SVM used for this application indicated higher performance than that of the ANN

model presented by Nigam and Graupe (2004) also.

5. The classification results and the values of statistical parameters indicated that the SVM with the ICA, PCA and LDA had con- siderable success in the EEG signals classification by comparing with the ANN. The proposed combined PCA, ICA and LDA meth- ods with SVM approach can be evaluated in classification of the non-stationary biomedical signals.
6. The testing performance of the SVM-based diagnostic system is found to be satisfactory and we think that this system can be used in clinical studies after it is developed. This application brings objectivity to the evaluation of EEG signals and its automated nature makes it easy to be used in clinical practice. Besides the feasibility of a real-time implementation of the expert diagnosis system, diagnosis may be made more accurately by increasing the variety and the number of parameters.

4. Conclusion

Diagnosing epilepsy is a difficult task requiring observation of the patient, an EEG, and gathering of additional clinical informa- tion. SVMs that classifies subjects as having or not having an epi- leptic seizure provides a valuable diagnostic decision support tool for physicians treating potential epilepsy, since differing etiol- ogies of seizures result in different treatments. Conventional classification methods of EEG signals using mutually exclusive time and frequency domain representations does not give efficient re- sults. In this work, EEG signals were decomposed into time–fre- quency representations using DWT and statistical features were calculated to represent their distribution. Using statistical features extracted from the DWT sub-bands of EEG signals, three feature extraction method; namely PCA, ICA, and LDA, were used with SVM and cross-compared in terms of their accuracy relative to the observed epileptic/normal patterns. The comparisons were based on two scalar performance measures derived from the confusion matrices; namely specificity and sensitivity. The result of EEG signal classification using SVMs shows that nonlinear feature extraction can improve the performance of classifier with respect to reduce the number of support vector. According to this result, the application of nonlinear feature extraction and SVMs can serve as a promising alternative for intelligent diagnosis system in the future. Also it is demonstrated that dimension reduction by PCA, ICA and LDA can improve the generalization performance of SVM.

References

Abe, S. (2005). *Support vector machines for pattern classification*. London: Springer.

Adeli, H., Zhou, Z., & Dadmehr, N. (2003). Analysis of EEG records in an epileptic patient using wavelet transform. *Journal of Neuroscience Methods, 123*, 69–87.

Andrzejak, R. G., Lehnertz, K., Mormann, F., Rieke, C., David, P., & Elger, C. E. (2001). Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Physical Review E, 64*, 061907.

Bronzino, J. D. (2000). Principles of electroencephalography (2nd ed.). In J. D. Bronzino (Ed.). *The biomedical engineering handbook*. Boca Raton: CRC Press LLC. Burbidge, R., Trotter, M., Buxton, B., & Holden, S. (1998). Drug design by machine learning: Support vector machines for pharmaceutical data analysis. *Computers and Chemistry, 26*, 5–14.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery, 2*(2), 1–47.

Cao, L. J., Chua, K. S., Chong, W. K., Lee, H. P., & Gu, Q. M. (2003). A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine. *Neurocomputing, 55*, 321–336.

Cortes, C., & Vapnik, V. (1995). Support vector networks. *Machine Learning, 20*(3), 273–297.

D'Alessandro, M., Esteller, R., Vachtsevanos, G., Hinson, A., Echauz, A., & Litt, B. (2003). Epileptic seizure prediction using hybrid feature selection over multiple intracranial EEG electrode contacts: A report of four patients. *IEEE Transactions on Biomedical Engineering, 50*(5), 603–615.

Duda, R. O., Hart, P. E., & Strok, D. G. (2001). *Pattern classification* (2nd ed.). John Wiley & Sons.

Fielding, A. H. (2007). *Cluster and classification techniques for the biosciences*. Cambridge, UK: Cambridge University Pres.

Kandaswamy, A., Kumar, C. S., Ramanathan, R. P., Jayaraman, S., & Malmurugan, N. (2004). Neural classification of lung sounds using wavelet coefficients. *Computers in Biology and Medicine, 34*(6), 523–537.

Marchant, B. P. (2003). Time–frequency analysis for biosystem engineering. *Biosystems Engineering, 85*(3), 261–281.

Nigam, V. P., & Graupe, D. (2004). A neural-network-based detection of epilepsy. *Neurological Research, 26*(1), 55–60.

Semmlow, J. L. (2004). *Biosignal and biomedical image processing: MATLAB-based applications*. New York: Marcel Dekker, Inc..

Subasi, A. (2006). Automatic detection of epileptic seizure using dynamic fuzzy neural networks. *Expert Systems with Applications, 31*, 320–328.

Subasi, A. (2007). EEG signal classification using wavelet feature extraction and a mixture of expert model. *Expert Systems with Applications, 32*, 1084–1093.

Ubeyli, E. D. (2008). Analysis of EEG signals by combining eigenvector methods and multiclass support vector machines. *Computers in Biology and Medicine, 38*, 14–22.

Van der Heijden, F., Duin, R. P. W., de Ridder, D., & Tax, D. M. J. (2004). *Classification parameter estimation and state estimation: An engineering approach using MATLAB*. England: John Wiley & Sons Ltd..

Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer.

Wang, X., & Paliwal, K. K. (2003). Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition. *Pattern Recognition, 36*, 2429–2439.

Widodo, A., & Yang, B. (2007). Application of nonlinear feature extraction and support vector machines for fault diagnosis of induction motors. *Expert Systems with Applications, 33*, 241–250.