



Two-Class *in Silico* Categorization of Intermediate Epileptic EEG Data

Abhijit Dasgupta¹, Ritankar Das¹, Losiana Nayak¹, Ashis Datta²,
and Rajat K. De¹(✉)

¹ Machine Intelligence Unit, Indian Statistical Institute,
203 Barrackpore Trunk Road, Kolkata 700108, India
rajat@isical.ac.in

² Epilepsy and Electrophysiology Division, Institute of Neurosciences,
185/1 A.J.C. Bose Road, Kolkata 700017, India

Abstract. Epilepsy treatment depends on multiple instances of EEG recordings. Often, clinicians encounter intermediate/borderline EEG signals in the recordings, and as a result, inconclusiveness arises regarding the epilepsy status of a patient. In this paper, we have addressed this issue with a computational solution. We have classified and created class-specific clusters of the EEG data belonging to epileptic patients and normal individuals using a scaled conjugate feed forward neural network (FNN) and average silhouette based k-means clustering algorithm respectively. Thereafter, we have categorized the intermediate data into the clusters of these two classes for a better clinical decision making using minimum squared Euclidean distance. The methodology proposed here can help the clinicians in dealing with intermediate EEG signals found in individuals suspected as suffering from epilepsy. It will also help in categorizing intermediate EEG data, and in turn facilitate clear diagnosis and better patient care in the case of undecided patients.

Keywords: Epilepsy · Epileptic network · Feed forward neural network · k-means clustering · Borderline EEG data

1 Introduction

Seizure is a common neurological disorder caused by abnormal electrical activities in the brain. Most often a seizure begins with micro-seizures starting asynchronously from small neuronal clusters. These micro-seizures then construct a macroscale hypersynchronization, *i.e.*, a seizure in combined form [11]. Epilepsy is a chronic disorder of the brain that can be caused by some underlying disease mechanism and is usually characterized by occurrence of seizure. However, the exact cause remains unknown in about 50% of patients [15].

It is estimated that about 70% of people living with epilepsy could live seizure-free if proper diagnosis and treatment are followed. However, in developing countries, epilepsy is still not considered as a disease of importance. The

negligence in treatment due to lack of awareness, resources, and manpower as caregivers often leads to premature mortality of epileptic patients [3].

Epilepsy is characterized by recurrent epileptic seizures. Electroencephalogram (EEG) is helpful in evaluating a patient with epilepsy [13]. It is usually interpreted by an experienced neurologist, epileptologist or an electrophysiologist. Often, the intermediate or borderline EEG recordings of individuals create a lot of confusion in clinical decision making. Whether these recordings behave as normal or epilepsy cannot be specifically determined, and thus it affects treatment. Clinicians are usually inclined towards further observation and treatment as and when required. Automated analysis and classification of such EEG data with the help of computational techniques can be helpful in interpreting such recordings as normal or epilepsy. Although a considerable amount of work has been done to classify EEG signals as either normal or epilepsy [1, 2, 14], none of them involves categorization of intermediate EEG data.

In this paper, we have analyzed EEG data of 91 healthy volunteers and 119 epilepsy patients to categorize 14 intermediate individuals. We have extracted 120 significant features from adjacency matrices obtained from the EEG data of all the categories. Thereafter, scaled conjugate gradient (SCG) feed forward neural network (FNN) is employed to classify the data in a frame of two-class problem. Silhouette based k-means clustering has been used to extract clusters of normal and epileptic EEG data followed by superimposition of the intermediate data to either epileptic or normal classes using minimum squared Euclidean distance.

2 Data

We have collected Electroencephalographic (EEG) data [8] of epileptic, intermediate and normal individuals, recorded using 16 channels/electrodes placed on the scalp. The location and naming of the electrodes follow the internationally accepted Modified Combinatorial Nomenclature (MCN) system [10].

Here, we have collected data from 91 healthy volunteers comprising 36 males and 55 females with a mean age of 19.1 years. Besides 119 patients suffering from epilepsy have been considered. Among them, 53 are males and 66 are females with a mean age of 18.75 years. Intermediate data (data which cannot be rightly classified as normal or epilepsy) has been collected from 14 individuals. Here, a total of 73 EEG recordings have been selected from 7 females and 7 males with a mean age of 44.21 years.

EEG Data have been recorded using Recorders & Medicare Systems Pvt. Ltd. (RMS) computerized EEG machine and Nihon Kohden Neurofax (NKN) EEG-1200, for an interval of 20–30 min each. During recording, all the participants have been asked to relax in a no-thinking state, and to lay motionless as far as possible keeping their eyes open.

3 Method

The methodology involves data filtering, data pre-processing, computation of adjacency matrix, creation of the final dataset, classifying normal and epileptic data and finally categorization of intermediate data as depicted in Fig. 1.

Data Filtering and Pre-processing

Data filtering steps include resampling, re-referencing, baseline correction and Independent Component Analysis (ICA) as per our previous investigations [5,6]. Here we have used EEGLAB toolbox [7] version 13 in MATLAB R2018a platform to implement the data filtering steps.

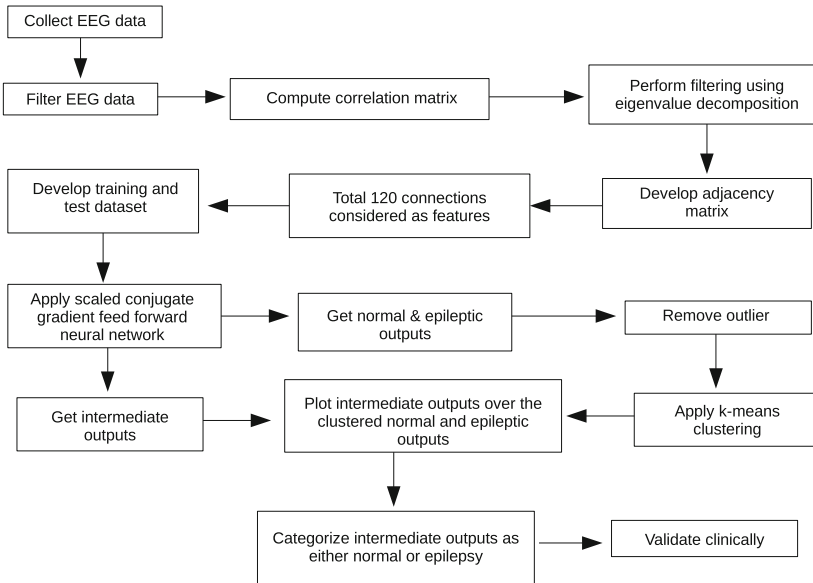


Fig. 1. Flowchart of the overall methodology.

In addition, we have employed a sliding window based eigenvalue decomposition technique to remove unwanted noise from the filtered time series data so that an adjacency matrix of the network, connecting positional nodes of the electrodes, can be generated. A vector of length l has been defined to measure the potential of each electrode on a continuous time series. Here, we have designed a window of size s to slide over the entire length l to divide the data into N_{trial} number of smaller segments/trials for each class, such as normal, epileptic and intermediate. Thus, we have obtained 1085, 834 and 499 trial data from 91 healthy volunteers, 119 epileptic and 14 intermediate patients respectively. It helps in increasing the number of inputs applied to the learning algorithm. Here, the sliding window has an overlap of 20%.

Computing Correlation Matrix. We have computed Pearson's correlation coefficient to measure the linear dependency of one channel to another. Here, for each window of size s , and two channels $\mathbf{x}^p = [x_1^p, x_2^p, \dots, x_s^p]^T$ and $\mathbf{x}^q = [x_1^q, x_2^q, \dots, x_s^q]^T$, the Pearson's correlation coefficient (c_{pq}) can be calculated as

$$c_{pq} = \frac{1}{s-1} \sum_{i=1}^s \left(\frac{x_i^p - \mu_{x^p}}{\sigma_{x^p}} \right) \left(\frac{x_i^q - \mu_{x^q}}{\sigma_{x^q}} \right) \quad (1)$$

Here, the terms μ_{x^p} and μ_{x^q} represent the mean values, while σ_{x^p} and σ_{x^q} stand for standard deviations of \mathbf{x}^p and \mathbf{x}^q respectively.

Thus, we obtain a correlation matrix $\mathbf{C} = [c_{pq}]_{n \times n}$ for each trial of each class, where $n = 16$ in our study. Here $c_{pq} = 1$ if $p = q$, otherwise c_{pq} lies between $+1$ and -1 . Positive values represent positive correlation and negative values represent negative correlation, while a value of zero represents no linear correlation between channels. Higher the value, higher is the correlation between channels/electrodes.

Filtering Correlation Matrix. In order to get a more appropriate noise free correlation among channels, we have carried out eigenvalue decomposition [9] on the correlation matrix. The eigenvalue spectrum of the correlation matrix contains a global part \mathbf{C}^{global} with the largest eigenvalue, regional part $\mathbf{C}^{regional}$ of our interest with intermediate eigenvalues, and finally a random part \mathbf{C}^{random} which contains a bulk of small eigenvalues. Thus, the correlation matrix (\mathbf{C}) can be expressed as

$$\mathbf{C} = \boldsymbol{\nu}_0 \lambda_0 \boldsymbol{\nu}_0^T + \sum_{i=1}^{n_h} \boldsymbol{\nu}_i \lambda_i \boldsymbol{\nu}_i^T + \sum_{j=n_h+1}^{n-1} \boldsymbol{\nu}_j \lambda_j \boldsymbol{\nu}_j^T \quad (2)$$

Here, the term n_h determines the separating boundary between $\mathbf{C}^{regional}$ and \mathbf{C}^{random} . The terms $\boldsymbol{\nu}_0$ and λ_0 stand for the eigenvector and eigenvalue of the global part respectively, while $\boldsymbol{\nu}_i$ and λ_i , for $i = 1, \dots, n_h$, represent the eigenvectors and eigenvalues of the regional part respectively. Similarly, the terms $\boldsymbol{\nu}_j$ and λ_j , for $i = n_h+1, \dots, n-1$, represent the eigenvectors and eigenvalues of the random part respectively.

Filtering the correlation matrix involves removal of both the global and random part. The global part contributes to the highest eigenvalues and affects all the channels during the recording of EEG. On the other hand, random noise that lies at the bottom of the eigenvalue spectrum, may interfere with one or more channels but does not affect all the channels as in the case of global noise. Thus, the filtered correlation matrix can be written as

$$\mathbf{C}_{filtered} = \sum_{i=1}^{n_h} \boldsymbol{\nu}_i \lambda_i \boldsymbol{\nu}_i^T \quad (3)$$

Here, n_h has been selected in such a way that it can be possible to nullify the effect of the random part by filtering out the eigenvectors with minimum

eigenvalues. There are 16 eigenvalues (λ_i , for $i = 1, \dots, 16$) for each correlation matrix of each trials for each class. The highest eigenvalue in the eigenvalue spectrum represents the global part of each correlation matrix. Besides, the random part contains the smallest eigenvalues. For epilepsy, we have chosen n_h to be 6, *i.e.*, after removing the global part, the next six highest eigenvalues contribute to the regional part. Similarly, we have set n_h as 7 and 8 for normal and intermediate trials respectively.

The determination of n_h is difficult but obtaining an exact value is not crucial, as a small change of n_h does not affect the result. In the eigenvalue spectrum, the eigenvalues corresponding to the regional part, are confined to a small portion. Therefore, if we neglect the eigenvalues that lie close to the boundary of the random region, it does not affect the final result much.

Computation of Adjacency Matrix

We set a cutoff value α from the filtered correlation matrix to obtain an adjacency matrix \mathbf{A} . The value α can be calculated as

$$\alpha = \frac{1}{N_{ind}} \sum_{i=1}^{N_{ind}} \left(\frac{1}{N_{trial}} \sum_{j=1}^{N_{trial}} \left(\frac{1}{n \times n} \sum_{k,l} |c_{kl}| \right) \right) \quad (4)$$

where N_{ind} represents number of individuals belonging to each class, and N_{trial} is the number of trials carried out on each individual. Subsequently, c_{kl} is an element of the $n \times n$ matrix $\mathbf{C}_{filtered}$. Thus, for $(a, b)^{th}$ element of $\mathbf{C}_{filtered}$ for each trial, $|c_{ab}| \geq |\alpha|$ and $a \neq b$, we consider an edge between a^{th} and b^{th} electrode with the corresponding correlation value ($|c_{ab}|$) as weights, otherwise we have considered no edge between the two electrodes.

Creation of the Final Dataset

The resultant adjacency matrix provides a weighted undirected graph $\mathbf{G}(V, E)$, where V is a finite non-empty set of electrodes and E is the set of edges. There exists a function from $V \times V \rightarrow E$ such that, $E = \{\{u, v\} | u, v \in V, \text{ and } u \neq v\}$. Thus, for a set of 16 electrodes, there are maximum of $(16 \times 15)/2 = 120$ significant connections possible. These connections are considered as features. Besides, the weight value assigned to each connection represents a feature value. Thus, the training input matrix for the learning algorithm (a feed forward neural network) is represented by $\mathcal{I}_{N_{trial} \times 121}$ such that $e_{ij} \in \mathcal{I}$ ($j \neq 121$) represents j^{th} feature of i^{th} trial and $e_{ij} \in \mathcal{I}$ ($j = 121$) represents the class level of i^{th} trial. Similarly, the test input matrix for the learning algorithm is denoted by $\mathcal{I}'_{N_{trial} \times 120}$ such that $e_{ij} \in \mathcal{I}'$ represents j^{th} feature of i^{th} trial.

We have used 1919 samples/trials to be classified into two classes, *viz.*, epilepsy and normal. Out of the total number of samples 70% (1343 samples) are randomly selected for training while 10% (192 samples) are considered for validation. The remaining 20% (384 samples) are used to test the prediction accuracy.

Classifying Normal and Epileptic Data

Here, we have classified epileptic and normal samples using a scaled conjugate gradient (SCG) [4, 12] feed-forward neural network for its quick and better

convergence rate. Besides, SCG has shown better classification accuracy compared to self-organizing feature map, multilayer perceptron and support vector machine.

We have randomly chosen an initial weight vector and set an initial conjugate vector to the steepest descent vector. Thereafter, we have calculated the Hessian matrix approximation and the scaling factor. The scaling factor has been readjusted based on the value of error approximation, whenever the Hessian matrix is not positive definite. Subsequently, the step size has been calculated. The process has been repeated iteratively until it meets the predefined performance goal or maximum validation failure.

During implementation of the aforesaid algorithm, we have set the maximum validation failures as 10, whereas the minimum performance gradient has been considered as $1/e$. Subsequently, we have considered the learning rate of 0.1. Here, the hidden layer consists of 100 nodes.

Categorization of Intermediate Data

The normal and epileptic data have again been fed into the trained feedforward neural network (FNN) to generate all the corresponding points on the output space. In order to eliminate outliers from the output space under consideration, we have used curve fitting toolbox implemented in MATLAB R2018a. Here, we have considered the data points on the output space as outliers those lie above or below 1.5 times the standard deviation from the mean.

The outlier removed data are then clustered into k clusters using average silhouette based k-means clustering algorithm. Further, the trained FNN generated output points for intermediate trials have been projected on resulting k clusters for normal and epileptic dataset. Here, the squared Euclidean distance from each intermediate output point d to each cluster center ψ_i , for $i = 1, \dots, k$, can be calculated as

$$\gamma^2 = (\psi_i - d) * (\psi_i - d)^T \quad (5)$$

Now, the point d has been clubbed into the cluster corresponding to minimum γ^2 . Similarly, all other output points for intermediate epileptic trails have been categorized as either normal or epilepsy. For each individual patients, if more than 50% of trails have been categorized as normal/epilepsy, then the individual recording has been treated as that of the same category. Whereas, if exactly 50% of trials have been categorized as normal/epilepsy, we have left the individual recording as undetermined.

4 Results and Discussion

Here, applying scaled conjugate gradient feed forward neural network with 100 nodes in the hidden layer, we have correctly classified 781 epilepsy and 1049 normal trials out of total 1919 trials obtained from 91 normal and 119 epileptic individuals. Subsequently, 53 epileptic trials have wrongly been classified as normal, while 36 normal trials have wrongly been classified as epileptic. Thus, it has shown an overall accuracy of 95.36%.

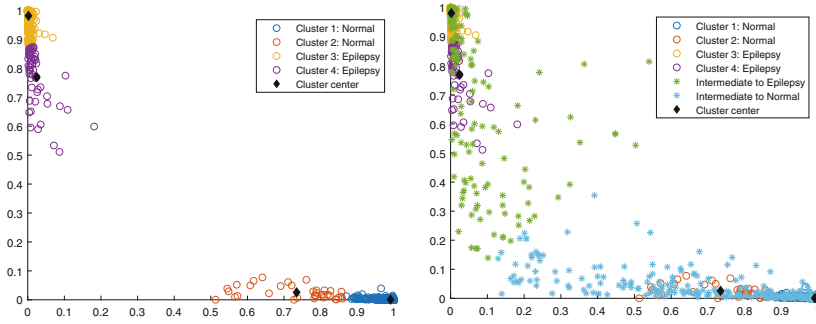


Fig. 2. The outlier removed output points from feed forward neural network for the two classes, *i.e.*, normal and epilepsy, have been plotted on the output space and clustered into four different clusters (left). Depending on minimum squared Euclidean distance from the cluster centres, the output points for intermediate data have been categorized as either normal or epilepsy (right).

Thereafter, intermediate data have been applied as inputs to the previously trained aforementioned FNN model. The output points for intermediate data have been used to categorize intermediate epileptic patients as either normal individuals or epileptic patients. Here, the output obtained from the trained FNN for all normal and epileptic trials have been clustered to serve as a template for categorizing the intermediate data as depicted in Fig. 2. In this context, it should be noted that out of 73 intermediate EEG recordings involving 14 individuals, clinicians are able to identify 38 recordings (obtained from nine individuals) as belonging to either epilepsy or normal category. Here, 31 predictions obtained by the proposed methodology have matched with the clinical recommendation. However, one recording have not be identified, while six recordings have wrongly been predicted, thus giving us an accuracy of 81.58% for intermediate recordings. On the other hand, while categorizing intermediate individuals into either normal or epilepsy class and comparing the results with the diagnosis by clinicians, we have found that the proposed methodology has correctly identified 7 individuals while 2 individuals have wrongly been categorized.

Table 1. Clinical view and computational prediction for intermediate recordings

Categorized as	Clinical view	Prediction through proposed methodology on recordings			
		Correct	Wrong	Undetermined	Accuracy (in %)
Normal	12	8	2	1	81.58
Epilepsy	26	23	4	0	
Undetermined	35	0	0	0	
Total	73	31	6	1	

Table 2. Clinically and computationally predicted decisions for intermediate individuals

Categorized as	Clinical view	Prediction through proposed methodology on individuals			
		Correct	Wrong	Undetermined	Accuracy (in %)
Normal	2	1	1	0	77.78
Epilepsy	7	6	1	0	
Undetermined	5	0	0	0	
Total	14	7	2	0	

Henceforth, we obtain an accuracy of 77.78% for intermediate individuals. Although clinicians have failed to identify the remaining 35 recordings, the proposed method is able to categorize 12 of them as epileptic and leave another one as undetermined, while rest have been categorized as normal. Thus, three clinically unidentified individuals have computationally been categorized as normal while two have been categorized as epileptic. Here, an individual has been assigned to a particular class containing his/her maximum recordings. The summary of our findings are found in Tables 1 and 2.

5 Conclusion

The primary objective of this article is to categorize intermediate EEG data into two classes; namely, normal and epilepsy. It would assist clinicians in their decision making regarding patients who cannot be rightly categorized. Clinically these individuals are usually treated according to the symptoms as and when observed. Therefore, it delays the process of disease identification and treatment.

Here we have tried to employ different machine learning tools to categorize intermediate recordings as either normal or epilepsy. It is quite successful in categorizing each intermediate recordings and has shown 77.78% accurate result for individuals. The prediction made by the methodology can thus help clinicians in accurately diagnosing individuals as normal, epilepsy or intermediate.

This study has been conducted on individuals above the age of 10 years. Among them, most of the people are right-handed. All recordings have been taken at no sleep and eye open state. Similar studies can be conducted on children and/or during sleep. Further, obtaining regional brain signatures from right brain dominant individuals and comparing them with our study may provide a more clear picture of brain regions that are usually affected during epilepsy.

Acknowledgement. AD acknowledges Digital India Corporation (formerly Media Lab Asia), Ministry of Electronics and Information Technology, Government of India, for providing him a Senior Research Fellowship under the Visvesvaraya Ph.D. scheme for Electronics and IT. RD acknowledges Council for Scientific and Industrial Research (CSIR), India for providing him a Senior Research Fellowship (09/093(0182)/2018 EMR-I). LN acknowledges University Grants Commission, India for a UGC Post-Doctoral Fellowship (No. F.15-1/2013-14/PDFWM-2013-14-GE-ORI-19068(SA-II)).

References

1. Bajaj, V., Pachori, R.B.: EEG signal classification using empirical mode decomposition and support vector machine. In: Deep, K., Nagar, A., Pant, M., Bansal, J.C. (eds.) *Proceedings of the International Conference on Soft Computing for Problem Solving (SocProS 2011)* December 20-22, 2011. AISC, vol. 131, pp. 623–635. Springer, New Delhi (2012). https://doi.org/10.1007/978-81-322-0491-6_57
2. Bajaj, V., Pachori, R.B.: Separation of rhythms of EEG signals based on Hilbert-Huang transformation with application to seizure detection. In: Lee, G., Howard, D., Kang, J.J., Ślęzak, D. (eds.) *ICHIT 2012*. LNCS, vol. 7425, pp. 493–500. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-32645-5_62
3. Bell, G.S., Neligan, A., Sander, J.W.: An unknown quantity—the worldwide prevalence of epilepsy. *Epilepsia* **55**(7), 958–962 (2014)
4. Chel, H., Majumder, A., Nandi, D.: Scaled conjugate gradient algorithm in neural network based approach for handwritten text recognition. In: Nagamalai, D., Renault, E., Dhanuskodi, M. (eds.) *CCSEIT 2011*. CCIS, vol. 204, pp. 196–210. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-24043-0_21
5. Dasgupta, A., Das, R., Nayak, L., De, R.K.: Analyzing epileptogenic brain connectivity networks using clinical EEG data. In: *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 815–821. IEEE (2015)
6. Dasgupta, A., Nayak, L., Das, R., Basu, D., Chandra, P., De, R.K.: Feature selection and fuzzy rule mining for epileptic patients from clinical EEG data. In: Shankar, B.U., Ghosh, K., Mandal, D.P., Ray, S.S., Zhang, D., Pal, S.K. (eds.) *PReMI 2017*. LNCS, vol. 10597, pp. 87–95. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-69900-4_11
7. Delorme, A., Makeig, S.: EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* **134**(1), 9–21 (2004)
8. Ebersole, J., Milton, J.: The electroencephalogram (EEG): a measure of neural synchrony. In: Milton, J., Jung, P. (eds.) *Epilepsy as a Dynamic Disease*. BIOMEDICAL, pp. 51–68. Springer, Heidelberg (2003). https://doi.org/10.1007/978-3-662-05048-4_5
9. Kim, D.H., Jeong, H.: Systematic analysis of group identification in stock markets. *Phys. Rev. E* **72**(4), 046133 (2005)
10. Klem, G.H., Lüders, H.O., Jasper, H., Elger, C., et al.: The ten-twenty electrode system of the International Federation. *Electroencephalogr. Clin. Neurophysiol.* **52**(3), 3–6 (1999)
11. Lowenstein, D.H.: Decade in review—epilepsy: edging toward breakthroughs in epilepsy diagnostics and care. *Nat. Rev. Neurol.* **11**(11), 616 (2015)
12. Møller, M.F.: A scaled conjugate gradient algorithm for fast supervised learning. *Neural Netw.* **6**(4), 525–533 (1993)
13. Noachtar, S., Rémi, J.: The role of EEG in epilepsy: a critical review. *Epilepsy Behav.* **15**(1), 22–33 (2009)
14. Subasi, A.: EEG signal classification using wavelet feature extraction and a mixture of expert model. *Expert Syst. Appl.* **32**(4), 1084–1093 (2007)
15. Zunt, J.R., et al.: Global, regional, and national burden of meningitis, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol.* **17**(12), 1061–1082 (2018)