Indian Journal of Science and Technology, *Vol 10(27), DOI: 10.17485/ijst/2017/v10i27/107757, July 2017*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Epileptic Seizure Prediction in EEG Records using Parallel Tree Based Learning and Feature Extraction

P. Ramina[1]* and M. Vanitha[2]

[1]Bharathidasan University, Trichy – 620024, Tamilnadu, India; rameenaa@gmail.com
[2]Department of Computer Applications, Alagappa University, Karaikudi – 630003, Tamilnadu, India

Abstract

Objectives: To propose an effective classifier to identify interictal from preictal signals and to reduce the EEG data size to fit into the processing model. **Methods/Analysis:** The first phase of the processing model identifies several features related to time-series data and identifies the features for data of specific dimension, called the epoch. A set of features are generated for each epoch and the corresponding class details are appended. The obtained data is used by the Classifier for the learning phase, followed by the test phase to identify the effectiveness of the classifier model. **Findings:** Experiments were conducted on EEG data obtained from American Epilepsy Foundation. The proposed features were extracted and the classifier model is trained and the prediction levels were recorded. It was observed that the true prediction levels of interictal signals showed high accuracy (0.91), while the true prediction levels of preictal signals exhibited moderate accuracy (0.66). The false predictions were identified to range from low to moderate (0.08 and 0.33). The accuracy levels were observed to be 0.75 and F-Measure was found to be 0.77. The architecture also exhibits moderate recall (0.66) and high precision levels (0.93). **Novelty/Improvement:** The proposed technique enables effective reduction of data by extracting features for an epoch, hence enabling customized data fine-tuning and scalability.

Keywords: Decision Tree, Feature Extraction, Random Forest, Seizure Prediction, EEG

1. Introduction

Epilepsy is a neurological disorder which affects nearly 1% of the world's population. It is usually characterized by sudden seizures. Though anticonvulsant medications[1] are available for patients, they are to be given at large doses due to the unpredictable nature of the seizures. This leads to a lot of side effects. Even in such a case, medications have become ineffective for 20-40% of the patients. The alternate solution proposed for this problem is surgical removal of the parts of brain related to seizure activity. However, even after such removals, several patients were found to exhibit spontaneous seizures[2]. The unavailability of a reliable solution reduces the quality of life of epileptic patients to a very large extent[3]. It becomes very difficult for them to carry on with their daily chores, as the occurrence of epilepsy is sudden and unpredictable[4].

Seizure forecasting is a boon to the patients. It can help them live their normal lives. Recent improvement in technology makes it possible to accurately measure brain activity. Electrical Brain Activity (EEG) signals serve as the basis for predicting seizures, as pre-seizure signals exhibit increased activity, and they vary considerably from normal signals. Appropriately identifying these signals can provide early warning to patients, reducing the necessity for regular drugs and surgical procedures[5]. However, several factors such as the streaming nature of the data, data hugeness and customized analysis has made the process complex[6].

EEG is a streaming signal that has to be processed. Hence several techniques based on time domain, wavelet domain[7], frequency domain, SVD, PCA[8] and ICA domain have been proposed. A real-time seizure prediction model called Brainatic was proposed in[9]. This technique relies on applying machine learning techniques on features extracted from the EEG signals. Classification is performed by Support Vector Machines (SVM), Multi-Layer Perceptron (MLP) neural networks and Radial Basis

*Author for correspondence

Function (RBF) neural networks. A data decomposition based epilepsy prediction technique was presented in[10]. This technique decomposes the channels into 8 sparse rational components and then applies 1D LGBP operator followed by data down sampling. A sequential technique utilizing decomposition and extreme learning algorithms for seizure detection was presented in[11]. A generic seizure detection system to identify various types of epileptic attacks was presented in[12]. A least square based processing technique that uses feature extraction techniques for seizure detection was presented in[13].

2. Data Analysis

Seizure detection/ prediction uses EEG signals generated from human brain as the base data. These are quick neuron transmissions from human brain, hence are data composed of high frequency signals. Properties of the EEG signals used in this paper is presented in Table 1.

Table 1. EEG Signal Properties

Property	Value
Data	240000 X 16
Sampling Rate	400
Samples in Segment	240000
Channels	16
Sequence	Sequence Number (1-6)

The data is made up of .mat files, containing a basic structure of metadata and then the actual data. The data is composed of 16 rows, representing the 16 electrode signals from the patient. The signals are divided into six sequences, each sequence containing 10 minute signals. In total, interictal and preictal signals for 1 hour are represented in the dataset. EEG signals are usually represented with four states, interictal, preictal, ictal and postictal. Ictal is the signal state during seizure, interictal is the normal state, preictal state represents signals until five minute prior to the seizure and postictal state represents signals right after the seizure has occurred. The interictal and preictal states are of importance to us, since they can be used in the effective prediction of seizures. Sampling rate of the signal is maintained at 400Hz, representing 400 signals for each second.

3. Our Approach

Brain signals can be analyzed to identify seizures prior to their occurrence. Brain signals exhibit four states, of which the preictal state is of major importance, as it plays a crucial role in predicting seizures. Seizure prediction is the process of identifying preictal signals from interictal signals[14]. This process is categorized as a binary classification problem, however, the classifier faces several major challenges. The first challenge being the huge data size. As discussed in the previous section, the minimum sampling level for EEG signals is 400Hz, hence for every second, 400 signals are generated from each channel. This leads to a huge amount of data being generated for every second. Further, for analysis, data pertaining to 60 minutes of EEG signals are to be used. Hence leading to a huge spike in the amount of data to be processed. Another major challenge faced by seizure prediction systems is that the process of prediction differs for each individual[15]. Hence user specific training is required for effective prediction, leading to the requirement of a large training data pertaining to individuals[16]. This also leads to a huge training time. The next major challenge faced by the classifier is data imbalance. It is essential that the data used for training is obtained from the user, however, the number of interictal signals are bound to be higher than the preictal signals creating imbalance.

This paper proposes a fast and effective seizure prediction model using frequency based feature data. The major aim of the proposed technique is to utilize frequency based properties of the rather than the actual data itself. Hence this mode of operation reduces size of the data to a large extent, making it appropriate for usage in a real-time environment. The proposed seizure prediction architecture is shown in Figure 1.

EEG data signals obtained from the specific patient is taken as the input data. The architecture requires a minimum of at least 20 hours of interictal data and at least 2-3 hours of preictal signal data. The data will be processed and only the feature vectors are to be taken for model creation, hence the size of data can be significantly larger.

3.1 Feature Vector Creation

The first phase of processing is the creation of feature vectors for training the model. The feature vectors correspond to a group of data for a specific time interval, called the epoch. The data itself is divided into 10 minute

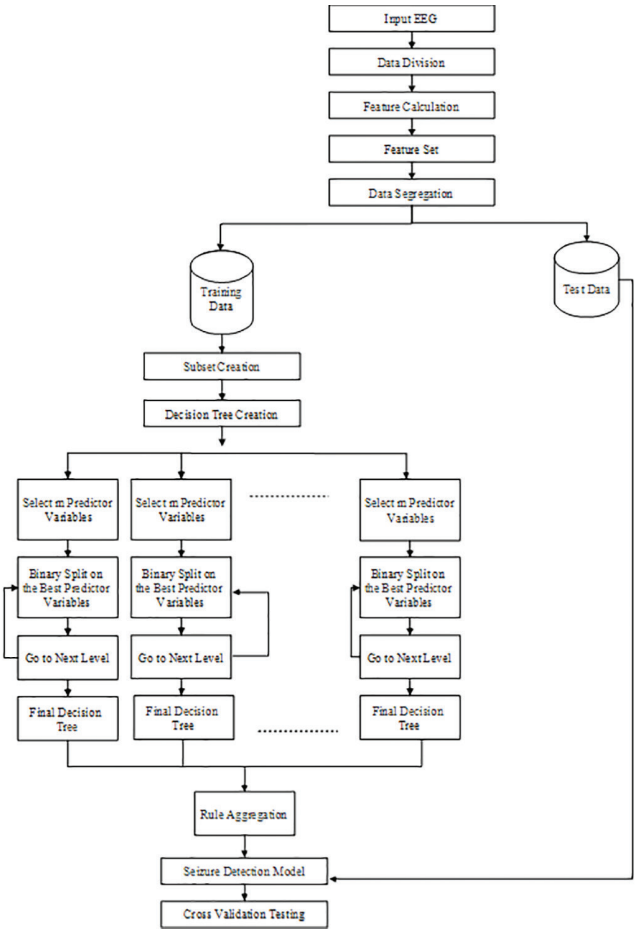

Figure 1. Seizure Prediction Architecture.

sequences. The default epoch time of 10 minutes can be used as such. However, this threshold is best determined depending on the size of the available data. On analyzing the data, it was identified that a maximum threshold limit of 10 minutes is to be maintained during the division. Dividing into intervals greater than 10 minutes will lead to inconsistencies, hence inappropriate predictions. This work uses an epoch of 1 minute intervals to calculate the feature vectors.

A brief description of the features to be used for creating feature vectors is presented below.

3.1.1 Fast Fourier Transformation

A Fourier transformation[17] is used to convert signals from its original domain to a representation in the frequency domain. This process tends to group signals preserving their frequency domain variations, hence reducing the number of computations from $O(n^2)$ to $O(n \log n)$. A normalization component is also incorporated into this module to equalize the importance levels of each component in the signal.

3.1.2 Shannon Entropy

Entropy in general is a non-linear measure depicting the degree of complexity in a time series. It defines how well an epoch can be differentiated from other epochs in a series. Shannon entropy is characterized by a degree of uncertainty associated with the occurrence of a result[18]. Information entropy S of a random variable X that takes the values $x_1, x_2, \ldots, x_N$ is defined as:

$$S_{en} = \sum_{i=1}^{n} p(x_i) \log_a \frac{1}{p(x_i)} = -\sum_{i=1}^{n} p(x_i) \log_a p(x_i),\ a > 1$$

3.1.3 Spectral Edge Frequency

Spectral edge frequency is the frequency at which a certain percent of the signals fall below the signal's frequency domain.

3.1.4 Cross Correlation

Cross correlation[19] presents the level of similarity existing between two signals. This falls between -1 and +1. A value of 1 indicates positive correlation, exhibiting a direct proportional relationship, a value of -1 indicates a negative correlation, exhibiting an inversely proportional relationship and a 0 represents no relationship.

3.1.5 Hjorth Parameters

Hjorth parameters[20] are indicators of statistical properties in a time series data. The parameters are activity, mobility and complexity.

The activity parameter represents the signal power, indicating the surface of power spectrum in the frequency domain.

$$Activity = var(y(t)).$$

Where y(t) represents the signal.

The mobility parameter represents the mean frequency.

$$Mobility = \sqrt{\frac{var\left(\frac{dy(t)}{dt}\right)}{var(y(t))}}.$$

Complexity parameter represents change in frequency with respect to a pure sine wave.

$$Complexity = \frac{Mobility\left(\frac{dy(t)}{dt}\right)}{Mobility(y(t))}$$

3.1.6 Skewness and Kurtosis

Skewness represents the symmetry of a probability density function around its mean and kurtosis describes the length of the tail of a curve.

3.1.7 Others

Other parameters used for analysis includes Shannon entropy dyad, cross correlation dyad, mean and median. The epochs are passed to the corresponding functions and the feature vectors[21] corresponding to the data are obtained. Since the epoch threshold is maintained at 1 minute, every sequence generates 10 feature vectors, corresponding to a data reduction ratio of 1:24000. A low epoch level provides higher accuracy, while a higher epoch value might lead to rounded off data, hence a lesser accurate prediction. However, this totally depends upon the input data, hence it is to be determined on a trial and error basis.

The feature set is created by aggregating all the feature vectors and the class value corresponding to each feature vector. The True value (1) represents preictal signals and the False value (0) represents interictal signals. This forms the base data for the proposed work.

3.2 Seizure Prediction Model Creation

The feature set created in the previous section is passed to this phase for model creation. This work uses Random Forest[22,23], a parallelized version of the DecisionTree[24] algorithm to perform the prediction. Decision trees build decision rules from the training data and utilizes the generated rules to perform prediction. However, it is considered to be a weak classifier. Random Forests are built with the basic principle that several weak classifiers can be combined to form a strong classifier.

Random Forests are composed of several Decision Trees, each being trained on a subset of the data. The number of decision trees to be used for a particular application is problem dependent and is carried out on the basis of user's expertise. This work utilizes a Random Forest with 10 decision trees. A subset of the data is passed to each of the trees for rule creation. Subset creation is carried out with a major constraint that at least 66% of the actual data is to be passed to each of the trees. This constraint is to ensure that all the trees are provided with sufficient number of entries for each of the class. This process plays a vital role in determining the prediction levels on imbalanced data. As discussed earlier, the EEG data is imbalanced, with preictal signals as the minority class due to rare occurrences of seizures. The architectural model of Random Forest provides effective performance irrespective of the level of imbalance contained in the data, making it flexible in terms of the imbalance levels.

Subset creation is followed by the rule creation phase. This leads to the creation of the actual decision trees. Data subsets are passed to each of the trees and each tree determines its base predictor variable. The predictor variable is selected from a sampled set of m predictors, sampled from M total predictor variables ($m<M$). The best predictor variable is identified from this set and a binary split is performed on it. This marks the beginning of the tree creation process. Predictor sampling and predictor identification is carried out for all the available predictors and a decision tree is created for each of the data subsets. Pruning forms a major role in the process of creating a decision tree. However, as only a part of the data is used for the tree creation process, pruning is not performed and the trees are retained as such.

The process of sampled predictor selection requires a major component m, the number of sample predictors to be selected. This can be performed in three ways. The first is the Random splitter selection method, where the value of m is always set to 1, the Breiman's bagger method[25], where all the available predictors are chosen ($m=M$) and the Random Forest method where $m<<M$. Brieman also suggested three possible ways of selecting the values of m, namely $\frac{1}{2}\sqrt{m}$, $\sqrt{m}$ or $2\sqrt{m}$.

The rule set obtained from each of the decision trees are aggregated to provide the final decision rules. Though the decision rules from each of the trees are considered to be weak, the combined decision rules were found to be stronger in terms of both reliability and accuracy.

The patient is usually provided with an EEG measuring device and an alerting device/wearable[26]. If the detected preictal signals crosses the defined threshold, the patient is alerted to take medications or move to the nearest safe point. Though the hardware based areas are not discussed in this paper, several contributions are available[27] concentrating in this domain. Pairing the proposed

architecture with such devices can prove to be very effective in improving the quality of life of epileptic patients.

4. Results and Discussion

Seizure data utilized in the proposed approach was obtained from the American Epilepsy Society[28]. Data is in the form of .mat files, with 6 sequences for each 1 hour signal. The process of seizure prediction is performed in two major phases; feature extraction and model building. Both the phases are encoded in Python, due to the flexibility offered by Python in operating .mat files and learning algorithms.

Training and testing is performed on the data and the confusion matrix[29] is created. ROC and PR Plots[30] are constructed from the data and the algorithm efficiency is analyzed.

ROC plot depicting the performance of the Random Forest classifier is presented in Figure 2. Data corresponding to each of the transactions is noted and the plot is created. It could be observed from the plot that as the classifier progresses, it exhibits very high true positive rates and low false positive rates. The true positive rates exhibit the performance of a classifier in correctly predicting the true values. According to the dataset, the true value represents preictal signals representing the probability of occurrence of seizure. Hence the algorithm is proposed to exhibit very good prediction levels in identifying pre-seizure signals.

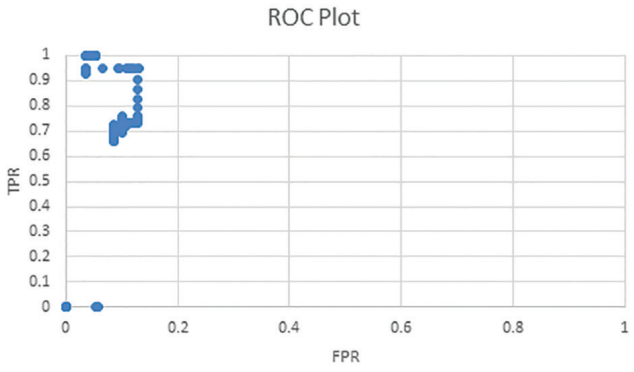

Figure 2. ROC Plot.

The Precision Recall (PR) Plot presented in Figure 3 presents a measure of relevance, i.e. the level of type 1 and type 2 errors exhibited by the prediction algorithm. Precision presents the fraction of the retrieved instances that are relevant and recall presents the fraction of relevant instances that are retrieved. Precision is considered as the measure of quality, while recall is considered as the measure of quantity. It could be observed from the plot that very high precision and recall levels are exhibited by the proposed algorithm. This can be used to prove the presence of very low type 1 and type 2 errors.

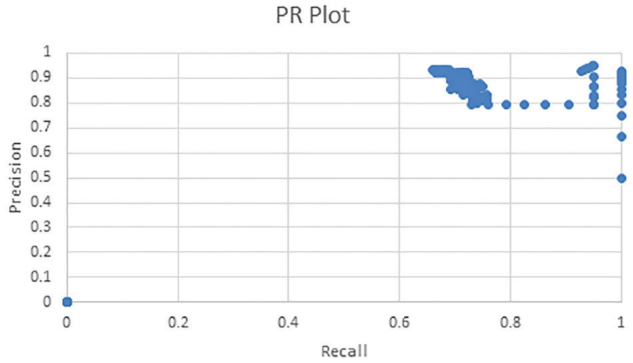

Figure 3. PR Plot.

Additional performance metrics used for analysis are presented in Table 2. The average accuracy of 74.8% and an F-Measure level of 0.77 was observed from the Random Forest Classifier. These exhibit moderate prediction rates of the algorithm as a whole. However, the true negative rate is 91%, exhibits very high prediction level of the negative classes (interictal signals). Further, the false negative rates were also found to be low at 33%. However, it could be observed from the results that there still is scope for improvement in the preictal prediction levels.

Table 2. Performance Metrics

Metric	Value
Accuracy	0.748744
F-Measure	0.774775
TNR	0.913043
FNR	0.338462

5. Conclusion

Epilepsy has become one of the most disturbing issues due to the sudden onset of seizures. However, the brain signals can be utilized to predict such onsets at least 60 minutes prior to the occurrence, providing sufficient buffer for the patients. This paper presents a feature extraction based EEG prediction technique that can be

used to effectively predict seizures. The process of classifier training is usually customized and differs from person to person. A common training data will not be effective during the prediction process. Hence the proposed work utilizes a parallelized ensemble approach, which is also robust towards data imbalance.

Future extensions of the proposed technique includes identifying several other features for inclusion in the training data. Though the proposed technique exhibits high true prediction levels, the false predictions reach moderate levels when aggregated. Future works will also concentrate in reducing the false prediction levels. Further, the proposed technique is a single time training set. However, a learning system that dynamically adapts to the patient's data would provide improved prediction levels. Future works will also concentrate on developing a continuous learning system that tunes itself according to the patient's data to provide enhanced prediction levels.

6. References

1. Nandhakumar J, Tyagi MG. Evaluation of seizure activity after phospho-diesterase and adenylate cyclase inhibition (SQ22536) in animal models of epilepsy. Indian Journal of Science and Technology. 2010 Jul; 3(7):710-7.
2. Kwan P, Brodie MJ. Early identification of refractory epilepsy. New England Journal of Medicine. 2000 Feb; 342:314-19. Crossref. PMid:10660394.
3. Camfield P, Camfield C. Idiopathic generalized epilepsy with generalized tonic-clonic seizures (IGE–GTC): a population-based cohort with N20 year follow up for medical and social outcome. Epilepsy and Behavior. 2010 May; 18(1-2):61-3. Crossref. PMid:20471324.
4. Prince PGK, Hemamalini R, Rajkumar RI. LabVIEW based abnormal muscular movement and fall detection using MEMS Accelerometer during the occurrence of seizure. Indian Journal of Science and Technology. 2014 Oct; 7(10):1625-31.
5. Ramgopal S, Thome-Souza S, Jackson M, Kadish NE, Fernandez IS, Klehm J, Bosl W, Reinsberger C, Schachter S, Loddenkemper T. Seizure detection, seizure prediction, and closed-loop warning systems in epilepsy. Epilepsy and Behavior. 2014 Aug; 37:291-307. Crossref. PMid:25174001.
6. Freestone DR, Karoly PJ, Peterson AD, Kuhlmann L, Lai A, Goodarzy F, Cook MJ. Seizure prediction: science fiction or soon to become reality? Current Neurology and Neuroscience Reports. 2015 Nov; 15(11):1-9. Crossref. PMid:26404726.
7. Hassan AR, Siuly S, Zhang Y. Epileptic seizure detection in EEG signals using tunable-Q factor wavelet transform and bootstrap aggregating. Computer Methods and Programs in Biomedicine. 2016 Dec; 137:247-59. Crossref. PMid:28110729.
8. Ventura A, Franco JM, Ramos JP, Direito B, Dourado A. Epileptic seizure prediction and the dimensionality reduction problem. 19th International Conference on Artificial Neural Networks. 2009 Sep; p. 1-9.
9. Teixeira C, Favaro G, Direito B, Bandarabadi M, Feldwisch-Drentrup H, Ihle M, Alvarado C, Le Van Quyen M, Schelter B, Schulze-Bonhage A, Sales F. Brainatic: A system for real-time epileptic seizure prediction. Brain-Computer Interface Research. 2014; p. 7-17. PMid:24708728 PMCid:PMC3983857.
10. Samiee K, Kovacs P, Gabbouj M. Epileptic seizure detection in long-term EEG records using sparse rational decomposition and local Gabor binary patterns feature extraction. Knowledge-Based Systems. 2017 Feb; 118:228-40. Crossref.
11. Li D, Xie Q, Jin Q, Hirasawa K. A sequential method using multiplicative extreme learning machine for epileptic seizure detection. Neurocomputing. 2016 Nov; 214:692-707. Crossref.
12. Ulate-Campos A, Coughlin F, Gainza-Lein M, Fernandez IS, Pearl PL, Loddenkemper T. Automated seizure detection systems and their effectiveness for each type of seizure. Seizure. 2016 Aug; 40:88-101. Crossref. PMid:27376911.
13. Zamir ZR. Detection of epileptic seizure in EEG signals using linear least squares preprocessing. Computer Methods and Programs in Biomedicine. 2016 Sep; 133:95-109. Crossref. PMid:27393803.
14. Alotaiby TN, Alshebeili SA, Alshawi T, Ahmad I, El-Samie FE. EEG seizure detection and prediction algorithms: a survey. EURASIP Journal on Advances in Signal Processing. 2014 Dec. Crossref.
15. Giannakakis G, Sakkalis V, Pediaditis M, Tsiknakis M. Methods for seizure detection and prediction: an overview. Modern Electroencephalographic Assessment Techniques: Theory and Applications. 2014 Aug; 91:131-57. Crossref.
16. Jory C, Shankar R, Coker D, McLean B, Hanna J, Newman C. Safe and sound? A systematic literature review of seizure detection methods for personal use. Seizure. 2016 Mar; 36:4-15. Crossref. PMid:26859097
17. Fast Fourier transform. Available from: https://en.wikipedia.org/wiki/Fast_Fourier_transform. Date accessed: 22/06/2017.
18. Borowska M. Entropy-Based Algorithms in the Analysis of Biomedical Signals. Studies in Logic, Grammar and Rhetoric. 2015 Dec; 43(1):21-32.
19. Cross correlation. Available from: https://en.wikipedia.org/wiki/Cross-correlation. Date accessed: 23/06/2017.

20. Hjorth B. EEG analysis based on time domain properties. Electroencephalography and Clinical Neurophysiology. 1970 Sep; 29(3):306-10. Crossref.
21. Zhang Z, Chen Z, Zhou Y, Du S, Zhang Y, Mei T, Tian X. Construction of rules for seizure prediction based on approximate entropy. Clinical Neurophysiology. 2014 Oct; 125(10):1959-66. Crossref. PMid:24690391.
22. Ho TK. Random decision forests. IEEE, Proceedings of the Third International Conference on Document Analysis and Recognition. 1995 Aug; 2:278-82.
23. Ho TK. The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1998 Aug; 20(8):832-44. Crossref.
24. Quinlan JR. Simplifying decision trees. International Journal of Man-Machine Studies. 1987 Sep; 27(3):221-34. Crossref.
25. Breiman L. Random forests. Machine Learning. 2001 Oct; 45(1):5-32. Crossref.
26. Rao MJ, Rao MK. An RTOS based Architecture for Patient Monitoring System with Sensor Networks. Indian Journal of Science and Technology. 2016 May; 9(17):1-5.
27. Patel AD, Moss R, Rust SW, Patterson J, Strouse R, Gedela S, Haines J, Lin SM. Patient-centered design criteria for wearable seizure detection devices. Epilepsy and Behavior. 2016 Nov; 64:116-21. Crossref. PMid:27741462
28. Melbourne University Seizure Prediction. Available from: https://www.kaggle.com/c/melbourne-university-seizure-prediction. Date accessed: 02/09/2016
29. Confusion matrix. Available from: https://en.wikipedia.org/wiki/Confusion_matrix. Date accessed: 14/05/2017.
30. Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters. 2006 Jun; 27(8):861-74. Crossref.