

Real-Time Customized Seizure Prediction on Streaming EEG Data using Attribute Extraction and Feature Identification Techniques

P. Ramina
Research Scholar
Bharathidasan University
Trichy, Tamil Nadu, India

M. Vanitha
Asst Professor
Alagappa University
Karaikudi, Tamil Nadu, India

ABSTRACT

Epilepsy is considered to be a neurological disorder caused by unoriented signal emissions from brain, leading to seizures. Prior identification of occurrence of seizures is made possible by measuring the signal emissions at certain parts of the brain, known as EEG. Fast detection of preictal signals can alert patients to prevent catastrophe. However, EEG signals are voluminous and have very high velocity rates, making the prediction process complex. This paper presents an effective seizure prediction model, that enhances predictions by identifying frequency based features and performs two level data reduction to enable faster processing. The processed data is then passed to GBT, a boosted ensemble model for prediction. Experiments were conducted with data obtained from American Epilepsy Society. Results indicate good performances in terms of ROC and PR. A comparison with an existing parallel bagging based seizure prediction model indicates improved accuracy levels in the proposed model.

Keywords

Seizure Prediction; Feature Identification; Attribute Elimination; Gradient Boosted Trees; EEG

1. INTRODUCTION

Epilepsy is a neurological disorder, characterized by unexpected seizures. Epileptic seizure prediction from electroencephalogram (EEG) signals is an important application in patient monitoring systems. EEG signal has become the most convenient measure to analyze signals from different lobes of the brain. The acquisition of EEG signals is based on electrical activities in the brain, which are measured and recorded using non-invasive methods [1]. These signals returned from the brain follows certain defined patterns, that corresponds to the state of the human nervous system. Though the signals differ between brains, they remain harmonious. Any disruption in this harmony leads to seizures [2]. Although disruptions occur prior to seizures, the onset of seizure is not immediate. External indicators cannot be observed due to the disruptions, however, measuring them can clearly provide an alert with convenient time buffer levels for the patient to implement appropriate precautionary measures.

Patients with epileptic disorder usually consume medications to control seizures. However, this does not always prevent epilepsy. According to WHO [3], 1.5% to 5% of the population might have seizure in their lifetime. A survey indicates that the main cause of death in seizures is due to drowning and accidents [4]. Avoiding such incidents is possible only by identifying seizures prior to the actual occurrence. This is possible by analyzing EEG signals from the brain. The major challenges faced in designing such a system is the hugeness and streaming nature of the data. EEG

signals are usually measured from several lobes in the brain, hence it generates signals in several channels, leading to huge amounts of data being generated. The velocity levels of the data being generated are also very high. Hence a system satisfying all these real-time constrains and also delivers results on time is the major need for the current domain.

2. RELATED WORK

Seizure prediction is usually performed as a time series analysis model. A real-time seizure prediction model based on Bayesian classifier was proposed by Behnam et al. in [5]. This is an offline seizure prediction model, that extracts histogram based statistical features to create signal components. The optimal feature components are then identified and Multi-Layer Perceptron (MLP) is used for the actual prediction process. Although this model exhibits good accuracy levels, the usage of MLP leads to high time requirements which does not fit to the real-time requirements of the seizure prediction domain. A mobile based framework for seizure prediction was proposed by Sareen et al. in [6]. EEG signals are captured through wireless technologies and analysis is performed in the cloud. This model utilizes k-means classifier for signal categorization. An EEG based low-complexity model for seizure prediction is proposed in [7]. This model utilizes Fourier transformation applied on six EEG bands for seizure prediction. This is an early warning system that operates on 20 second signal batches. Other similar early warning systems include early warning system for critical transitions [8], a simulation based early warning system [9], a time series based early warning system [10] and dynamical warning system [11]. A poincare plane based feature extraction model for predicting epileptic seizures was proposed by Sharif et al. in [12]. This model utilizes SVM for the prediction process. A Fourier transformation and chaos theory based preictal prediction model was proposed by Fei et al. in [13]. This model is a modified form of Lyapunov exponent [14], to effectively operate on time series data. Other EEG based seizure prediction systems include a validation based model by Yang et al. [15], a closed loop warning system proposed by Ramgopal et al. [16] and a regularity based seizure prediction model by Chien et al. [17].

3. SEIZURE PREDICTION ON STREAMING EEG DATA

Seizure prediction is the process of predicting preictal signals from interictal signals. Interictal state represents normal harmonious state of neural transmissions, while preictal states represents transmissions prior to seizures. Two other states; ictal and postictal are also represented in the signals, depicting seizure state and the state after the occurrence of seizure. Appropriate prediction of preictal state is of major importance, as this is the crucial stage, that when

appropriately detected can be used to alert for the onset of seizure. Real time prediction of preictal signals is important, as faster predictions can provide higher buffer time, which in-turn will be a boon for the patients. As the statistics states that seizures themselves are not fatal, however, they are fatal only due to their sudden onset and inappropriate counter-measures adopted. Higher time buffers can provide better situations where the patients can perform counter-measures and even avoid the onset of seizure.

The major disadvantage of the existing approaches is the tradeoff that had occurred between time and accuracy of predictions. In several approaches, time has been hugely compromised for the want of accuracy. This leads to late alerts, which tend to be potentially harmful. The proposed approach incorporates data aggregation strategies into the operational process such that the amount of data to be processed by the seizure prediction module is limited, hence exhibiting good accuracy levels within considerable time.

3.1 Data Preparation with Multilevel Feature Reduction

Data preparation is the first and one of the essential phases when dealing with EEG data. The major requirement of this module is the hugeness and generation rate of the data involved. EEG data is generated from 16 electrodes connected to the lobes of the human brain. The signals are divided into sequences, with each sequence containing 10 minutes of recorded signal information. Sampling rates are usually maintained at 400Hz. Hence 400 signals are generated by each channel per second, hence a total of 24000 signals are generated by each channel per minute. This aggregates to generation of 1.4 million signals per channel per hour. Generation of preictal signals from the lobes differ between individuals. Hence in-order to accurately identify the onset of seizures, it becomes mandatory to utilize data generated from all the 16 electrodes.

Utilization of information from the electrodes alone is not sufficient for accurate predictions, as variations between the interictal and preictal signals are the major indicators of the onset of seizure. Hence frequency based features are to be extracted from the channel information. However, not all features correspond to variations between interictal and preictal signals. Hence a feature shortlisting phase is incorporated into the model to reduce the processing overhead. The major reason for deriving features and then trimming them down is that signal properties differ between individuals. Features corresponding to preictal variations in one individual might not correspond to preictal variations in another individual. A different set of features might represent the variations in the second individual. Hence it is necessary to incorporate the feature shortlisting module within the architecture.

3.1.1. Data Aggregation

Data is usually processed as epochs. A single epoch corresponds to signals generated for 1 minute. Hence for every epoch, each channel generates 24000 signals. However, these signals are mostly similar to each other with slight and acceptable variations. Hence, a variance based data aggregation can reduce the data volume. Epochs are analyzed and if the variations within a single epoch is low, signal aggregation is performed, resulting in a huge reduction of data. However, a single epoch can contain both interictal and preictal data, hence an individual record analysis based aggregation becomes mandatory.

The data aggregator module processes epochs in a sequential manner, as they arrive. Epochs are collected for the required duration (1 minute) and single epoch batches are passed to the data aggregator. A variance matrix (V_{mat}) is created between two consecutive records for the entire epoch. The records in V_{mat} are analyzed for huge variations. Any huge variations are flagged for aggregation. Records between the previous and the current variations are aggregated into a single record. Algorithm for the data aggregation process is provided below.

Algorithm

1. Input epochs
2. For each epoch e
 - a. For each signal r in e
 - i. $V_{mat}[r] = \text{Variance between } e[r] \text{ and } e[r+1]$
3. For each record i in V_{mat}
 - a. If $i > \text{Thresh}_{var}$
 - i. Flag i
4. For each flag f
 - a. Aggregate records between $e[f]$ to $e[f+1]$

3.1.2. Feature Identification

The feature identification module is triggered after generation of 10 minute sequences, i.e. after generating 10 epoch modules. Although the epochs are aggregated and the number of epoch records are reduced, they are still maintained as ten individual epochs. The epochs are combined to generate a single epoch record set and is used for feature generation. It has been observed from literature [] that, frequency based analysis of EEG data exhibits better seizure detection levels compared to direct analysis. Hence frequency parameters are analyzed and created for the 10-minute epoch signals. The generated frequency parameters are used for seizure prediction rather than the actual EEG signals. The frequency based functions used for generating the EEG data include Fast Fourier Transformation (FFT), Shannon Entropy, Spectral Edge Frequency, Cross Correlation, Hjorth Parameters, Skewness, Kurtosis, Shannon entropy dyad, cross-correlation dyad, Mean and Median. The significance of these parameters and their operational nature have been discussed in detail by Ramina et al. in [18]. The epochs are converted to features and each feature vector contains their corresponding values calculated from the epoch signals.

3.1.3. Attribute Extraction

The feature identification process results in a huge data explosion, as the size of every feature vector is equivalent to that of the epoch signals. Although several features were created, not all features effectively correspond to the interictal or preictal states. Hence extracting the significant features can further reduce the data size, leading to faster seizure prediction. Each of the feature vectors are passed to the attribute extraction module, using which the appropriate channels are retained, while others are filtered. This module uses L1-based feature selection model for attribute extraction.

All the feature vectors are passed to the feature selection model and significant attributes are extracted from the vectors. All the feature vectors have same dimensions, however, after this phase, the feature vectors differ in their dimensions. Feature vectors with zero dimensions are eliminated from the list. The data preparation module results in a hugely reduced feature data, that can provide effective

predictions and also can effectively reduce the time of prediction.

3.2 Seizure Prediction using Gradient Boosted Trees

Data obtained from the data preparation phase is processed by the seizure prediction module for analysis. Its precise nature has the ability to provide better and faster predictions. This paper uses Gradient Boosted Trees (GBT) constructed with Decision Trees for seizure prediction. The feature vectors are passed to GBT and predictions are obtained. However, each feature vector generates its own prediction corresponding to its data. Hence voting is used to obtain the final prediction.

Boosting is the process of learning and re-learning from the data by back-propagation. Let T be the learning model, d be the training data and y be the trained model.

$$y = T(d)$$

However, the trained model y is not free from errors. The error component e is calculated by finding the difference between the actual prediction and the prediction from the trained model. This is given by

$$e = \Delta y - y$$

where Δy is the actual prediction.

Boosting back-propagates the errors to iteratively build better models such that the error levels are reduced within defined threshold levels. The second training level of GBT is given by

$$y' = T(d) + e$$

The process of training (eq. 1), error identification (eq. 2) and error incorporation (eq. 3) is iteratively performed to obtain the final model.

The major advantage of using boosting based model for seizure prediction is that, the domain of seizure prediction is dynamic and exhibits considerable variations between patients. Hence an iterative model can help achieve reduced error levels depending on the input signal data.

4. RESULTS AND DISCUSSION

Seizure data used in the proposed model was obtained from Kaggle provided by the American Epilepsy Society [19]. Data is in the form of .mat files, with 6 sequences for each 1 hour signal. Data preparation and seizure prediction phases are implemented 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 [20] is created. ROC and PR Plots [21] are constructed from the data and the algorithm efficiency is analyzed.

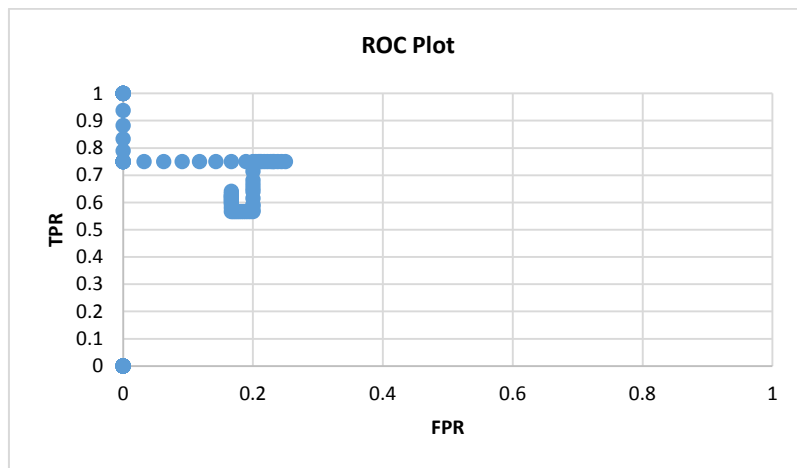


Fig 1: ROC Plot

ROC plot representing the efficiency of the proposed prediction model is shown in figure 1. It could be observed that the proposed model exhibits high TPR levels at an

average of 0.8 and low FPR levels at an average of 0.1. The plots can be observed to be grouped in the top left corner of the plot indicating effective predictions.

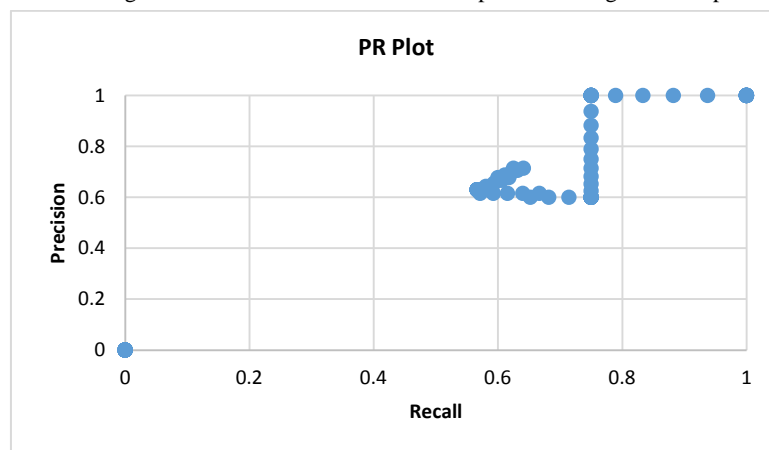


Fig 2: PR Plot

PR plots exhibiting the efficiency of the prediction of positive classes (seizure signals) is presented in figure 2. It could be observed that the precision levels range from 0.6 to 1, and the

recall levels range between 0.58 to 1. This exhibits effective retrieval rates of the proposed model.

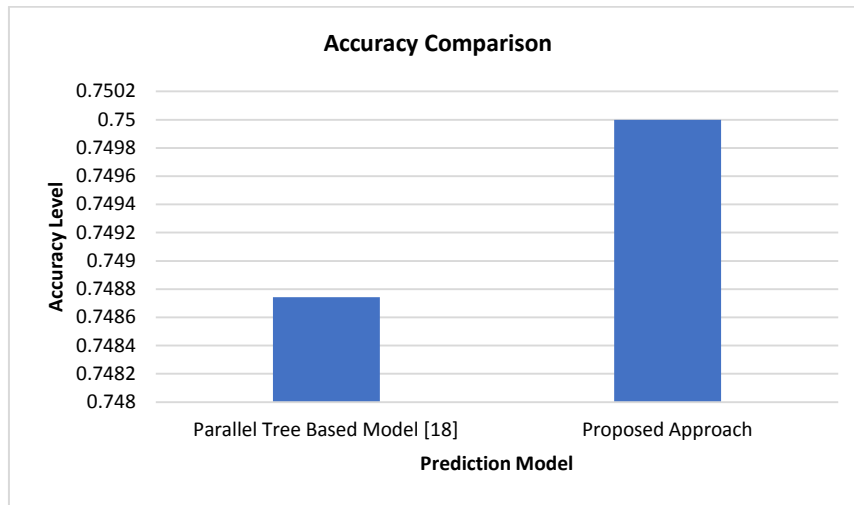


Fig 3: Accuracy Comparison

A comparison between the proposed approach and the parallel tree based approach proposed in [18] has been performed and is shown in figure 3. It could be observed that the proposed approach exhibits higher accuracy levels compared to the parallel tree based model, exhibiting the efficiency of the proposed approach.

Table 1: Performance Metrics

	Performance
Accuracy	0.75
F-Measure	0.666666667
TNR	0.833333333
FNR	0.375
FPR	0.166666667
TPR	0.625

Other performance metrics exhibited by the proposed model are shown in table 1. It could be observed that the proposed model exhibits an F-Measure level of 0.66 and an interictal prediction level of 0.833. The false positive levels were also found to be low at 0.16. However, it could be observed that the preictal prediction levels were moderate at 0.6325, which needs to be improved.

5. CONCLUSION

This paper presents an effective seizure prediction model that incorporates a data processing phase and the seizure prediction phase. The data processing phase performs feature identification and attribute aggregation and elimination. The attribute aggregation phase combines similar signal records to reduce the number of attributes being processed. The feature identification phase identifies frequency based features for enhanced prediction, while the final attribute elimination phase eliminates channels that do not contribute to the final prediction. The processed data is passed to GBT for seizure prediction. GBT is constructed with Decision Trees. The iterative error reduction model of GBT produces effective results when compared to other existing approaches. The

major advantage of this model is that it uses features rather than the actual data, leading to reduced workloads, hence faster and better predictions. Although the proposed model exhibits high performance in terms of accuracy, the seizure prediction level remains moderate. Future works will incorporate advanced strategies to improve the prediction levels. Future directions will concentrate on enhancing the Boosting process to incorporate multiple criterion to enhance the seizure prediction rates.

6. REFERENCES

- [1] Nasehi, S. and Pourghassem, H. 2012. Seizure detection algorithms based on analysis of EEG and ECG signals: a survey. *Neurophysiology* 44 (2), 174–186.
- [2] Shoeb, A., Schachter, S., Schomer, D., Bourgeois, B., and Guttag, J. 2005. Detecting seizure onset in the ambulatory setting: demonstrating feasibility. *Conf Proc of IEEE Eng Med Biol Soc*, 3546–3550.
- [3] World Health Organization. <http://www.who.int/mentalhealth/management/neurological/en/>, 2015.
- [4] Bellon, M., Panelli, R.J. and Rillotta, F. 2015. Epilepsy-related deaths: An Australian survey of the experiences and needs of people bereaved by epilepsy. *Seizure*, 29, 162–168.
- [5] Behnam, M. and Pourghassem, H. 2016. Real-time seizure prediction using RLS filtering and interpolated histogram feature based on hybrid optimization algorithm of Bayesian classifier and Hunting search. *Computer methods and programs in biomedicine*, 132, pp.115-136.
- [6] Sareen, S., Sood, S.K. and Gupta, S.K., 2016. An Automatic Prediction of Epileptic Seizures Using Cloud Computing and Wireless Sensor Networks. *Journal of medical systems*, 40(11), pp.226-226.
- [7] Chu, H., Chung, C.K., Jeong, W. and Cho, K.H., 2017. Predicting epileptic seizures from scalp EEG based on attractor state analysis. *Computer methods and programs in biomedicine*, 143, pp.75-87.

- [8] Scheffer, M., Bascompte, J., Brock, W.A., Brovkin, V., Carpenter, S.R., Dakos, V., Held, H., Van Nes, E.H., Rietkerk, M. and Sugihara, G., 2009. Early-warning signals for critical transitions. *Nature*, 461(7260), pp.53-59.
- [9] Dakos, V., Carpenter, S.R., Brock, W.A., Ellison, A.M., Guttal, V., Ives, A.R., Kefi, S., Livina, V., Seekell, D.A., van Nes, E.H. and Scheffer, M., 2012. Methods for detecting early warnings of critical transitions in time series illustrated using simulated ecological data. *PLoS one*, 7(7), p.e41010.
- [10] Cotilla-Sanchez, E., Hines, P.D. and Danforth, C.M., 2012. Predicting critical transitions from time series synchrophasor data. *IEEE Transactions on smart grid*, 3(4), pp.1832-1840.
- [11] Chen, L., Liu, R., Liu, Z.P., Li, M. and Aihara, K., 2012. Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers. *Scientific reports*, 2.
- [12] Sharif, B. and Jafari, A.H., 2017. Prediction of epileptic seizures from EEG using analysis of ictal rules on Poincaré plane. *Computer Methods and Programs in Biomedicine*, 145, pp.11-22.
- [13] Fei, K., Wang, W., Yang, Q. and Tang, S., 2017. Chaos feature study in fractional Fourier domain for preictal prediction of epileptic seizure. *Neurocomputing*, 249, pp.290-298.
- [14] Aarabi, A. and He, B., 2012. A rule-based seizure prediction method for focal neocortical epilepsy. *Clinical Neurophysiology*, 123(6), pp.1111-1122.
- [15] Yang, A., Arndt, D.H., Berg, R.A., Carpenter, J.L., Chapman, K.E., Dlugos, D.J., Gallentine, W.B., Giza, C.C., Goldstein, J.L., Hahn, C.D. and Lerner, J.T., 2015. Development and validation of a seizure prediction model in critically ill children. *Seizure*, 25, pp.104-111.
- [16] Ramgopal, S., Thome-Souza, S., Jackson, M., Kadish, N.E., Fernández, I.S., Klehm, J., Bosl, W., Reinsberger, C., Schachter, S. and Loddenkemper, T., 2014. Seizure detection, seizure prediction, and closed-loop warning systems in epilepsy. *Epilepsy & behavior*, 37, pp.291-307.
- [17] Chien, J.H., Shiao, D.S., Halford, J.J., Kelly, K.M., Kern, R.T., Yang, M.C., Zhang, J., Sackellares, J.C. and Pardalos, P.M., 2011. A signal regularity-based automated seizure prediction algorithm using long-term scalp EEG recordings. *Cybernetics and Systems Analysis*, 47(4), pp.586-597.
- [18] P. Ramina and M. Vanitha, 2017, "Epileptic Seizure Prediction in EEG Records using Parallel Tree Based Learning and Feature Extraction", *Indian Journal of Science and Technology*.
- [19] Melbourne University AES/MathWorks/NIH Seizure Prediction <https://www.kaggle.com/c/melbourne-university-seizure-prediction>. Date accessed: 10/06/2016
- [20] Confusion matrix https://en.wikipedia.org/wiki/Confusion_matrix Date accessed: 10/06/2016
- [21] Fawcett, T., 2006. An introduction to ROC analysis. *Pattern recognition letters*, 27(8), pp.861-874.