A novel method of EEG data acquisition, feature extraction and feature space creation for Early Detection of Epileptic Seizures

Sylvia Bugeja, Lalit Garg, and Eliazar E. Audu,

University of Malta, Msida, Malta

Abstract— **In this paper we describe a simple and very fast method of data acquisition, feature extraction and feature space creation for epileptic seizure detection. The scalp electroencephalogram (EEG) dataset [1, 2] collected at the Children's Hospital Boston from 22 pediatric patients having 192 intractable seizures (available as CHB-MIT database) is used to assess this simple approach against existing ones [1, 3], with very positive results reaching up to 99.48% Sensitivity.**

I. Introduction

Epilepsy is a neurological disease, which affects around 50 million people of the world's population [4]. People suffering from epilepsy experience involuntary recurrent seizures, which happen due to abnormal electrical activity in the brain. Unfortunately, epileptic seizure detection happens quite late when the patient is already experiencing bad symptoms. For instance, a patient can experience an epileptic seizure without any warnings, with a probability that the patient suffers a dangerous fall with severe injuries. With the increased development of effective prevention treatments, early diagnosis of epileptic seizures is becoming necessary because the patient can undergo treatments, which can delay or prevent the disease progression.

A number of studies [5] have been carried out in the past to explore the feasibility of a practical real-time epilepsy seizure detector. The aim of this paper is to propose a novel method of data acquisition, feature extraction and feature space creation for epilepsy seizure detection. This method differs from previous studies mainly on two things; the first is providing a simple yet very effective training set acquisition for epileptic seizure detection and the second is testing this novel approach using a high number of seizure instances, precisely a total of 192 seizures from total 22 pediatric patients.

II. ELECTROENCEPHALOGRAM (EEG)

Electroencephalogram (EEG) is a multi-electrode recording of the current flows, which are produced by the millions of neurons residing in our brain. In order to record scalp EEG, the electrodes are symmetrically placed on the scalp to measure the brain's spatial and temporal data. The spatial data consists of the brain's electrical activity emerging from a particular brain region, while the temporal data describes how the brain's electrical activity changes over time [6-8]. EEG data can be used to detect abnormal brain activity

related to Epilepsy, which is manifested by reduction in amplitude, or frequencies beyond the normal limit, or production of spike patterns [9].

A. Artifacts

Artifacts are signal distortions, which are not related to abnormal EEG, and they can be experienced by people who do not suffer from Epilepsy. Artifacts can be either patient-related or technical-related. Patient-related artifacts are common biological signals that can disturb the EEG signals, such as electrical activity from heartbeats, which produce sharp wave artifacts. Technical-related artifacts are related to malfunctioning electrodes or electromagnetic interferences [12, 13]. For better EEG interpretation, one can exclude artifacts from an EEG recording; else, their characteristics should be studied in detail, in order to distinguish them from abnormal EEG.

III. EPILEPTIC SEIZURE DETECTION

Previous studies have used different methods to detect epileptic seizures, however their main goal was one, that to derive a number of feature vectors from the EEG signals and to classify them into labels which show an epileptic seizure or not.

A. Feature Vector Design

The goal of feature vector design is to transform EEG signals into feature vectors by extracting most important features of a signal from EEG data that provide sufficient information to distinguish between seizures and non-seizures state. The features selected are magnitude (spike), spectral energy variation within the clinical relevance frequency of 0.5- 25Hz, and morphology of the signals, which are mapped into label vectors, called the feature (vector) space. Since EEG signals are transient, and highly dynamic, feature vectors are formed for each time epoch. The Multi-level Wavelet Decomposition is a popular technique used in previous studies [3, 10, 11, 12, 13, 15-18] for feature vector extraction. This technique decomposes an EEG signal into a number of sub-band signals each depicting a different waveform morphology within a particular frequency range. Spectral features from each sub-band signal are extracted in order to form a feature vector, which will represent the original EEG signal. For instance, one type of spectral feature is the energy falling within a sub-band signal. The maximum, minimum and or mean frequencies of these energy values are captured for better discrimination between abnormal and

S. Bugeja is with the University of Malta, Malta (e-mail: sylvia.bugeja.08@um.edu.mt).

L. Garg is also with the University of Malta, Malta (phone: +356-2340-2112; e-mail: lalit.garg@um.edu.mt).

E. E. Audu is also with the University of Malta, Malta (e-mail: ideologye@yahoo.com).

978-1-4577-0220-4/16/$31.00 ©2016 IEEE

normal EEG signals. Fig. 1 below shows a transformation of EEG-signal into feature space during feature extraction process and Fig. 2 mathematically represents the formation of feature vector.

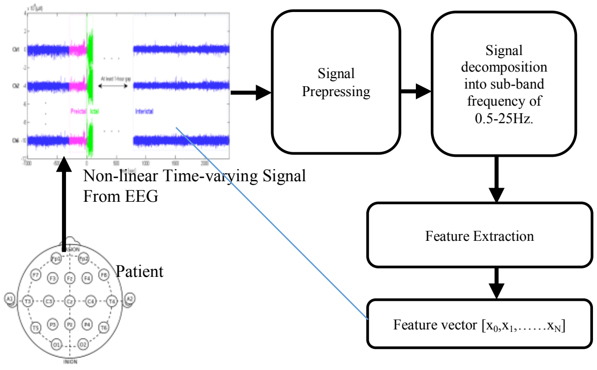
Signal Prepressing
Signal decomposition into sub-band frequency of 0.5-25Hz.
Feature Extraction
Feature vector [x_0,x_1,……x_N]
Non-linear Time-varying Signal From EEG
Patient
Preictal
Ictal
Interictal

Figure 1. EEG feature extraction process

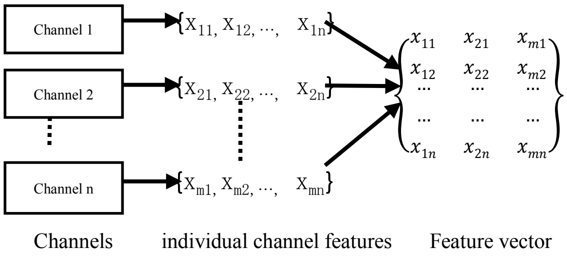
Channel 1
Channel 2
Channel n
{X_{11}, X_{12}, ···, X_{1n}}
{X_{21}, X_{22}, ···, X_{2n}}
{X_{m1}, X_{m2}, ···, X_{mn}}
x_{11} x_{21} x_{m1}
x_{12} x_{22} x_{m2}
x_{1n} x_{2n} x_{mn}
Channels
individual channel features
Feature vector

Figure 2. Feature Vector (Space)

In their study, Shoeb et al. [3] describe that since epileptic seizure characteristics in EEG signals are non-stationary, it is important that spectral features are calculated at multiple time epochs. In fact, studies [3, 11, 16] use a sliding window to capture the spectral features within multiple two-second epochs.

B. Feature Vector Classification

Feature Vector Classification methods such as support vector machine (SVM) and extreme learning machine (ELM) have been used by various studies [3, 11, 16, 17, 19] to classify feature vectors into labels which either show an epileptic seizure or not. The architecture of a SVM consists of a single-layer feed forward network were an input feature vector is mapped into a higher dimensional space and the weights are adjusted only from the last hidden layer to the output layer. ELM works with a generalized single hidden feed forward, were if the activation functions in the hidden layer are infinitely differentiable, the input weights layer and hidden layer biases can be randomly chosen without tuning. This makes ELM perform better and with less human intervention. These supervised machine-learning systems need training data to learn how to classify feature vectors into their appropriate labels. This training data will consist of labeled feature vectors, where each labeled feature vector is characterized by its corresponding known class.

Previous studies use a leave-one-out cross validation scheme, to estimate their detector's performance. This is calculated as follows. Let X denote the total number of seizure and non-seizure records in a dataset. The classification system is first trained on X-1 records and tested on the untrained record. This process is repeated X times until all records are used once for testing.

IV. PROPOSED METHOD

The main advantage of this method is in creating a simple and yet very effective training set acquisition for epileptic seizure detection. Having a simple training set makes the classifier's training phase faster. Also, most studies [11, 12, 15, 18-20] use a small data set [21] having a total of only 39 minutes seizure data. We tested our method on an EEG dataset of 22 patients (5 males and 15 females ages 1-22) collected at the Children's Hospital Boston (available as CHB-MIT database) [1, 2]. The dataset contains 977 hours of electroencephalogram (EEG) data grouped into 23 test cases (two cases were from the same female patient but collected at 1.5 years' time interval) with 192 intractable seizures, which amount to 188 minutes seizure data. This means that results obtained from this study are more accurate. This is because it is important that the final feature vectors' data is consistent. Choosing the CHB-MIT dataset was very vital because it's one of the characteristics, which differs our study from previously epileptic seizure detector systems. This is because previous studies use datasets, which have small amount of seizure data. The international 10-20 system of EEG electrode positions was used to record the EEG data. The majority of the cases have 23 channel signals, but there are a few which contain either 24 or 26 channel signals [1]. For better feature vector design and classification, we designed the same channel numbering and positioning system for each patient.

A. Data Pre-Processing

Data was downloaded from the PhysioNet website [2]. Artifact channels such as ECG data were removed from the EEG data for better epileptic seizure detection. Since some patient's data was taken at different time intervals with different multi-electrode systems (for example, two cases from the same female patient were collected at 1.5 years' time interval), the data was altered such that consistent number of channel data was stored for each patient. Also, the EEG data was epoched into two seconds temporal segments.

B. Training Subset Acquisition

The pre-processed EEG data was then divided into N minute subsets, where each subset contained one seizure data. This was done by fetching N/2 minutes of EEG data before the start of a seizure and N/2 minutes of EEG data after the end of a seizure. Thus, if a patient has 2 seizures and SL_i is the length of seizure i, then $\{(N+SL_1) + (N+SL_2)\}$ minutes subsets are fetched for that particular patient. This means that, assuming N=10-minutes long and both SL_i are 1-minute-long, a one hour patient's data is reduced to 22 minutes training data which is simpler but at the same time containing the most important features which were there in the original seizure data. In this study N=20-minute and N=20-minute training subsets are separately used as training data in order to investigate the effect the length of the training data has on the results obtained. The advantage of this simple training set acquisition is that although the training sets are simpler they are still effective since they contain the most important features which are the original seizure data and some non-seizure data located before and after the seizures. Also, all the non-seizure data was included as part of the testing data set.

Multi-level Wavelet Decomposition was used to extract waveforms of frequency range between 1 Hz to 10 Hz. The mean frequencies of the energy values falling within these waveforms were used to create the feature vectors of the training subsets. Also, a sliding window capturing 3 contiguous 2 second epochs is used to capture the time-evolution characteristics of the EEG signals.

C. *Testing Phase*

Both SVM and ELM were used to classify the 10-minute and 20-minute feature vectors into two classes; a seizure class and a non-seizure class. Also, a leave-one-out (three fold) cross validation scheme is used to test their performance. The training subsets are divided into three subsets. Two out of the three subsets are then used as training dataset and whilst the third subset is used as a testing dataset. Precisely, each 10-minutes training subset residing in the third subset is added to the testing dataset and used once as a seizure test case. The three-fold cross validation process is repeated three times until all test seizures in each of the latter three subsets are used once for testing. Therefore, the testing dataset would always contain a single seizure. The advantage of this fold cross validation method is that although the detector is trained using a simple training dataset, it is tested with a huge amount of seizure and non-seizure EEG data, a total of 198 seizures and a total of 977 non-seizure EEG data.

a) D. *Performance metrics*

Sensitivity, Specificity, and Latency used as Performance metrics. Sensitivity refers to the percentage of test seizures, which are correctly identified by the detector. Sensitivity is defined in (1):

$$Sensitivity = \left(1 - \frac{T_{seizures\ detected}}{T_{seizures}}\right) * \ 100 \qquad (1)$$

where $T_{seizures\ detected}$ is the total number of seizures detected/identified correctly, while $T_{seizures}$ is the total number of seizures occurred.

Specificity refers to percentage of non-test seizures which where correctly identified as non-seizures. Specificity is defined in (2):

$$Specificity = \left(1 - \frac{T_{falsepositives}}{T_{non-seizure}}\right) * \ 100 \qquad (2)$$

where, $T_{falsepositives}$ is the total number of false positive epochs, while $T_{non-seizure}$ is the total number of non-seizure epochs.

Latency refers to the delay from the time a seizure actually occurs to the time when the detector declares the seizure activity. Latency is defined in (3)

$$Latency = (t_{Sezure\ detected} - t_{Sezure\ occured}) \qquad (3)$$

where $\mathrm{T}_{\mathrm{falsepositives}}$ is the time when the detector declared that the seizure is detected, while $t_{Sezure\ occured}$ is the time when the seizure is actually occurred.

V. Results

Results obtained from this method are compared with Shoeb et al. study [3]. This is because although less seizure test cases were used, than in our study, it is still the only study which tested it's detector with a large amount of seizure test cases, precisely a total of 173 seizures from the same dataset.

With 192 test-seizures, both 10, 20 and 30-minutes subsets reached high Sensitivity values with SVM reaching 95.33%, 95.42%, and 97.98% respectively. ELM reached 99.48%,99.48% and 98.99% Sensitivity values when trained on both 10, 20 and 30-minutes subsets. When compared with Sheob et al.'s study [17], which when tested on 173 seizures reached a 96% Sensitivity; results obtained from this study prove that the simple training set acquisition proposed in this method is very effective and efficient (fast) as summarized in the Table 1 below. Fig. 3 and Fig. 4 show the performance comparison between the results.

As shown in Table 1 and Fig. 5 above, a 0.97 seconds Latency was obtained by the ELM when trained on both 10-minute and 20-minute subsets. This latency result is also very positive when compared with the 3 seconds latency result obtained by Shoeb et al.'s study [3]. Furthermore, with 30-minute subsets, a latency of 1.26 seconds was obtained, which is still better than the previous study [3].

TABLE I. 10, 20 AND 30-MINUTES SUBSETS' RESULTS

	10-Minute Subsets		20-Minute Subsets		30-Minutes Subset	
	SVM	ELM	SVM	ELM	SVM	ELM
Sensitivity(%)	95.33	99.48	95.42	99.48	97.98	98.99
Specificity(%)	87.11	74.21	89.90	77.16	83.73	81.39
Latency(Seconds)	3.18	0.97	2.88	0.97	2.95	1.26

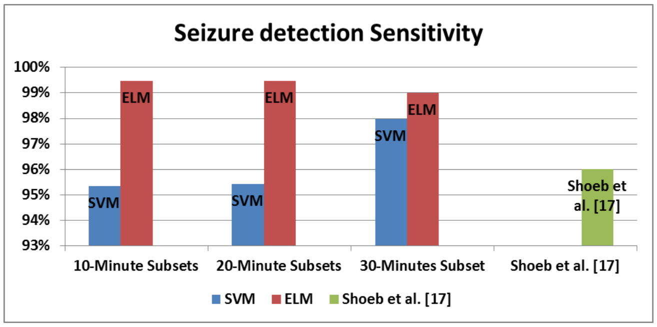

Figure 3. Seizure detection Sensitivity: SVM vs ELM and Shoeb et al. [17].

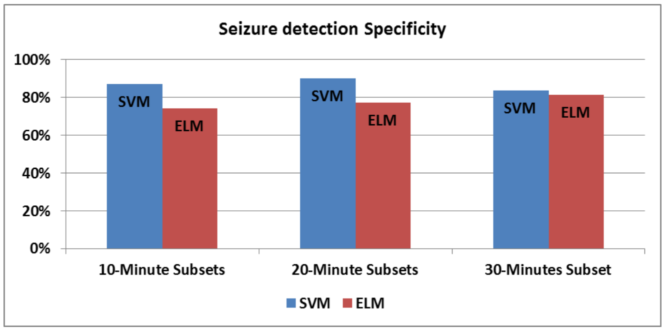

Figure 4. Specificity: SVM and ELM performance metrics.

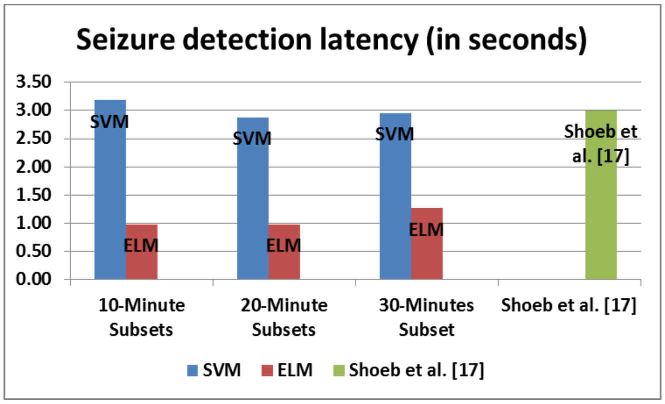

Figure 5. Latency: SVM and ELM.

Shoeb et al. [17] use a different definition of specificity than its definition in (2), and therefore, we are not comparing their specificity results with the specificity of the seizure detector we have proposed.

Although the lengthier the subsets training data the better the Specificity, Specificity still needs to be improved. This can be improved by testing the detector with all the original EEG data. This means that the detector will still be trained with simple training data sets but instead of testing it with a 10-minute, 20-minute or 30-minute subset data, the detector will be tested with all the original EEG Data. This will improve the Specificity since using (2) the proportion between $T_{falsepositives}$ and $T_{non-seizure}$ will be larger, where $T_{falsepositives}$ is the number of false positives declared by the detector and $T_{non-seizure}$ is the total number of non-seizure epochs in the testing data.

VI. Conclusion

This paper proposed a simple and effective training set acquisition method, which was tested using Multilevel Wavelet Decomposition as a feature vector design process and both SVM and ELM as feature classification methods. Multi-channel data was fed at once to each classification technique. The proposed method was tested using a testing methodology, which does not yield overly optimistic results. In other words, the proposed method was tested with a high number of seizure and non-seizure EEG data, precisely with more than 185 seizure instances and 977 hours of EEG data.

Results obtained from the proposed method is remarkable and demonstrate that this simple training set acquisition is not only very effective but even better than other training methodologies. Results obtained from the proposed method provide a very good foundation for simple and effective epileptic seizure detectors to be built in the near future.

Results obtained from the proposed method show that the 10-minutes training subsets, perform as good as lengthier training subsets. Also, when classifying multi-channel data at once, the ELM classification technique performs better than the SVM classification technique. ELM's results reach more than 99% 'Sensitivity' and less than 2 seconds 'Latency' which are better than the 96% 'Sensitivity' and 3 seconds 'Latency' obtained by Shoeb et al. [3, 17].

References

[1] A. H. Shoeb. "Application of Machine Learning to Epileptic Seizure Onset Detection and Treatment". Ph.D. Thesis, Harvard-MIT Program of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, MA, September 2009.

[2] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. -K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals", *Circulation*, vol. 101, no. 23, pp. e215-e220, June 2000.

[3] A. H. Shoeb and J. V. Guttag, "Application of machine learning to epileptic seizure detection", *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp.~975-982, 2010.

[4] *Epilepsy, Fact sheet*, World Health Organization (WHO). February 2016. [Online] Available: http://www.who.int/mediacentre/factsheets/fs999/en/.

[5] S. Bugeja, and L. Garg, "Applications of Machine Learning Techniques for the Modelling of EEG Data for Diagnosis of Epileptic Seizures", *The 3rd Workshop on Recognition and Action for Scene Understanding (REACTS 2015) Valletta, Malta*, September 5 2015.

[6] C.D. Binne, and P.F. Prior," Electroencephalography" Journal of neurology, neurosurgery and psychiatry, Vol.57, No.11, pp.1308-1319, 1994.

[7] S. Parmet, C. Lynm and R.M. Golub, "JAMA Patient Page Epilepsy", The Journal of the American Medical Association, Vol.305, No.16, 1722, 2011.

[8] M. Teplan, M. "Fundamentals of EEG measurement", Measurement *science review*, Vol.2, No.2, pp. 1-11, 2002.

[9] A. Shoeb, J. Guttag and S. Cash "Patient-Specific Seizure Onset Detection", [Online] Available: http://dspace.mit.edu/bitstream/handle/1721.1/17991/57194544.pdf.

[10] S.M. Akareddy and P.K. Kulkarni, "EEG signal classification for epilepsy seizure detection using improved approximate entropy". *International Journal of Public Health Science (IJPHS),* Vol.2, No.1, pp.23-32, 2013.

[11] N. Fatma Guler, and E.D. Ubeyli, "Multiclass support vector machines for EEG-signals classification", *Information Technology in Biomedicine, IEEE Transactions on*, Vol.11, No.2, pp.117-126, 2007.

[12] L. Guo, D. Rivero, and A. Pazos, "Epileptic seizure detection using multiwavelet transform based approximate entropy and artificial neural networks", *Journal of neuroscience methods*, Vol.193, No.1, 156-163, 2010

[13] T. Lajnef, S. Chaibi, A. Kachouri, and M. Samet, "Epileptic Seizure Detection: Approximate Entropy and Discrete Wavelet Transform based method", *Third International Conference: E-Medical Systems*, 2010.

[14] T.-P. Jung and S. Making, "Artifact removal from EEG", Available Online fromhttp://sccn.ucsd.edu/~jung/artifact.html.

[15] A.M. Murugavel, and S. Ramakrishnan, "Wavelet Domain Approximate Entropy-Based Epileptic Seizure Detection", *The 5th International Conference on Information Technology*, 2011.

[16] C.P. Shen, C.M. Chan, F.S. Lin, M.J. Chiu, J.W. Lin, J.H, Kao, and F. Lai, "Epileptic seizure detection for multichannel EEG signals with support vector machines", *Bioinformatics and Bioengineering (BIBE), IEEE 11th International Conference*, pp.39-43, 2011.

[17] A. Shoeb, H. Edwards, J. Connolly, B. Bourgeois, S. Ted Treves, and J. Guttag, "Patient-specific seizure onset detection", *Epilepsy & Behavior*, Vol.5, No.4, pp. 483-498, 2004.

[18] H. Vavadi, A. Ayatollahi, and A. Mirzaei, "A wavelet-approximate entropy method for epileptic activity detection from EEG and its sub-bands", *Journal of Biomedical Science and Engineering*, Vol.3, No.12, pp 1182-1189, 2010.

[19] L.L. Chen, J. Zhang, J.Z. Zou, C.J. Zhao, and G.S. Wang, "A framework on wavelet-based nonlinear features and extreme learning machine for epileptic seizure detection", *Biomedical Signal Processing and Control*, Vol.10, pp. 1-10, 2013.

[20] Y. Song, J. Crowcroft, and J. Zhang, "Automatic epileptic seizure detection in EEGs based on optimized sample entropy and extreme learning machine", *Journal of neuroscience methods*, Vol.210, No.2, 132-146, 2012.

[21] EEG time series. (2005). [Online] Available: http://www.meb.uni-bonn.de/epileptologie/science/physik/eegdata.html.