Seizure Detection and Prediction Project Proposal CS534

Group members: Daniela Chanci, Anu Trivedi, David Wang

October 6, 2021

1 Introduction

Epilepsy is a common brain disorder that affects approximately 70 million people in the world [1]. It is characterized by the predisposition to generate seizures due to an abnormal and excessive neuronal discharge [2]. These epileptic seizures can cause a number of behaviors, such as shaking and loss of awareness, and they can even result in accidents and death. There are four phases of seizures: (1) interictal state, between seizures; (2) preictal state, before seizure; (3) ictal, during seizure; and (4) postical, after seizure [3]. In this regard, we can use physiological signals, such as electrocardiogram (ECG) and electroencephalogram (EEG) to monitor and study the seizures. Brain activity has shown that seizures develop several minutes to hours prior to their clinical onset [4].

In this project, we use the EEG signal, combined with machine learning algorithms, for the accurate detection and prediction of seizures, i.e., the goal is to identify the ictal and preictal states. In the following sections, we present the specific aims, provide a brief description of the dataset, and present the proposed approaches, as well as the selected off-the-shelf machine learning algorithms for subsequent comparison and evaluation.

2 Specific Aims

To classify different EEG segments into interictal - ictal states, and estimate the seizure likelihood, we propose the following specific aims:

Specific Aim 1: Implement a machine learning algorithm for seizure detection using EEG data.
Specific Aim 2: Implement a machine learning algorithm for seizure prediction using EEG data.
Specific Aim 3: Validate the developed machine learning algorithms using off-the-shelf algorithms.

3 Dataset

The two datasets used for this project are obtained from two Kagle competitions. The first one is the UPenn and Mayo Clinic's Seizure Detection Challenge. It is divided into training and testing data that consists of 1-second EEG clips with the corresponding label: ictal or interictal. The intracranial EEG (iEEG) data was collected from 4 dogs with epilepsy, sampled at 400 Hz. This dataset also includes iEEG from 8 human subjects with epilepsy, sampled at 500 and 5000 Hz [5]. The second one is the American Epilepsy Society Seizure Prediction Challenge. For this dataset, the data collection task was conducted in the manner as in the previous one, except for the sampling frequency of the human subjects EEG, which was only 5000 Hz. It is divided into training and testing data that consists of 10 minutes EEG clips with the corresponding label: preictal or interictal. The preictal clips are EEG recordings within one hour before the seizure. The five minutes before the seizure were not considered [6].

4 Methodology

4.1 Data Preprocessing

Before starting with the implementation of the machine learning algorithms, it is essential to preprocess the raw data from the datasets. In this regard, we will first filter the EEG signal to reduce the noise. Then, for the seizure prediction dataset, which contains EEG segments of 10 minutes, we will create new smaller windows of 1 minute to alleviate the computational requirements and facilitate the interpretation of the data. Last, we will apply different feature selection techniques to reduce the dimensionality, and obtain representative features in the time domain and in the frequency domain that can be used to successfully train the classifiers.

4.2 Advanced Machine Learning Algorithms - "Stretch Algorithm"

With the goal of obtaining satisfactory results for seizure detection and prediction, we propose the implementation of the following algorithms:

- Learning vector quantization (LVQ) is a nearest prototype classification (NPC) algorithm related to K-nearest neighbo (kNN) [7, 8]. kNN and NPC are both local classification techniques, but NPC methods do not use the entire training dataset, instead they use specifically selected prototype vectors. Besides performance, this also reduces the computational complexity of the algorithm, making LVQ a popular choice for real-time applications, and thus may be suitable for seizure detection and prediction [9].

- Classification and Regression Trees (CART) algorithm is a classification algorithm made to build binary decision trees by splitting node into two child nodes repeatedly based on the Gini's impurity index. The Gini index calculates the probability of misclassification of a feature so it can determine the optimal split from the root node and any subsequent splits. In terms of efficacy, for identifying risk factors, CART may outperform logistic regression. [10] CART identified more risk factors by using tree-splitting methods to represent the risk factors instead of using odd ratios for significant factors as seen in logistic regression. [10] Alongside with this, the added Gini impurity index will give increased accuracy to seizure prediction compared to logistic regression.

4.3 Off-the-shelf Machine Learning Algorithms

In order to evaluate our algorithm, we will compare our results with the ones obtained with off-the-shelf machine learning algorithms. In this regard, to simplify this task, we selected algorithms that are readily available in the scikit-learn library. A brief explanation of each of them is provided below:

- **K-nearest neighbor (kNN)** is a simple supervised classification algorithm, popular in cases where there is minimal prior knowledge about the data distribution [11]. The output is the label of the corresponding class. In the seizure detection problem, we consider two classes: interictal and ictal. A common validation strategy for this algorithm is to obtain the confusion matrix.

- **Logistic regression** is a predictive modeling algorithm used primarily as a binary classifier with posteriors bounded by two class labels. The result should be a projection that maximizes the separation between the two classes of each of the addressed problems for a bigger prediction horizon [12].

- **Support Vector Machines (SVM)** is a classical and robust machine learning classifier. We will try different kernels to find the optimal hyperplane that separates one class from the other [13]. In the detection problem it will classify the EEG segments into ictal and interical, while in the prediction proble, it will classify the segments into preictal and interictal.

4.4 Evaluation Metrics

To quantitatively assess the performance of our machine learning models and compare them with the off-the-shelf methods, we will use the following evaluation metrics: accuracy, recall, precision, and the area under the receiver operating characteristic (AUROC) curve.

5 Timeline

October 12	Getting familiar with the data (Daniela)
October 12	Data preprocessing (David)
November 2	Implementation and training of our stretch algorithm (All)
November 2	Evaluation of off-the-shelf algorithms (Anu)
November 16	Validation of the models (All)
December 6	Report and presentation preparation (All)

References

[1] Roland D Thijs et al. "Epilepsy in adults". In: *The Lancet* 393.10172 (2019), pp. 689–701.

[2] Carl E Stafstrom and Lionel Carmant. "Seizures and epilepsy: an overview for neuroscientists". In: *Cold Spring Harbor perspectives in medicine* 5.6 (2015), a022426.

[3] Waleed Abood and Susanta Bandyopadhyay. "Postictal Seizure State". In: *StatPearls [Internet]* (2020).

[4] Brian Litt and Javier Echauz. "Prediction of epileptic seizures". In: *The Lancet Neurology* 1.1 (2002), pp. 22–30. ISSN: 1474-4422. DOI: https://doi.org/10.1016/S1474-4422(02)00003-0. URL: https://www.sciencedirect.com/science/article/pii/S1474442202000030.

[5] *UPenn and Mayo Clinic's Seizure Detection Challenge*. URL: https://www.kaggle.com/c/seizure-detection/data.

[6] *American Epilepsy Society Seizure Prediction Challenge*. URL: https://www.kaggle.com/c/seizure-prediction/data.

[7] Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York, 2001.

[8] Sambu Seo and Klaus Obermayer. "Soft learning vector quantization". In: *Neural computation* 15.7 (2003), pp. 1589–1604.

[9] Takashi Komori and Shigeru Katagiri. "Application of a generalized probabilistic descent method to dynamic time warping-based speech recognition". In: *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 1. IEEE. 1992, pp. 497–500.

[10] Reiczigel Nagy, Schrott Harnos, et al. "Tree-Based Methods as an Alternative to Logistic Regression in Revealing Risk Factors of Crib-Biting in Horses". In: *Journal of Equine Veterinary Science* 30.1 (2010), pp. 21–26. DOI: https://doi.org/10.1016/j.jevs.2009.11.005.

[11] Leif E Peterson. "K-nearest neighbor". In: *Scholarpedia* 4.2 (2009), p. 1883.

[12] Tham Nusinovici, Ting Yan, Sabanayagam Li, et al. "Logistic regression was as good as machine learning for predicting major chronic diseases". In: *Journal of Clinical Epidemiology* 122 (2020), pp. 56–69. DOI: https://doi.org/10.1016/j.jclinepi.2020.03.002.

[13] Derek A Pisner and David M Schnyer. "Support vector machine". In: *Machine Learning*. Elsevier, 2020, pp. 101–121.