Project Name:

Unsupervised Learning of EEG States for the Application of Seizure Detection in Epilepsy

Project Number:

2022659

Supervisor:

Roman Genov

Group members:

Mehak Kalra (1004702654)

Fatima Siddiqui (1005069916)

Khalil Scott (1005352244)

Mina Assaad (1004926510)

1.0 Executive Summary

Clinicians researching seizures and epilepsy often spend significant time in cleaning the data as they must annotate the Electroencephalography (EEG) data to differentiate epileptic and normal brain activity. To reduce the time spent labelling the data, the Intelligent Sensory Microsystems Laboratory (ISML) is looking to automate the process using machine learning (ML). The team has been tasked with designing a software toolbox that uses digital signal processing (DSP) and ML to preprocess unlabelled EEG data. The scope of the project is mainly focused on designing a preprocessing labelling algorithm for EEG data, along with training and inference with off-the-shelf algorithms, and a visual interface to showcase the trained data. The project is currently in its initial state and the team has come up with various requirements to develop a design solution. The first set of requirements is the functions, which detail the underlying workflow that must be executed by the software toolbox. The second set is the objectives, which are the quantifiable goals the toolbox should reach. The last set is the constraints, which are the hard limits that the toolbox must abide by.

2.0 Motivation

Currently, brain activity is manually annotated by clinicians as EEG data is collected from patients. The process of annotating data can take hours of the clinicians' time, therefore we are attempting to automate the process using ML algorithms [1]. However, EEG data can be very complex because the signals are nonlinear, constantly changing, and noisy, which makes it difficult for ML algorithms to label the data [2]. The wave-like characteristics of EEG data mean there are multiple varieties of preprocessing methods that could either prioritize its frequency, periodicity, or both [3]. There are three potential approaches to preprocessing EEG data: clustering, deep learning, and stacked autoencoders. Clustering involves first reducing the complexity of the data, then using modified versions of clustering algorithms, such as K-means, to label the data. The reduction can be done either by using an approximation of the data, or by performing feature extraction [4]. Within deep learning, a convolution neural network will be used to learn the wave-let features of the EEG. Stacked autoencoders can be used to reduce dimensionality of the data to extract complex features and input them into a fine-tuned classifier to label the data [5].

3.0 Problem Statement

EEG data can be used to detect future occurrences of epileptic seizures using supervised ML. This process requires training labels to differentiate existing epileptic seizures from normal physiological brain activity. This project aims to automate the labelling of EEG data by using an algorithm to assist clinicians in the data annotation process.

4.0 Project Goal

The goal of this project is to create a software toolbox that assists clinicians in the data labelling process. This toolbox will utilize DSP and ML to perform feature extraction and label recorded data. The performance of the software toolbox will be tested against the ground truth labels in the CHB-MIT EEG dataset [1].

5.0 Scope of Work

The project's scope will be largely focused on developing ML algorithms to label EEG data and separate epileptic seizures from normal brain activity. After we have completed this main component in the project, we will include additional smaller components like a visual interface for the users on which they can upload Neurodata Without Borders (.nwb) files containing EEG data and choose an off-the-shelf supervised ML algorithm to train on labelled data and view the results, as requested by our supervisor [1]. We will experiment on three different preprocessing algorithms from general time series, clustering, and deep learning and compare results using our listed objectives. The final preprocessing solution will be the model with the highest overall performance. The project will be designed using PyTorch, SKlearn, and Google Collab. The dataset we will be using to train, validate, and test our ML algorithm is the CHB-MIT Scalp EEG Database. The dataset consists of 969 hours of EEG signals taken from 22 subjects with 173 seizures being recorded [6]. The dataset contains labels, which will be used to validate the accuracy of our ML algorithm. An additional dataset that will be used to perform a secondary test is an EEG dataset with HFO markings, which contains 90 hours of EEG signals recorded from 30 patients [7].

6.0 Requirements specification

This section highlights the main functions, objectives and constraints of the project.

6.1 Functions

The functions detail what the software toolbox must do to automate the labelling of EEG data. The functions have been separated into categories: the primary and secondary functions. The primary functions are the software toolbox's main tasks to operate as a whole, while the secondary functions outline the sub-tasks that will enable the primary functions.

Table 1. Primary and Secondary Functions of the Software Toolbox

Primary Functions	**Secondary Functions**
1. Pre-process input data for ML algorithms	1. Perform feature extraction on input data

2. Integrate with off-the-shelf supervised ML algorithms for training 3. Display training results and overall performance data in a visualization interface	2. Generate labels for unlabelled input data 3. Provide clinical insights into how labels are determined [1]

6.2 Objectives

To measure the success of the software toolbox, this section outlines the objectives that should be met when testing the final solution. Table 2 lists the objectives that will be tested, how they will be measured, and the numerical goal that should be reached.

Table 2. Objectives of the Software Toolbox

Objective	**Metric**	**Goal**
Accuracy	Percentage of data labelled correctly	Should be greater than 80% [8], as requested by supervisor
Labeling speed	Amount of data that can be labelled per unit time	Should be faster than the sampling speed of 256 Hz [9], as requested by supervisor

6.3 Constraints

Constraints are the limits the software toolbox is restricted to to be a valid design. The following table lists the constraints imposed by the project supervisor, as well as legal requirements.

Table 3. Constraints and Metrics of the DSP Software Toolbox

Constraint	**Metric**	**Reason for Constraint**
Must have a visual interface	Not Measurable	Stated by supervisor
Must accept .nwb files as input	Whether the toolbox can process .nwb files	
Must follow the Personal Information Protection and Electronic Documents Act	Whether the toolbox follows the principles laid out by the code, including the collection, use, and retention of personal information from patients	An organization is legally obligated to collect and use personal information in fair and legal ways, and only with consent [10]

7.0 Conclusion

This document highlights the problem clinicians face while manually labelling data to separate epileptic seizures from other brain activity. This is the first milestone for the project, and it identifies the scope and

the main requirements for the later stages. The project will focus on developing a machine learning algorithm that would automate the process and save clinicians' time. The main functions of the algorithm would be to perform feature extraction on input data and generate labels for it. The model will be checked against various objectives including accuracy and labelling speed. The document also lists some restrictions on the potential solutions, such as requiring a visual interface and accepting .nwb files as input.

8.0 References

[1] (2022, July 15). *Re: Interest in Project for ECE496*.

[2] Dai C;Wu J;Pi D;Becker SI;Cui L;Zhang Q;Johnson B; "Brain Eeg Time-series clustering using maximum-weight clique," *IEEE transactions on cybernetics*. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/32149677/. [Accessed: 24-Oct-2022].

[3] S. Scholl, "Fourier, gabor, Morlet or wigner: Comparison of time-frequency transforms," *arXiv.org*, 17-Jan-2021. [Online]. Available: https://arxiv.org/abs/2101.06707. [Accessed: 24-Oct-2022].

[4] "Assignment," Assignment - Visual Analytics and Applications. [Online]. Available: https://wiki.smu.edu.sg/18191isss608g1/Assignment. [Accessed: 24-Oct-2022].

[5] "Stacked autoencoders based deep learning approach for automatic epileptic seizure detection," *IEEE Xplore*. [Online]. Available: https://ieeexplore.ieee.org/document/8703357. [Accessed: 24-Oct-2022].

[6] "Department of Electrical & Computer Engineering emails: Arxiv:1908 ..." [Online]. Available: https://arxiv.org/pdf/1908.10432v1.pdf. [Accessed: 24-Oct-2022].

[7] "ArXiv:2108.01030v1 [eess.SP] 2 Aug 2021." [Online]. Available: https://arxiv.org/pdf/2108.01030.pdf. [Accessed: 24-Oct-2022].

[8] Jin Jing, P. D. (2020, January 1). Interrater reliability of experts in identifying interictal epileptiform discharges in eegs. JAMA Neurology. Retrieved September 23, 2022, from https://jamanetwork.com/journals/jamaneurology/fullarticle/2752670

[9] CHB-MIT Scalp Eeg Database. CHB-MIT Scalp EEG Database v1.0.0. (2010, June 9). Retrieved September 23, 2022, from https://physionet.org/content/chbmit/1.0.0/

[10] Branch, L. S. (2022, September 15). Consolidated federal laws of canada, Personal Information Protection and Electronic Documents act. Personal Information Protection and Electronic Documents Act. Retrieved September 23, 2022, from https://laws-lois.justice.gc.ca/eng/acts/p-8.6/page-7.html