Graph-based EEG Analysis

EE-452: Network Machine Learning

EPFL - Spring 2024/2025

1 Introduction

Epilepsy affects nearly 50 million people worldwide, according to a 2019 World Health Organization report [4]. It is a cerebral disease consisting of recurrent *seizures*, i.e. transient occurrence of signs and/or symptoms due to abnormal excessive or synchronous neuronal activity in the brain [1]. It is estimated that 70% of the people with epilepsy could live "seizure-free" if properly diagnosed and treated, either through medicine assumption or brain surgery. Being able to reliably detect, and if possible predict, their attacks could improve the life standard of the untreatable subjects.

The main non-intrusive tool used by doctors to monitor the neuronal activity of people with epilepsy is Electroencephalography [2], which captures neuronal activity with electrodes placed over the scalp. More precisely, voltage fluctuations resulting from ionic current within neurons are measured, producing multiple time series, or *channels*, which together form an *electroencephalogram* (EEG).

These signals pose significant modeling challenges due to their high dimensionality, noise, and non-stationary nature, necessitating careful preprocessing and feature extraction.

Traditional methods for EEG analysis rely on signal processing techniques, such as Fourier transforms and wavelet decompositions, or deep learning models designed for time-series data, such as recurrent neural networks (RNNs) and transformers. However, these approaches often overlook the inherent spatial relationships between EEG electrodes, which can be critical for understanding brain connectivity patterns.

A promising alternative is to represent EEG data as a graph, where nodes correspond to electrodes and edges encode functional or structural relationships between brain regions. This graph-based representation enables the application of Graph Neural Networks (GNNs) and other geometric deep learning techniques to capture complex dependencies in EEG signals. By leveraging graph structures, these methods can enhance interpretability and improve predictive performance in tasks such as seizure detection, mental state classification, and cognitive workload assessment.

Main Objective: This project aims to explore the utility of graph-based methods for EEG time-series data processing. Students should compare graph-based and non-graph-based approaches, evaluating their advantages and limitations. Through this comparison, students will gain insights into how graph structures influence model performance and interpretability. These insights should be clearly articulated in the project report and presentation.

2 Dataset & Preprocessing

In this project, we use a subset of the Temple University Hospital EEG Seizure Corpus (TUSZ) [3], which is one of the largest public EEG seizure databases. We will study data from 50 patients for training and 25 patients for testing. Each patient has multiple EEG *sessions*, which represent continuous recordings of electrical potentials, in mV, measured by 19 electrodes at 250 Hz according to the 10-20 system, represented in Figure 1. We will analyze non-overlapping windows of 12 s, with corresponding labels for the training set, indicating whether a window contains regular brain activity or seizures.

For this task, graphs might represent structural or functional properties of the brain. Figure 1 represents a distance-based graph, whose pairwise distances are available on the `distances_3d.csv` file. The students are encouraged to try different graphs for this task, getting inspired by the geometry and the physics of the problem.

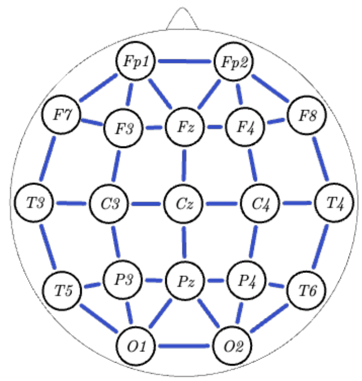

Figure 1: Standard 10-20 montage for EEG acquisition. Blue edges represent the graph based on distances between electrodes on the scalp.

Usage Conditions: Importantly, by participating in this project, the students agree to abide by the original dataset conditions, which are fully described in the official dataset agreement. In particular, the names of the participating students will be communicated to the data owner and the students agree with the following conditions:

> (1) "The user should acknowledge the provider of this data using the publication listed in the documentation included with the specific corpus." (i.e., [3] should be mentioned in the provided report.)
>
> (2) "The user will not release data to a third party or redistribute the data."
>
> (3) "The user agrees that no attempts will be made to re-identify the subjects, who have been anonymized in this distribution."
>
> (4) "The user will not use the data for malicious purposes. This data can be used for research and technology development, but not for uses beyond these broad classifications."
>
> (5) "The data recipient will delete the data from all computer systems when finished with the data."

3 Kaggle Competition

To enhance engagement and encourage experimentation, we will host a Kaggle Competition where students can test their different baselines in a competitive setting. This competition serves as an opportunity for students to benchmark their models and refine their approaches. Additionally, to ensure a meaningful challenge, we will award a bonus to the top-performing groups. This incentive encourages innovation while fostering an engaging learning experience. See Section 6 for more details.

While the Kaggle leaderboard provides a way to gauge performance relative to other teams, it should primarily serve as a reference for comparison rather than the sole evaluation metric. We strongly encourage groups to conduct and report local cross-validation performances using a validation sets within the provided training data. This ensures that model selection is not solely dependent on Kaggle rankings but is guided by robust internal evaluation practices.

Finally, your models should be **exclusively** trained on the data provided on Kaggle.

4 Computational Resources

All enrolled students will have access to GPU resources on the SCITAS cluster, Izar. If you are not familiar with SCITAS, please review the documentation available at SCITAS Documentation. The dataset (approximately 6.25 GB) will be available on the cluster. Alternatively, you may choose to use your own computational resources or Google Colab.

5 Project Expectations & Deliverables

Students are expected to submit **three deliverables: a report, a presentation, and the corresponding code**.

5.1 Code

We recommend using libraries already utilized in the exercise session notebooks (such as *PyTorch*, *PyTorch Geometric*, *NetworkX*, *DGL*, ...), but students are free to use any existing open-source code to build their models. The code should be well-documented and reproducible.

5.2 Report

Each group must submit a strictly 4-page report, excluding references, following a structured format similar to a scientific paper. We recommend using the IEEE Conference Template for consistency and clarity. The report should include the following sections:

- **Introduction**: Problem motivation and objectives.
- **Method**: Approach and technical details.
- **Results**: Experiments and findings.
- **Conclusion**: Summary and insights.

5.3 Presentation

Each group will give a presentation of up to 10 minutes, summarizing their work. The structure of the presentation should closely follow the report, covering the problem statement, methodology, key findings, and conclusions.

5.4 Submission Deadlines

- **Report and Code Submission**: To be submitted on Moodle by 30th May 2025.
- **Presentation Dates**: Between 2nd June and 13th June 2025.

6 Evaluation Criteria

The project grade will be based on three components: presentation, code, and report. All team members will receive the same grade.

6.1 Grading Breakdown

- **80%: Presentation and Report**
  - Description of methods (20%)
  - Originality of the approach (10%)
  - Interpretation of results (25%)
  - Presentation: communication and clarity (15%)
  - Report: writing quality, clarity, and overall presentation (10%)
- **20%: Code**
  - Functionality: the code should run without issues (10%).
  - Code quality and documentation: readability, structure, and comments (10%).

6.2 Bonus Points

We will award bonus points to the top 10 performing teams. The bonus will be distributed in a linearly spaced manner, with the 1st team receiving the maximum bonus of 0.5 and the 10th team receiving the minimum of 0.05, with intermediate teams awarded proportionally decreasing bonuses. These bonus points will be added to the final project grade (capped at 6).

References

[1] Fisher, R.S., Boas, W.v.E., Blume, W., Elger, C., Genton, P., Lee, P., Engel Jr., J.: Epileptic Seizures and Epilepsy: Definitions Proposed by the International League Against Epilepsy (ILAE) and the International Bureau for Epilepsy (IBE). Epilepsia **46**(4), 470–472 (2005)

[2] Schomer, D.L., da Silva, F.L.: Niedermeyer's Electroencephalography: Basic Principles, Clinical Applications, and Related Fields. Lippincott Williams & Wilkins (2012)

[3] Shah, V., von Weltin, E., Lopez, S., McHugh, J.R., Veloso, L., Golmohammadi, M., Obeid, I., Picone, J.: The Temple University Hospital Seizure Detection Corpus. Frontiers in Neuroinformatics **12** (Nov 2018). https://doi.org/10.3389/fninf.2018.00083

[4] World Health Organization: Epilepsy: A public health imperative. Tech. Rep. WHO/MSD/MER/19.2, World Health Organization (2019)