EEG Preprocessing Pipeline for Epilepsy Detection

Graduate Thesis

October 1, 2025

1 EEG Preprocessing Pipeline

1.1 Overview

This section describes the EEG preprocessing pipeline developed for epilepsy detection using the Temple University Hospital (TUH) EEG Epilepsy Corpus dataset. The raw EEG recordings in EDF format undergo a series of standardized signal processing steps to produce artifact-cleaned, normalized epochs suitable for downstream machine learning and graph neural network analysis.

1.2 Pipeline Description

The core preprocessing is implemented in `preprocess_core.py`, encapsulating the following steps:

- **Channel Cleaning and Montage Assignment:** Raw channel names are standardized by removing prefixes and suffixes, and a fixed set of 22 core 10-20 layout electrodes is selected. A manual addition of nonstandard electrodes T1 and T2 is performed to enhance spatial coverage.
- **Referencing and Filtering:** Signals are common average referenced. A notch filter at 60 Hz removes power line interference, followed by bandpass filtering between 0.5 and 100 Hz to capture relevant EEG frequencies while excluding noise.
- **Independent Component Analysis (ICA, Optional):** ICA is optionally applied to remove artifacts such as eye blinks and muscle noise. This is currently disabled but easily enabled for customizable cleaning.
- **Resampling and Cropping:** Signals are resampled to 250 Hz to standardize temporal resolution. The first 10 seconds of non-epileptic controls are cropped to remove potential recording startup artifacts.
- **Epoching:** The continuous EEG is segmented into fixed-length 2-second epochs without overlap, facilitating time-locked analysis.
- **Artifact Rejection:** Epochs with peak-to-peak amplitude exceeding the 95th percentile threshold are rejected to reduce contamination by noise and artifacts.

- **Z-score Normalization:** Each channel within each epoch is normalized via z-scoring (mean subtraction and scaling by standard deviation) to stabilize feature variance across epochs and patients.
- **Labeling:** Each epoch is labeled as epileptic or non-epileptic based on file path information, enabling supervised learning tasks.
- **Power Spectral Density (PSD) Analysis:** PSD estimates before and after preprocessing are computed and saved for quality assurance and spectral characterization.

1.3 Script Usage

`preprocess_single.py` wraps the core preprocessing function and facilitates single EEG file processing via command line:

```
python src/preprocess_single.py --edf path/to/file.edf --out path/to/output_dir --psd_
```

While `preprocess_batch.py` enables batch processing of entire datasets, recursively scanning for EDF files, preserving input directory hierarchy in outputs, and allowing limit on the number of patients processed for testing:

```
python src/preprocess_batch.py --input_dir data_raw --output_dir data_pp --psd_dir fig
```

This command will preprocess EEG recordings for at most 10 unique patients, saving cleaned epochs, labels, raw signals, metadata, and PSD plots with folder structure preserved.

1.4 Rationale for Pipeline Steps

Each step is chosen to ensure clean, standardized, and comparable EEG input for advanced analyses:

- Cleaning and selecting core channels ensures spatial consistency.
- Referential and frequency filtering remove common noise sources without discarding physiological signals.
- Resampling harmonizes sampling frequencies across variable recordings.
- Epoching and artifact rejection reduce temporal noise and improve model robustness.
- Normalization stabilizes feature scales to facilitate machine learning convergence.
- Label extraction automates ground truth generation for classification tasks.
- PSD visualization supports spectral quality control and data exploration.

By following this reproducible pipeline, the processed EEG data becomes suitable for robust, interpretable epilepsy prediction modeling.