Machine learning approach for marking seizures on epileptic EEG

Authors:

Vadim Grubov[1], Sergey Afinogenov[2], Vladimir Maksimenko[1], Nikita Utyashev[3]

[1]**Immanuel Kant Baltic Federal University**, Baltic Center for Artificial Intelligence and Neurotechnology

[2]**Financial University under the Government of the Russian Federation**, Faculty of Information Technology and Big Data Analysis

[3]**Ministry of Healthcare of the Russian Federation**, National Medical and Surgical Center named after N. I. Pirogov

Participants and Data acquisition

Dataset by National Medical and Surgical Center named after N. I. Pirogov of Russian Healthcare Ministry

- Subjects: 30 adults, 15 males and 15 females, age 33.4±9.4
- Diagnosis: focal epilepsy
- Length: 8-57 hours depending on patient's condition
- Seizures: 1-5 for each subject

EEG recorded with "Micromed" encephalograph (Micromed S.p.A., Italy)

- 25 EEG channels according to "10-20" system
- Sampling rate: 128 Hz
- Filters: band-pass (1 and 60 Hz), notch (50 Hz)
- Artifact removal: Independent Component Analysis (ICA)

Data preprocessing

Time-frequency analysis of EEG signals using continuous wavelet transform (CWT) with Morlet mother wavelet function.

We consider wavelet power (WP) as:

$$W_n(f,t) = |w_n(f,t)|, \quad (1)$$

We calculated averaged WP (AWP) by averaging WP values over N = 25 EEG channels:

$$E(t) = \frac{1}{N}\sum_{n=1}^{N} W_n(f,t) \quad (2)$$

We divided each EEG recording into 60-second intervals T_m, where m = 1,2… M, M = L//60, L – the length of EEG recording in seconds.

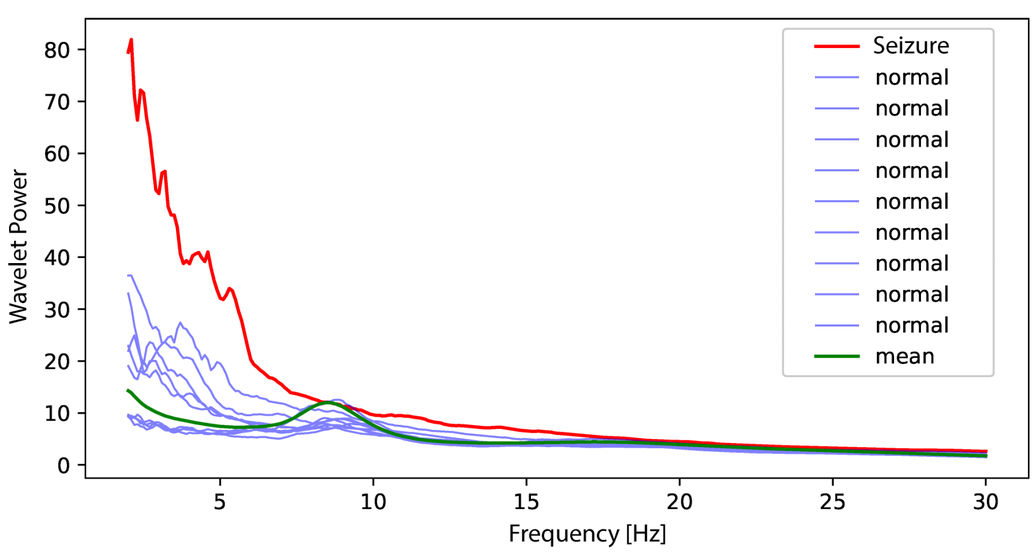
AWP values were calculated for each time interval T_m and averaged over the whole length of the interval to obtain "downsampled" AWP (DAWP):

$$e_m = \frac{1}{\Delta T}\int_{t\in T_m} E(t)dt, \quad (3)$$

Feature engineering

The initial feature space consisted of DAWP spectra, but we aimed to introduce several additional features.

Extended research on epileptic EEG reveals certain peculiarities of seizures in comparison to normal EEG.



Typical DAWP spectra of a single patient for normal (blue), epileptic (red) and mean (green) activity

Properties of EEG spectrum differ between epileptic and normal activity.

According to these properties we introduced two features:

- *Mean* – mean DAWP across 2-30 Hz frequency range
- *Variance* – variance of DAWP in spectrum

Additional features to assess normal and epileptic data similarity were introduced using cosine similarity:

- *SimToMean* – cosine similarity between DAWP spectrum at given time interval T_m and mean DAWP spectrum for the patient
- *SimToNeigh* – mean cosine similarity between DAWP spectrum at given time interval T_m and each of DAWP spectra from neighboring intervals

Some parts of the spectrum are more prone to reflect epileptic activity, so we introduced:

- *FreqDiff* – difference between DAWPs averaged over low (2-5 Hz) and high (5-30 Hz) frequencies

We used Principal Component Analysis (PCA) to reduce initial DAWP spectrum down to two components – *PCA0* and *PCA1*, that contain 97.18% of all information from the initial data.

Correlation analysis showed high correlation between *Mean* and *PCA0*, so we excluded *Mean*.

Final set consisted of 6 features: *PCA0, PCA1, Variance, SimToMean, SimToNeigh, FreqDiff*

Results

We used **RandomForest** algorithm to develop two-class classifier with 4 possible outcomes:

- **True Positive (TP)** – correctly identified seizure;
- **True Negative (TN)** – correctly identified normal activity;
- **False Positive (FP)** – incorrectly identified epileptic seizure, i.e. episode of normal activity identified as seizure;
- **False Negative (FN)** – missed epileptic seizure, i.e. seizure identified as episode of normal activity.

We assessed *Recall* and *Precision* as:

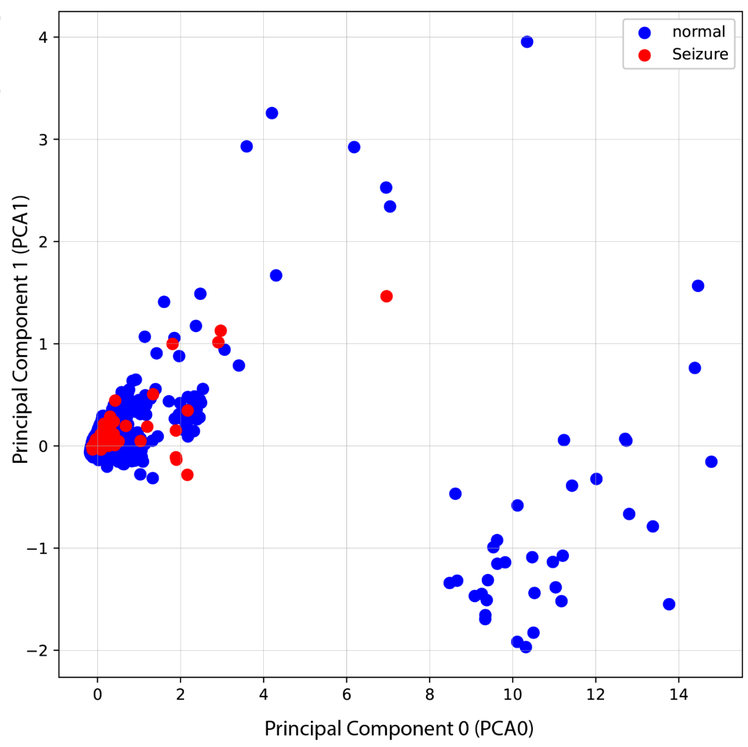
$$Recall = \frac{TP}{TP+FN} \quad (4)$$

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

On the studied dataset the classifier provides *Recall* = 78.67±1.33% and *Precision* = 5.33±0.22%. These results are comparable to our previous work on similar dataset.

We also analyzed of feature significance and ranked the features.

Feature	%
Variance	31.68
SimToNeigh	26.93
FreqDiff	24.65
PCA0	7.97
PCA1	7.04
SimToMean	1.73



The DAWP projection on the two principal components (PCA0 and PCA1)

Three most significant features – *Variance, SimToNeigh and FreqDiff* – together contribute 83.26% to classification.

At the same time, features *PCA0* and *PCA1*, that contain 97.18% of all information from the "raw" data, contribute only ~15%.

This is an important result: most significant features are based on the knowledge of EEG data and peculiarities of seizure activity, while the features derived mathematically have low significance for classification.