1ST WORKSHOP ON ARTIFICIAL INTELLIGENCE FOR BIOMEDICAL DATA - AIBIO

Self-Attention as a Predictor of EEG Anomalies

Natalia Koliou (presenter),
Maria Sierra, Christoforos Romesis,
Stasinos Konstantopoulos, and Luis Montesano

MANOLO
CLOUD-EDGE EFFICIENT & TRUSTWORTHY AI

Funded by
the European Union

25 Oct 2025

Overview

1. Introduction & Background
2. Research Methodology
3. Experiments
4. Results & Discussion
5. Conclusion & Future Work

Introduction & Background

Introduction & Background

- EEG signals are highly sensitive, low-amplitude recordings easily contaminated by **artifacts** (noise not generated by brain activity).
- Artifacts may arise from physiological (muscle tension, sweating) or technical sources (electrode detachment).
- What counts as "noise" is often **task-dependent**: a feature irrelevant for one application may be meaningful for another.
- Traditional denoising via **reconstruction/prediction errors** (Autoencoders, LSTMs) ignores contextual relevance.
- **Our idea:** Train an attention-based model on a downstream task (e.g., sleep stage classification) and use its attention patterns to infer anomalies without explicit artifact labels.

Research Methodology

Self-Attention as a Predictor of EEG Anomalies

Research Methodology

- Counterfactual explanations provide **instance-based insights** for time-series classifiers.
- The goal is to explain why a classifier predicts a certain label:
  - Dataset: $D = \{x^{(i)}\}_{i=1}^{N}$, each $x^{(i)} \in \mathbb{R}^d$.
  - Classifier: $C : \mathcal{X} \to \mathcal{Y}$, labels $\mathcal{Y} = \{1, \ldots, n\}$.
  - For input x, predicted label $y = C(x)$.
- For misclassified instances x_m with $C(x_m) \neq y^*(x_m)$, find **minimal modification** x'_m such that $C(x'_m) = y^*(x_m)$ and x'_m is close to x_m.
- $\Delta x_m = x'_m - x_m$ reveals dominant patterns influencing the classifier's decision.

Experiments

Data Preprocessing

We use the **BOAS EEG dataset**, with 128 full-night recordings from 2 channels (256 Hz), to explain classifier errors across 4 sleep stages (N1, N2, N3, REM).

- **EEG Data:**
  - 30s epochs ($n_s = 7680$ samples, $n_c = 2$).
  - FFT $\rightarrow$ 0.4-30 Hz filtering $\rightarrow$ segment into $n'_s = 300$ spectral slices.
  - Extract 3 features per slice (frequency, phase, amplitude).
- **Classifier:** Two-stage convolutional network over sequences of $k = 5$ epochs:

$$X_t^f = [e_{t-k+1}, \ldots, e_t] \rightarrow \hat{y}_t$$

- **Autoencoders:** Hybrid networks with convolutional and attention layers:

$$X_f^t = e_t \oplus \text{pos} \longrightarrow \hat{X}_f^t \text{ (reconstructed input)}$$

Training Process & Hyperparameters

- We pass the test dataset through each class-specific autoencoder, then classify the restyled outputs using the original classifier.
- High **classification accuracy** on restyled signals indicates the autoencoders have learned the patterns that the classifier uses for each sleep stage.

Metric	N1	N2	N3	REM
Accuracy	0.9997	0.9998	0.9260	0.9986

Results & Discussion

Results & Discussion

We use the **BOAS EEG dataset**, with 128 full-night recordings from 2 channels (256 Hz), to explain classifier errors across 4 sleep stages (N1, N2, N3, REM).

- **EEG Data:**
  - 30s epochs ($n_s = 7680$ samples, $n_c = 2$).
  - FFT $\rightarrow$ 0.4-30 Hz filtering $\rightarrow$ segment into $n'_s = 300$ spectral slices.
  - Extract 3 features per slice (frequency, phase, amplitude).
- **Classifier:** Two-stage convolutional network over sequences of $k = 5$ epochs:

$$X_t^f = [e_{t-k+1}, \ldots, e_t] \rightarrow \hat{y}_t$$

- **Autoencoders:** Hybrid networks with convolutional and attention layers:

$$X_f^t = e_t \oplus \text{pos} \longrightarrow \hat{X}_f^t \text{ (reconstructed input)}$$

Conclusion & Future Work

Conclusion

- **xSTAE** explains classifier errors by restyling misclassified EEG instances into correctly classified ones.
- Autoencoders balance **identity loss** (keep instance similar) and **classification loss** (push to correct label), revealing features the classifier missed.
- Contributions:
  - Ground counterfactual explanations in time-series EEG classification.
  - Identify spectral representations suitable for EEG signals.
  - Validate on open BOAS dataset and release full experimental setup.

Future Work

- Explore **alternative identity losses** to highlight meaningful changes (e.g., bigger local changes, selective brainwave bands).
- Conduct **expert trials** to refine interpretability of restyled EEGs.
- Investigate linking insights from misclassifications to **actionable guidance** at the data or confidence level, keeping xSTAE model-agnostic.

Thank you for your attention!

Thank you for your attention! Any questions?

Natalia Kolliou, Christoforos Romesis, Stasinos Konstantopoulos

Maria Sierra, Luis Montesano

This research was co-funded by the European Union under GA no. 101135782 (MANOLO project). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or CNECT. Neither the European Union nor CNECT can be held responsible for them.

Funded by the European Union

MANOLO

CLOUD-EDGE EFFICIENT & TRUSTWORTHY AI

`https://manolo-project.eu`