

# 1ST WORKSHOP ON ARTIFICIAL INTELLIGENCE FOR BIOMEDICAL DATA - AIBIO



## xSTAE:

Explaining Classifier Decisions through EEG Signal Style Transfer Autoencoding

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



Funded by  
the European Union

25 Oct 2025



# Overview

## 1. Introduction & Related Work

## 2. Problem Statement

## 3. Proposed Methodology

## 4. Experiments

## 5. Conclusion & Future Work

# Introduction & Related Work

---

*xSTAE: Explaining Classifier Decisions through EEG Signal Style Transfer Autoencoding*

# Introduction & Related Work

- Understanding EEG-based sleep stage classification is crucial for sleep research and healthcare.
- Traditional models struggle with **misclassifications**, especially for minority sleep stages.
- Existing work interprets time-series classifiers, either by integrating explainers into the model (e.g., XTF-CNN, DeepVix) or using post-hoc techniques to highlight which input patterns drive predictions (e.g., timeXplain, LIME).
- **Counterfactual explanations** have been successful in visual domains but are underexplored for time-series EEG.
- We address this gap by generating **counterfactual EEG examples** that show what a misclassified instance should have looked like to be correctly classified.

# Problem Statement

---

*xSTAE: Explaining Classifier Decisions through EEG Signal Style Transfer Autoencoding*

# Problem Statement

- 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:  $C : \mathcal{X} \rightarrow \mathcal{Y}$ , labels  $\mathcal{Y} = \{1, \dots, n\}$ .
  - 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.

# Proposed Methodology

---

*xSTAE: Explaining Classifier Decisions through EEG Signal Style Transfer Autoencoding*

# Proposed Methodology

- **xSTAE**: A generative framework using class-conditional **autoencoders**.
- Each autoencoder  $E_{\text{tgt}} : \mathcal{X} \rightarrow \mathcal{X}_{\text{tgt}}$  is trained to reconstruct any input while restyling it toward a target class  $y_{\text{tgt}} \in \mathcal{Y}$ .
- For input  $x$  with  $y_{\text{tgt}} \neq C(x)$ , the autoencoder generates a counterfactual:

$$x' = E_{\text{tgt}}(x) \quad \text{s.t.} \quad C(x') = y_{\text{tgt}}$$

- Comparing  $x$  and  $x'$  reveals the patterns in  $x$  responsible for the classifier's original decision.
- Training uses dual loss function:
  1. **Identity**: Ensures  $x'$  remains similar to  $x$ , using a distance function  $d(x, x')$ .
  2. **Classification**: Ensures  $x'$  is classified as the target, by comparing  $C(x')$  with  $y_{\text{tgt}}$ .



# Experiments

---

*xSTAE: Explaining Classifier Decisions through EEG Signal Style Transfer Autoencoding*

# Setup



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:

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

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

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

# Quantitative Results



- 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

# Qualitative Discussion

- **Example case:** Original EEG segment was labeled as N2 but misclassified as N1.
- xSTAE restyles the segment toward the correct class (N2), producing a counterfactual.
- Comparison reveals why the classifier erred:
  - Original signal lacked sufficiently prominent spikes.
  - Restyled signal emphasizes features around timesteps 55-60, 70-75, 90-95.
  - Classifier "expects" these more pronounced spikes to identify N2.
- Misclassification occurs not from wrong feature detection, but from **insufficient weighting** of existing patterns.

# Conclusion & Future Work

---

*xSTAE: Explaining Classifier Decisions through EEG Signal Style Transfer Autoencoding*

# 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.

# 1ST WORKSHOP ON ARTIFICIAL INTELLIGENCE FOR BIOMEDICAL DATA - AIBIO



## Thank you for your attention!

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



Funded by  
the European Union

25 Oct 2025