Mini-Project (ML for Time Series) - MVA 2025/2026

Yasmin van den Broek yasmin.van-den-broek@ensta.fr
Tiffney Aina tiffneyaina@gmail.com

January 6, 2026

1 Introduction and contributions

Electroencephalography (EEG) has become increasingly important in brain-computer interface (BCI) technology and is currently used for numerous applications, including stroke rehabilitation, sleep stage detection, and emotion regulation in mental health treatments. EEG signals encode information across both spatial dimensions (electrode positions) and temporal dimensions (neural oscillations), making it important that BCI systems effectively capture subtle temporal dynamics. However, this task is challenging due to high inter-subject variability and susceptibility to artifacts from movement, eye blinks, and environmental factors.

The EEG-Deformer architecture [8] addresses these challenges through a hierarchical approach designed to capture both coarse and fine-grained temporal patterns in EEG signals. The architecture, consisting of three main components, is composed by a shallow feature encoder for initial temporal processing, a Hierarchical Coarse-to-Fine Transformer (HCT) that integrates Fine-grained Temporal Learning (FTL) with global attention mechanisms, and a Dense Information Purification (DIP) module that leverages multi-level temporal information to enhance decoding accuracy. Previous approaches based on CNNs or standard Transformers captured either local or global temporal dependencies, but not both simultaneously within each layer, which limited their ability to fully exploit EEG signals' multi-scale temporal dynamics.

While EEG-Deformer demonstrated strong performance on motor imagery tasks using research-grade equipment, its effectiveness under realistic deployment constraints remained unexplored. We address this gap through two complementary tracks. First, we evaluate EEG-Deformer on DREAMER [4], which employs consumer-grade EEG (Emotiv EPOC, 14 channels) for emotion recognition. We investigate whether architectural sophistication provides tangible benefits when signal quality and channel density are limited, conducting systematic ablation studies on model depth and temporal structure sensitivity. Second, we address the authors' acknowledged robustness limitation through controlled degradation experiments on Fatigue, testing both temporal and spatial resilience. We perform Gaussian noise injection to evaluate spectral corruption tolerance, and introduce channel masking to test the model's capacity to leverage multivariate spatial covariance when specific channels are lost—simulating realistic BCI "lead-off" artifacts. This reveals a critical asymmetry: while temporal robustness remains strong (95% baseline at 20 dB SNR), spatial fragility is severe, with accuracy collapsing from 76.6% to 50.9% at just 10% channel masking.

Contributions and methodology. Responsibilities were equally distributed: Yasmin focused on DREAMER implementation with architectural ablations (depth, temporal structure) and baseline comparisons. Tiffney investigated data diagnostics SVD and outlier detection, as well as, robustness through controlled degradation on Fatigue, including SNR manipulation and channel masking to evaluate temporal versus spatial resilience. We utilized the Deformer class from the original implementation [2], adapting it for DREAMER. Beyond this, we developed new code for data loading, preprocessing, ablation frameworks, temporal perturbations, baseline models, and degradation pipelines. Our experiments extend the original paper by: (1) evaluating on consumer-grade EEG not tested by authors; (2) conducting systematic depth ablations; (3) analyzing temporal structure sensitivity; (4) testing acknowledged robustness limitations through SNR degradation and spatial masking, revealing spatial fragility unexplored in the original work.

2 Method

2.1 DREAMER Implementation

For the DREAMER dataset [4], emotion recognition is formulated as binary classification. Given EEG segments $X \in \mathbb{R}^{C \times T}$ where $C = 14$ channels and $T = 384$ samples (3 seconds at 128 Hz), we learn a mapping $f : \mathbb{R}^{C \times T} \to \{0, 1\}$ predicting binary emotion labels. Continuous ratings $r \in [1, 5]$ are binarized as:

$$y = \begin{cases} 1, & \text{if } r > \tau \quad \text{(high emotional intensity)} \\ 0, & \text{if } r \leq \tau \quad \text{(low emotional intensity)} \end{cases} \tag{1}$$

where $\tau = 3$ balances class distribution. Each segment undergoes z-score normalization per channel, with statistics computed from training data to prevent leakage.

We minimize cross-entropy lossoptimized via AdamW with cosine annealing, following the original EEG-Deformer configuration. Training uses 80%/20% splits with early stopping.

Our experiments include: (1) baseline performance across valence, arousal, and dominance; (2) comparison with simpler architectures—a three-block CNN (kernel sizes [7,5,3], filters [32,64,128]) and bidirectional LSTM (hidden size 128); (3) depth ablation using single-layer models; and (4) temporal structure sensitivity via three perturbations: reversed time, block-shuffled (swapped halves), and randomly permuted time points.

2.2 Robustness Analysis: SNR Degradation

To directly address the authors' acknowledged limitation regarding robustness, we systematically degrade the Fatigue dataset by injecting additive Gaussian noise at controlled SNR levels.

We implemented a controlled noise injection pipeline utilizing zero-mean Additive White Gaussian Noise (AWGN). For each EEG trial $X \in \mathbb{R}^{C \times T}$, the process follows three steps:

1. Signal Power Calculation. We compute the average power of the clean trial:

$$P_{\text{sig}} = \frac{1}{C \cdot T} \sum_{c=1}^{C} \sum_{t=1}^{T} X_{c,t}^2 \tag{2}$$

2. Target Noise Power. For a specified target SNR_{dB}, the required noise power is derived by rearranging the SNR definition:

$$P_{\text{noise}} = \frac{P_{\text{sig}}}{10^{(\text{SNR}_{\text{dB}}/10)}} \tag{3}$$

3. Gaussian Injection. We generate a noise matrix $N \sim \mathcal{N}(0, \sigma^2)$ where $\sigma = \sqrt{P_{\text{noise}}}$, yielding the corrupted trial:

$$\tilde{X} = X + N \tag{4}$$

2.3 Spatial Robustness: Random Channel Masking

To evaluate structural resilience, we simulate sensor failure via a stochastic masking operator. For an input $X \in \mathbb{R}^{C \times T}$, we define a masking ratio $\rho \in \{0.1, 0.2, 0.5\}$ and sample a subset of indices $\mathcal{M} \subset \{1, \ldots, C\}$ where $|\mathcal{M}| = \lfloor \rho C \rfloor$. The perturbed input $\tilde{X}$ is defined as:

$$\tilde{X}_{c,t} = \mathbb{1}_{\{c \notin \mathcal{M}\}} \cdot X_{c,t} \quad \forall\, t \in [1, T] \tag{5}$$

where $\mathbb{1}$ is the indicator function. By zeroing entire time series without re-training, we test the model's ability to leverage learned inter-channel correlations—a vital property for multivariate time-series robustness. Indices are sampled uniformly without replacement per subject, ensuring structural stress across the entire cohort.

3 Data

We evaluate the EEG-Deformer on two datasets representing opposite ends of the EEG quality spectrum: consumer-grade DREAMER and research-grade Fatigue.

DREAMER Dataset Working with the DREAMER [4] dataset introduces practical challenges not found in the original benchmarks. Captured via 14-channel consumer-grade Emotiv EPOC (128 Hz) with saline-soaked felt electrodes, the signal suffers from lower signal quality, power-line interference, and contact instability compared to research-grade gel-based systems. Label quality introduces further complexity: self-reported valence/arousal scores on a 1–5 scale result in subjective label noise. We binarized these scores ($\tau = 3$) to create a balanced split, but ambiguous boundary cases (scores near 3) and limited sample size (18 trials per subject) create risks of pseudo-replication and overfitting.

Fatigue Dataset The Fatigue dataset [1] provides a research-grade contrast, consisting of 32-channel EEG recorded at 500 Hz and clinically preprocessed with 1–50 Hz band-pass filtering and automatic artifact removal (AAR). The resulting signals are substantially cleaner than DREAMER, and labels derived from reaction times yield clearer class separation. Nevertheless, baseline accuracies still vary widely across subjects (0.62–0.84), indicating strong inter-individual differences that warrant deeper diagnosis beyond aggregate performance.

Our data analysis followed a retrosynthetic approach. When applying moderate (20 dB) and severe (5 dB) Gaussian noise, results were not monotonic: among four tested subjects, two showed a decrease in accuracy (e.g. sub40), while two improved relative to baseline (e.g. sub21). This ambiguity suggested that robustness could not be explained by noise sensitivity alone, prompting further inspection of the data structure using singular value decomposition (SVD) and outlier statistics.

To assess spatial redundancy, we computed the SVD of each subject's EEG trials. As illustrated in 1, we observe a clear rank gap across subjects. Robust individuals (e.g. sub21) exhibit steep singular-value spectra, with 90% of spatial variance captured by the first ∼15 components, indicating strong inter-channel redundancy. Interestingly, these subjects are precisely those whose accuracy increased at 20 dB SNR, consistent with mild noise acting as a regularizer that reinforces dominant spatial modes. In contrast, fragile subjects such as sub40 show flatter spectra, reflecting near full-rank data where each channel carries unique information, leaving little spatial redundancy to exploit.

Outlier analysis further clarifies this distinction. Despite its poor robustness, sub40 contains fewer temporal outliers (0.30%) than sub21 (0.91%), ruling out transient spikes or temporal noise as the failure mode. Instead, sub40's signal is temporally clean but spatially sparse: the model overfits to a small set of critical sensors. When these channels are masked, the learned spatial representation collapses, revealing that robustness is governed not by temporal cleanliness but by spatial redundancy.

4 Results

We evaluate EEG-Deformer through two key experiments: (1) architecture analysis on the consumer-grade DREAMER dataset, testing whether the hierarchical coarse-to-fine design generalizes beyond the original benchmark tasks, and (2) robustness analysis under controlled noise perturbations, directly addressing the authors' acknowledged limitation that "efforts should focus on improving the model's robustness."

4.1 Experiment 1: Architecture Analysis on DREAMER

EEG-Deformer achieves an accuracy of 75.56% (valence), 72.78% (arousal), and 73.59% (dominance) on DREAMER, presented on 1 of Appendix B, demonstrating that it adapts successfully to consumer-grade EEG. However, we can notice substantial overfitting (23–25% train-test gap), with training accuracy reaching 98% while test performance remains at 73–76%. When comparing architectures, the simple CNN baseline elaborated for comparison matches the transformer (74.82%, compared to 74.53% from EEG-Deformer) with significantly fewer parameters, while LSTM displays a lower accuracy (60.36%), suggesting emotion recognition relies on local spatial-temporal patterns rather than the long-range dependencies transformers are designed to capture.

Results from systematically varying transformer depth, displayed in 2 in Appendix B, reveal that a single-layer model achieves optimal test accuracy. Deeper models exhibit worse overfitting (gap increases from 15% to 24%) without improving generalization (all configurations converge to about 75%). This directly challenges the paper's core claim that "HCT blocks are specifically designed to capture both coarse- and fine-grained temporal dynamics", as, on consumer-grade EEG, this hierarchical structure provides no measurable benefit.

To investigate what temporal structure models actually exploit, we evaluated sensitivity to temporal perturbations: reversed time, block-shuffled (swapped halves) and random permutation (Table 3, Appendix B). The experiments show that the models tolerate reversed and block-shuffled data (±1.5%) but fail under random permutation (-17%). This behavior indicates reliance on local temporal coherence, as the architecture preserves adjacent time points, rather than global sequential order. This explains both the CNN's performance, which was comparable to EEG-Deformer's and the LSTM's failure, since sequential modeling provides no benefit when global order is uninformative.

4.2 Experiment 2: Robustness Analysis

Spectral resilience under Gaussian noise We evaluated EEG-Deformer's robustness to temporal and spectral corruption by retraining on the Fatigue dataset with additive Gaussian noise at 20 dB and 5 dB SNR (Fig. 2). Relative to the clean baseline of 0.7659 ± 0.09, performance remained stable, with a mean accuracy of 0.734 even at the highest noise level. Interestingly, subject-specific accuracy often converged back toward baseline at 5 dB, while intermediate noise (20 dB) produced more variable effects across individuals. This behavior suggests that strong noise acts as a form of regularization, encouraging the model to disregard stochastic fluctuations and focus on stable spectral patterns associated with fatigue. Overall, these results indicate that EEG-Deformer is well adapted to severe temporal noise, effectively preserving discriminative rhythmic structure despite heavy interference.

Spatial Fragility: Structural Stress Test We contrasted this temporal resilience with a zero-shot random channel masking protocol (10%, 20%, 50%) to directly probe the structural assumptions of the Local-Global Graph (LGG) module (Fig. 3). In contrast to the Gaussian noise experiments, even mild spatial corruption caused a sharp performance drop: mean accuracy fell to 0.509 at 10% masking and declined below the 0.50 chance level at 50% masking (0.4726). This abrupt collapse indicates that the learned spatial representation lacks redundancy—once a small fraction of channels is removed, the model is unable to compensate using the remaining sensors. The effect is highly subject-dependent, with large variance across individuals (±0.15), revealing a clear robustness gap: while some subjects retain distributed spatial signatures, others rely on a small set of critical electrodes. Overall, these results show that although EEG-Deformer is effective at extracting temporal structure in noisy conditions, its spatial adaptation is rigid, treating sensors as discrete graph nodes rather than exploiting the continuous, correlated topology of EEG measurements.

4.3 Model Adaptation and Hypothesis Validation

The three experiments reveal a consistent pattern in how EEG-Deformer interacts with real EEG time series. On DREAMER, architectural ablations show that deeper hierarchical modeling does not improve generalization: performance saturates at 73–76% regardless of depth, CNN baselines match the Transformer, and temporal perturbations demonstrate reliance on local coherence rather than multi-scale dynamics. This challenges the claim that hierarchical coarse-to-fine modeling is beneficial when data quality and channel density are limited.

In contrast, robustness tests on the Fatigue dataset show a clear asymmetry between temporal and spatial adaptation. Under additive Gaussian noise, accuracy remains within 95% of baseline even at 5 dB SNR, indicating strong spectral resilience and validating the temporal feature-extraction hypothesis. However, random channel masking causes an abrupt collapse from 0.766 to 0.509 accuracy at only 10% sensor loss, with large subject-specific variance, revealing a rigid spatial representation.

Overall, EEG-Deformer is well adapted to temporal structure in EEG but poorly adapted to spatial incompleteness. The results indicate that the model behaves as a powerful temporal filter while lacking the spatial redundancy required for robust multivariate decoding under realistic sensor failure.

References

[1] Zehong Cao, Chun-Hsiang Chuang, Jung-Kai King, and Chin-Teng Lin. Multi-channel EEG recordings during a sustained-attention driving task. *Scientific Data*, 6(1):1–8, 2019.

[2] Yi Ding, Yong Li, Hao Sun, Rui Liu, Chengxuan Tong, Chenyu Liu, Xinliang Zhou, and Cuntai Guan. Eeg-deformer: A dense convolutional transformer for brain-computer interfaces. *IEEE Journal of Biomedical and Health Informatics*, 29(3):1909–1918, 2025.

[3] Crystal A Gabert-Quillen, Emily E Bartolini, Benjamin T Abravanel, and Charles A Sanislow. Affective picture viewing: how emotional responses change over time. 52:S40–S40, 2015.

[4] Stamos Katsigiannis and Naeem Ramzan. Dreamer: A database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices. *IEEE Journal of Biomedical and Health Informatics*, 22(1):98–107, 2017.

[5] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. 2017.

[6] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. 2019.

[7] Yonghao Song and Qingqing Zheng. Eeg-conformer: Convolutional transformer for eeg decoding. https://github.com/eeyhsong/EEG-Conformer, 2023. Accessed: 2024-12-20.

[8] Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. Eeg conformer: Convolutional transformer for eeg decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2023.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.

A Data Diagnosis

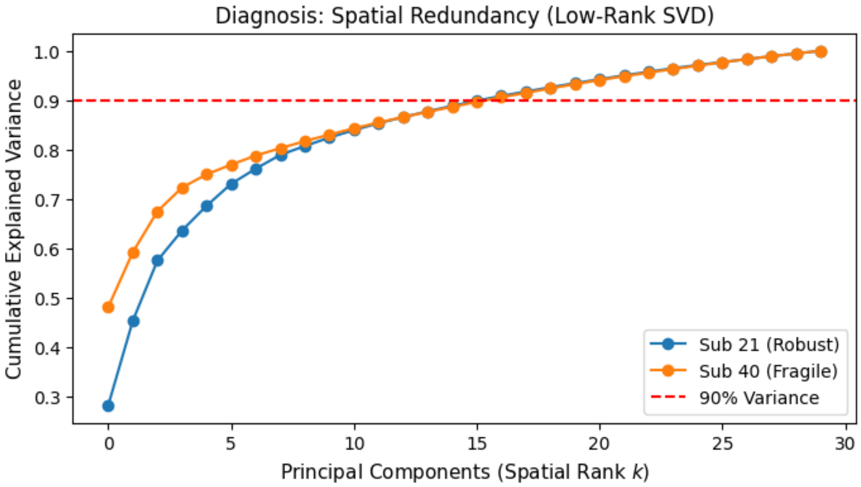

Figure 1: Cumulative singular value spectra (SVD) for two representative subjects from the Fatigue dataset. Subject 21 (robust) exhibits a steep spectrum, with most spatial variance captured by a small number of components, indicating high inter-channel redundancy. Subject 40 (fragile) shows a flatter spectrum, reflecting near full-rank data where information is distributed across many channels. This difference explains the contrasting robustness behaviors observed under noise and channel masking: low-rank, redundant signals can tolerate perturbations (and may even benefit from mild noise), whereas full-rank signals collapse when spatial information is corrupted.

B DREAMER implementation results

Table 1: Multi-task performance and architecture comparison (valence)

Configuration	Parameters	Best Acc.	Train-Test Gap
EEGDeformer across tasks (depth=4, 100 epochs):			
Valence	477K	75.56%	23.13%
Arousal	477K	72.78%	25.50%
Dominance	477K	73.59%	25.01%
Architecture comparison (valence, 30 epochs):			
CNN (3 layers)	47K	**74.82%**	4.13%
EEGDeformer (depth=4)	477K	74.53%	7.19%
BiLSTM (2 layers)	559K	60.36%	0.44%

Table 2: Depth ablation study (valence, 100 epochs for L=1; 30 epochs for others)

Depth (L)	**Parameters**	**Best Acc.**	**Train Acc.**	**Overfitting Gap**
1 layer	∼120K	**75.87%**	91.13%	15.26%
2 layers	∼240K	75.34%	93.45%	18.11%
4 layers	477K	75.56%	98.69%	23.13%
6 layers	∼720K	75.42%	99.12%	23.70%

Table 3: Temporal structure sensitivity (valence, 30 epochs)

Condition	**CNN**	**EEGDeformer**	**CNN Drop**	**Deformer Drop**
Original	74.20%	75.95%	—	—
Reversed	74.53%	74.42%	+0.33%	-1.53%
Block-shuffled	74.27%	75.69%	+0.06%	-0.26%
Random-shuffled	58.85%	58.85%	**-15.35%**	**-17.09%**

C Robustness Analysis Results

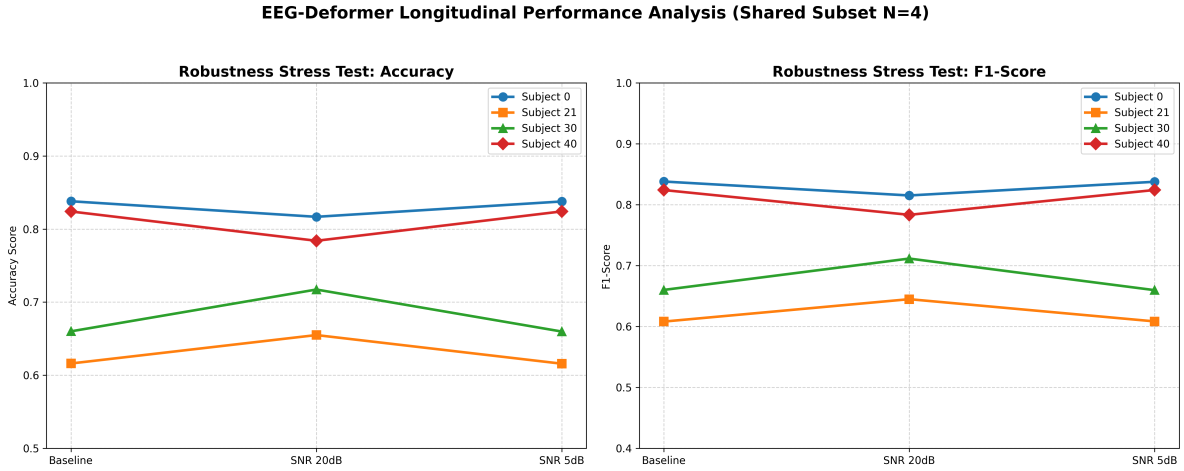

Figure 2: EEG-Deformer performance under controlled Gaussian noise injection. Accuracy and F1-scores degrade minimally at 20 dB and remain above 0.80 for most subjects even at 5 dB, confirming partial resilience to low-SNR EEG conditions.

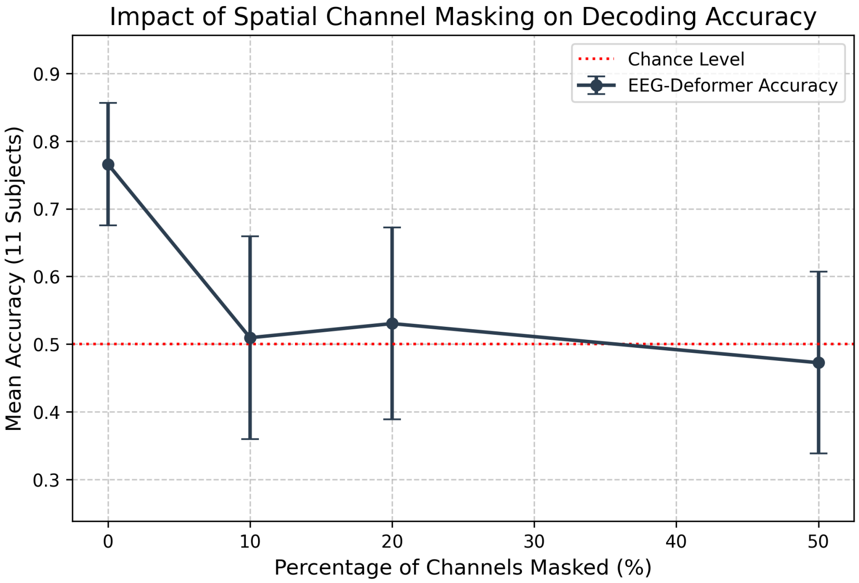

Figure 3: EEG-Deformer performance under controlled Gaussian noise injection. Accuracy and F1-scores degrade minimally at 20 dB and remain above 0.80 for most subjects even at 5 dB, confirming partial resilience to low-SNR EEG conditions.