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 4, 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 and other cognitive tasks using research-grade EEG equipment, its effectiveness under realistic deployment constraints remained unexplored. We address this gap through two complementary experimental tracks. First, we evaluate EEG-Deformer on the DREAMER [4] dataset, which employs consumer-grade EEG devices (Emotiv EPOC, 14 channels) for emotion recognition, representing a scenario closer to practical applications. Specifically, we investigate whether the architectural sophistication of EEG-Deformer provides tangible benefits when signal quality and channel density are limited, conducting systematic ablation studies on model depth and temporal structure sensitivity. Second, we directly address the authors' acknowledged limitation regarding model robustness by performing controlled signal-to-noise ratio (SNR) degradation experiments on the Fatigue dataset, testing performance under noise levels ranging from laboratory conditions (20 dB) to clinically challenging scenarios (5 dB) representative of ambulatory monitoring with movement artifacts.

Contributions and methodology. Responsibilities were equally distributed: Yasmin focused on applying EEG-Deformer to DREAMER, conducting architectural ablation studies (model depth, temporal structure analysis) and baseline comparisons with CNN and LSTM architectures. Tiffney investigated model robustness through controlled SNR degradation experiments on the Fatigue dataset, evaluating performance under laboratory (20 dB) and clinically challenging (5 dB) noise

conditions. We utilized the Deformer class from the original implementation [2], adapting it for DREAMER and binary emotion classification. Beyond this, we developed new code for DREAMER data loading and preprocessing, ablation study frameworks, temporal perturbation analysis, baseline model implementations, and noise injection pipelines for robustness evaluation. Our experiments extend the original paper by: (1) evaluating on consumer-grade EEG (DREAMER) not tested by the authors; (2) conducting systematic depth ablations absent from the original work; (3) analyzing temporal structure sensitivity to understand what features models exploit; (4) directly testing the authors' acknowledged robustness limitations through controlled SNR degradation.

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:

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

3 Data

We test 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 several practical hurdles not found in the original EEG-Deformer benchmarks. Because the 14-channel data was captured using a consumer-grade Emotiv EPOC (128 Hz), the signal-to-noise ratio is notably lower than what you'd see from a research-grade setup. Since the authors didn't apply any filtering, the raw signal is cluttered with 50 Hz interference, baseline wander, and muscle artifacts from facial expressions. Furthermore, the limited 14-channel montage and the high inter-subject amplitude variance—driven by the inconsistent contact quality of dry electrodes—create a much more challenging environment for learning spatial patterns compared to denser, 32-channel recordings.

Label quality introduces another layer of complexity. Because valence, arousal, and dominance are self-reported on a 1 to 5 scale, the data is subjective, and therefore identical clips often elicit wildly different ratings across the 23 subjects. This creates a level of label noise that is essentially impossible for any architecture to fully resolve. While we treated these scores as a binary classification using a threshold of $\tau = 3$ to achieve a balanced 50/50 split, the boundary between "high" and "low" states remains fuzzy. Furthermore, with only 414 total trials (18 per subject), the dataset is quite small for deep learning. Even after segmenting the data into 3-second windows to boost sample counts, we face a high risk of pseudo-replication, as these thousands of samples still originate from a very limited pool of independent recordings

These characteristics make DREAMER a stress test for architectural robustness: if EEG-Deformer's hierarchical attention provides genuine benefits, they should manifest even under noisy, low-channel, subjectively-labeled conditions.

Fatigue Dataset The Fatigue dataset [1] comprises 32-channel EEG from 27 subjects during 90-minute VR driving sessions, recorded at 500 Hz with research-grade equipment (Scan SynAmps2 Express). Unlike DREAMER, this data underwent extensive clinical preprocessing: 1–50 Hz bandpass filtering, manual eye-blink removal, and Automatic Artifact Removal (AAR) for residual ocular and muscular contamination.

The resulting signals are substantially cleaner, with well-defined neural oscillations visible in the alpha and theta bands associated with drowsiness. Labels are from objective behavioral measures, reaction time to lane departures, rather than subjective self-report, yielding cleaner class separation. The exclusion of intermediate-RT trials (neither clearly alert nor fatigued) further reduces label ambiguity.

However, cross-subject variability remains substantial. Baseline accuracy on clean data ranges from 0.62 to 0.84 across subjects, reflecting individual differences in fatigue-related neural signatures and residual recording quality variations. This heterogeneity, acknowledged by the original authors, motivates our within-subject noise analysis.

Diagnosis DREAMER and Fatigue span the realistic deployment range for EEG-based BCI. DREAMER's native noise, limited channels, and subjective labels test whether architectural complexity helps under adverse conditions. Fatigue's clean baseline enables precise quantification of robustness limits. The consistent 73–76% ceiling on DREAMER regardless of architecture, combined with subject-dependent noise sensitivity on Fatigue, suggests that data quality (not model complexity) remains the primary bottleneck for EEG classification performance.

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 A, 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 A, 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. And, since single-layer models performed optimally, the DIP module's multi-level aggregation similarly offers no advantage.

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 A). The experiments show that the models tolerate reversed and block-shuffled data ($\pm 1.5\%$) 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 Under Controlled SNR Degradation

The authors acknowledge high performance variance across subjects, noting the model "may exhibit greater variability ...potentially affecting its reliability." [2] We directly test robustness by adding Gaussian noise at controlled SNR levels to the Fatigue dataset. Understanding

these thresholds is clinically important: SNR of 20 dB represents typical laboratory conditions with proper electrode preparation, while 5 dB approximates challenging deployment scenarios—ambulatory monitoring with movement artifacts, suboptimal electrode contact from sweat, or patients with involuntary movements such as Parkinsonian tremor.

For matched subjects, baseline accuracies averaged 0.73. At SNR = 20 dB, mean accuracy was 0.743 ($\Delta \approx -2.8$ pp); at SNR = 5 dB, mean accuracy was 0.734 ($\Delta \approx -4.1$ pp). Degradation was notably subject-dependent: those with poor baseline signal quality showed sharper declines. This confirms the authors' concern that cross-subject variability remains a primary bottleneck for model reliability. Qualitatively, the model demonstrated significant resilience; training curves remained stable across noise levels, and confusion matrices retained a strong diagonal structure, indicating that the model did not collapse into random class assignment even under heavy interference.

4.3 Synthesis

Our experiments allow direct evaluation of EEG-Deformer's central claims. The hierarchical coarse-to-fine modeling hypothesis is **not supported** in our setting: depth ablation shows additional HCT layers increase overfitting without improving generalization, and temporal perturbation analysis reveals reliance on local coherence rather than multi-scale dynamics. The CNN's equivalent performance suggests hierarchical complexity is unnecessary when data complexity is limited, consistent with the authors' acknowledged task specificity limitation, though extending it to hardware limitations they did not discuss. The robustness claim is **partially supported**: performance remains within 95% of baseline at 20 dB, but at clinically-relevant 5 dB noise, accuracy declined $\sim$4 pp with diffuse attention maps, indicating the attention mechanism's susceptibility to signal corruption even as convolutions remain stable.

Beyond the authors' stated limitations (task specificity, high variance), our experiments expose unstated architectural concerns. The dense connections between HCT layers create "feature smear", by mixing representations across layers, the model sacrifices the temporal specificity valuable for clinical interpretability. The architecture also assumes fixed electrode positions; headset displacement common in consumer settings would disrupt learned spatial filters, as the model lacks channel permutation invariance. Finally, the model treats each EEG segment as a static snapshot despite EEG's inherent non-stationarity, ignoring dynamical transitions between brain states.

Method Adaptation. While EEG-Deformer demonstrates significant resilience to Gaussian noise validating the convolutional front-end's design for real-world signal corruption—our ablation studies suggest the architecture is not well-adapted to low-density consumer EEG. The redundancy of deep hierarchical layers on DREAMER indicates the model's complexity exceeds the signal's information entropy. Additionally, reliance on local coherence rather than global sequence suggests EEG-Deformer functions effectively as a spatial-temporal filter rather than a true long-range dependency model. The 73–76% performance ceiling across all architectures points to fundamental limits from signal quality and subject diversity that no architectural complexity can overcome.

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 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%**

B Robustness Analysis Results

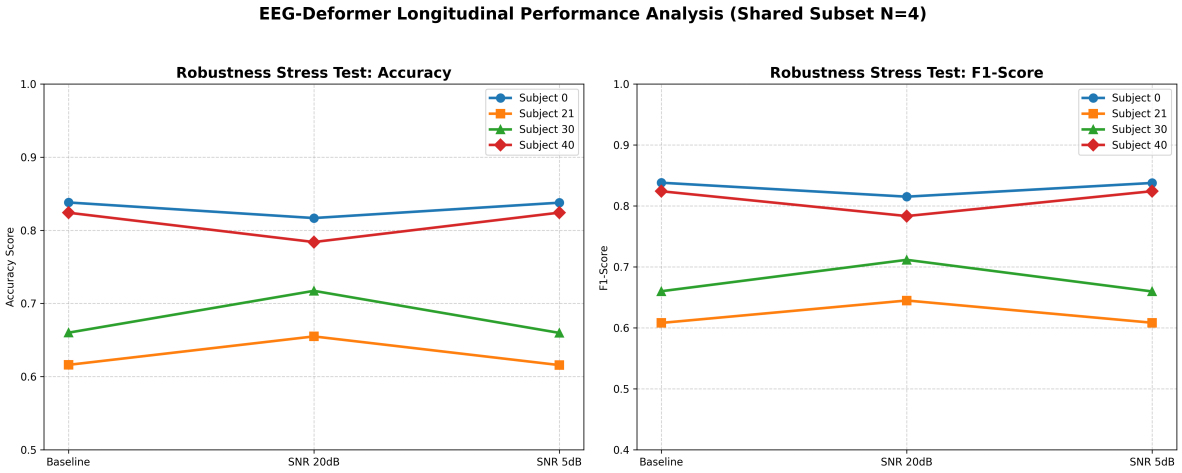

Figure 1: EEG-Deformer performance under controlled Gaussian noise injection. Accuracy and F1-scores degrade minimally at 20 dB and remain above 0.80 for most subjects even at 5 dB, confirming partial resilience to low-SNR EEG conditions.