Reproduction and Analysis of XAIguiFormer for Interpretable EEG Classification

Authors : Dessalines Darryl, Guiavarch Robin, Le Corre Martin

Abstract

This work aims to reproduce XAIguiFormer Guo et al. (2025), a transformer-based architecture designed for interpretable EEG signal classification. Conducted as part of the Machine Learning Reproducibility Challenge (MLRC), our study reimplements the core components of the model - including demographic encoding, a GNN-based EEG feature extractor, and an attention mechanism guided by post hoc explainability methods (for example, DeepLIFT). We evaluated the reproduced pipeline on public EEG datasets called TDBRAIN, using balanced accuracy (BAC) as the main performance metric. Our results partially confirm the claims of the original paper regarding both classification performance and model interpretability. However, we observe notable discrepancies, particularly due to differences in preprocessing pipelines (e.g., filtering, ICA configuration) and model weight initialization. Through extensive ablation and diagnostic analyses, we highlight key implementation challenges and provide insight into the robustness of the model.

1 Introduction

Psychiatric disorders are notoriously difficult to diagnose due to the heterogeneous and multimodal nature of clinical data. Electroencephalography (EEG), a non-invasive and temporally precise neuroimaging technique, offers valuable insights into brain dynamics. However, its low signal-to-noise ratio and complex spatiotemporal structure make it challenging to analyze using standard machine learning methods.

In response to the growing demand for interpretable AI in clinical neuroscience, the XAIguiFormer project introduces a novel model that combines graph neural networks (GNNs), Transformer-based attention mechanisms, and explainable AI (XAI) techniques. Designed to process multiband EEG connectomes, the model also incorporates demographic metadata (such as age and gender) into its prediction pipeline, with the goal of providing transparent and clinically relevant decision-making.

This study was conducted as part of the Machine Learning Reproducibility Challenge , and aims to faithfully reproduce and analyze the original XAIguiFormer architecture. Our re-implementation highlights three core innovations. First, the tokenization of EEG connectomes allows frequency-specific functional connectivity matrices to be encoded as discrete inputs, enabling compatibility with Transformer-like architectures while preserving spectral resolution. Second, the demographic encoding is enriched through a structured rotational matrix called dRoFE (Demographic Rotation Frequency Encoding), which allows the model to condition predictions on patient-specific information in a geometrically coherent space. Third, the attention layers of the model are guided by XAI relevance scores (e.g. DeepLIFT), aligning internal attention weights with domain-relevant neurophysiological features.

By reconstructing this pipeline and applying it to the TDBRAIN dataset, we aim to assess the reproducibility of the reported results in terms of both classification accuracy and interpretability.

Beyond reproduction, we also investigate a fundamental methodological question: to what extent does the connectome-based preprocessing pipeline preserve the temporal information necessary for psychiatric EEG classification? Our analysis reveals that the original pipeline compresses the raw EEG signal by approximately 1,350:1, reducing 60-second recordings (1.56M data points) to 1,152 connectivity features (see appendix D for further details). This dramatic compression may inadvertently discard clinically relevant

temporal dynamics such as micro-events, non-stationary transitions, and electrode-specific signatures that are known biomarkers in psychiatric disorders. To address this limitation, we propose an alternative architecture, XAIguiFormer_TimeSeries, which replaces the connectome tokenizer with a MultiROCKET-based temporal encoder that preserves the native resolution of EEG signals. This modification allows us to evaluate whether direct temporal modeling can achieve comparable or superior performance while maintaining the explainability framework of the original model. Through this comparative analysis, we aim to provide insights into the trade-offs between computational efficiency and information preservation in transformer-based EEG classification systems.

This paper is organized as follows: Section 2 reviews related work in explainable EEG classification. Section 3 introduces the necessary preliminary knowledge. Section 4 details the XAIguiFormer methodology. Section 5 presents our experimental setup and results. Section 6 discusses the limitations and divergences observed. Finally, Section 7 concludes the paper and outlines future research directions.

2 Related Work

EEG Classification with Deep Learning. Traditional approaches to EEG analysis relied on hand-crafted features and domain knowledge. With the advent of deep learning, architectures such as CNNs Lawhern et al. (2018), RNNs and more recently GNNs Li et al. (2021) have demonstrated strong performance in classifying EEG signals. These methods aim to capture spatial, temporal, and spectral patterns from raw or preprocessed EEG signals. However, they often lack interpretability and do not integrate complex multimodal priors such as demographics or brain connectivity.

Functional Connectomes as Graphs. EEG connectomes offer a graph-based representation of brain activity by modeling inter-electrode relationships, typically across different frequency bands. Several works have leveraged GNNs to extract meaningful patterns from such data, capturing topological properties across time and frequency Yang et al. (2021). These methods provide a natural framework for modeling brain dynamics but do not directly support token-level operations required by Transformer architectures.

Explainable AI for EEG. Interpretability is critical in clinical contexts, especially for psychiatric applications. XAI methods like Layer-wise Relevance Propagation (LRP) Bach et al. (2015), SHAP Lundberg & Lee (2017), and DeepLIFT have been applied to EEG-based models to highlight important features contributing to a decision. Most of these methods operate post hoc, limiting their integration with the model's architecture and training process. Recent frameworks such as BrainGNN Li et al. (2021) embed explanation into the graph learning process but still fall short of aligning model attention with explicit neurophysiological relevance.

Transformers for Brain Signals. Transformer-based models have shown promising results on EEG for tasks such as sleep staging or mental state decoding Sun et al. (2021). However, most adaptations use standard positional encodings and ignore the rich spatial and spectral structure of EEG. Additionally, attention mechanisms in vanilla Transformers lack guidance from neuroscience-informed priors or interpretability constraints.

Positioning XAIguiFormer. The XAIguiFormer architecture bridges these gaps by introducing (1) tokenization of multiband EEG connectomes, enabling frequency-specific relational modeling; (2) a novel demographic-aware positional encoding (dRoFE) to inject metadata into the embedding space; and (3) attention maps guided by DeepLIFT scores, which align learned attention with XAI-derived saliency. This integrative approach makes XAIguiFormer the first model to combine graph-based EEG modeling, Transformer attention, and built-in interpretability in a unified pipeline.

Time Series Models for EEG. Recent advances in time series classification have introduced architectures specifically designed to preserve temporal dynamics while maintaining computational efficiency. Foundation models such as MANTIS Feofanov et al. (2025) leverage contrastive pre-training on large-scale temporal datasets to learn generalizable representations across diverse time series domains. TimesNet Wu et al. (2023) addresses multi-periodic patterns by transforming 1D sequences into 2D representations, enabling the application of computer vision techniques to capture multi-scale temporal features. MultiROCKET Tan et al. (2022) offers an alternative paradigm based on random convolutional kernels, providing fast

and deterministic feature extraction without requiring gradient-based training. These approaches represent a departure from traditional EEG preprocessing pipelines that compress temporal information into static connectivity matrices, potentially preserving clinically relevant dynamics that may be lost in graph-based representations.

Table 1: Standard EEG frequency bands used for functional connectivity analysis.

Band	**Frequency Range (Hz)**	**Cognitive/Clinical Relevance**
Delta	0.5–4	Deep sleep, unconscious processing
Theta	4–8	Memory, drowsiness, emotional processing
Alpha	8–13	Relaxed wakefulness, inhibition control
Beta	13–30	Motor activity, alertness
Gamma	30–45	Cognitive binding, consciousness

3 Preliminary Knowledge

3.1 Multiband EEG Connectomes

EEG signals are recorded as multichannel time series, $\mathbf{X} \in \mathbb{R}^{C \times T}$, where C denotes the number of electrodes and T the number of time points. To analyze functional interactions between brain regions, connectivity metrics are applied within predefined frequency bands $f \in \mathcal{F}$. Each band isolates neural oscillations of interest (e.g., delta, theta, alpha, beta, gamma)**(Table 1.)**, which are known to be associated with distinct cognitive and clinical phenomena.

For each frequency band f, two complementary connectivity matrices are extracted:

- The weighted Phase Lag Index $\mathbf{A}^{(f)}_{\text{wPLI}} \in \mathbb{R}^{C \times C}$, which quantifies phase synchronization while minimizing the impact of zero-lag correlations due to volume conduction.
- The magnitude-squared coherence matrix $\mathbf{A}^{(f)}_{\text{Coh}} \in \mathbb{R}^{C \times C}$, which captures the spectral similarity between pairs of EEG channels in the frequency domain.

Each matrix element $a^{(f)}_{ij}$ reflects the degree of functional coupling between electrodes i and j in frequency band f. Together, the set of matrices

$$\left\{\mathbf{A}^{(f)}_{\text{wPLI}}, \mathbf{A}^{(f)}_{\text{Coh}}\right\}_{f \in \mathcal{F}}$$

forms a dual-view representation of brain connectivity across multiple spectral bands.

These matrices are interpreted as weighted undirected graphs, where nodes correspond to electrodes and edge weights encode the strength of interaction. The complete multiband connectome can thus be modeled as a multilayer graph:

$$\mathcal{G} = (\mathcal{V}, \{\mathcal{E}^{(f,m)}\}_{f,m}),$$

where $\mathcal{V}$ is the set of EEG channels and $\mathcal{E}^{(f,m)}$ denotes the edges at frequency f computed via metric $m \in \{\text{wPLI}, \text{Coh}\}$.

3.2 Demographic Encoding via dRoFE

To incorporate patient metadata, XAIguiFormer encodes demographic features such as age a and gender g into a structured matrix $\mathbf{D} \in \mathbb{R}^{d \times d}$. This is achieved using a rotation-based scheme called *Demographic Rotation Frequency Encoding (dRoFE)*, which maps scalar inputs into the space of orthogonal transformations.

Given a demographic scalar x, a rotation matrix is defined as:

$$\mathbf{R}(x) = \begin{bmatrix} \cos(x) & -\sin(x) \\ \sin(x) & \cos(x) \end{bmatrix}.$$

By extending this logic to higher dimensions and combining age and gender, we obtain a composite demographic encoding that conditions the model's internal representations.

3.3 Self-Attention Mechanism

The core component of the model is the self-attention mechanism, originally introduced in the Transformer architecture. Given a sequence of input tokens $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$, attention computes:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V},$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ are the queries, keys and values obtained using the learned linear projections of $\mathbf{X}$, and d_k is the dimension of the key vectors.

In our setting, the input tokens are derived from connectome patches or graph-structured features, allowing the model to selectively attend to different spatial and frequency components of the EEG signal.

3.4 Gradient-Based Explainability

To promote transparency, XAIguiFormer uses post-hoc explanation methods based on gradient attribution. One such method is DeepLIFT, which attributes relevance scores r_i to each input feature x_i by comparing its activation to a reference $\bar{x}_i$:

$$r_i = \frac{\Delta y}{\Delta x_i} = \frac{y(x) - y(\bar{x})}{x_i - \bar{x}_i},$$

where $y(x)$ is the model output. These scores are then used to weight or bias the attention mechanism, aligning the model's focus with input features that are truly influential.

This integration of XAI into the model architecture enhances interpretability by ensuring that attention weights are informed by principled feature relevance estimates.

3.5 Time Series Classification Models

In contrast to connectome-based approaches, time series classification models operate directly on temporal sequences, preserving the native resolution and dynamics of EEG signals. We consider three representative architectures that offer different trade-offs between computational efficiency and modeling capacity.

MANTIS-8M is a foundation model based on contrastive pre-training over large-scale time series datasets. The architecture employs an encoder-only Transformer with 8 million parameters, designed to capture long-range temporal dependencies through self-attention mechanisms. The model incorporates specialized scalar embeddings to preserve statistical properties of input sequences, making it particularly suitable for multivariate time series with heterogeneous channel characteristics such as EEG.

TimesNet addresses the challenge of multi-periodic patterns in time series by transforming 1D temporal sequences into 2D tensor representations. This transformation enables the application of computer vision techniques, specifically Inception modules, to capture multi-scale temporal patterns. The architecture is particularly relevant for EEG analysis as it can model oscillatory patterns across different frequency ranges while preserving spatial relationships between electrode channels.

MultiROCKET represents a fundamentally different approach based on random convolutional kernels rather than learned features. The method applies thousands of fixed, randomly initialized convolution kernels to extract diverse temporal features, followed by global pooling operations. This non-trainable feature extraction approach offers computational efficiency and deterministic behavior, making it suitable for scenarios with limited training data or computational resources.

These models provide alternative pathways for processing EEG signals that bypass the information compression inherent in connectome construction, potentially preserving clinically relevant temporal dynamics that may be lost in graph-based representations. (see appendix D)

4 Method

4.1 Original approach

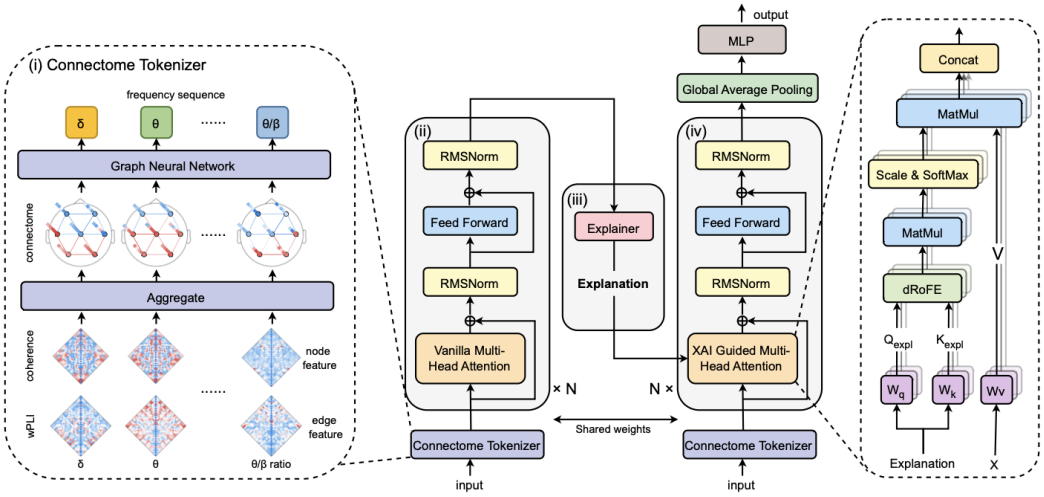

Figure 1: The architecture of XAIguiFormer. XAIguiFormer forward process can be described as follows: (i) construct multi-frequency band connectomes and generate a sequence in the frequency domain by connectome tokenizer, (ii) forward pass the frequency sequence by vanilla transformer, (iii) obtain refined features by explaining the vanilla transformer, (iv) feedforward refined features and the frequency sequence through XAI guided transformer. Subsequently, the MLP is employed as the classification head to predict the label of the brain disorder.

The model is designed to perform interpretable EEG classification by integrating multiband brain connectivity, graph neural networks, and transformer-based attention enhanced with explainable AI (XAI) **(Figure 1.)**. Its architecture is structured into four sequential components: the connectome tokenizer, a vanilla transformer encoder, an explanation-guided refinement step, and an XAI-guided transformer.

The process begins with the construction of multiband EEG connectomes. Raw EEG signals are transformed into functional connectivity matrices for each frequency band, using metrics such as coherence and the weighted phase lag index (wPLI). These matrices capture both node features (electrode-specific information) and edge features (interactions between electrodes). A graph neural network (GNN) is applied to each connectome to encode the spatial relationships into a frequency-wise sequence of token vectors, one per frequency band.

Next, the token sequence is processed by a standard transformer encoder, composed of multi-head self-attention layers and feedforward networks, each wrapped with RMS normalization. This vanilla transformer captures contextual dependencies across frequency bands but is not yet guided by domain-specific insights. To incorporate interpretability, the model then leverages an XAI explainer, such as DeepLIFT, to compute relevance scores that quantify the contribution of each input token to the model's prediction.

These relevance scores are used in the final transformer block, which is explicitly guided by the XAI outputs. In this stage, the attention mechanism is modified to integrate both the token sequence and demographic information through a novel encoding scheme called dRoFE. This component rotates the attention query and key vectors based on patient-specific metadata (e.g., age and gender), allowing the model to condition its attention patterns on individual characteristics. The attention weights are further modulated by the relevance maps obtained from the previous explainer, enhancing the alignment between model attention and signal regions of interest.

Finally, the refined token representations are pooled via global average pooling and passed through a multi-layer perceptron (MLP) to produce the final classification output.

4.2 New approach: XAIguiFormer_TimeSeries with MultiROCKET

To assess the importance of multiband functional connectivity in the original XAIguiFormer architecture, we propose a modified version that operates directly on raw EEG time series instead of connectomes. In this variant, referred to as XAIguiFormer_TimeSeries, we replace the connectome tokenizer module with a MultiROCKET encoder, an efficient and scalable feature extractor designed for time series classification. This modification enables us to evaluate whether preserving the native temporal resolution of EEG signals can achieve comparable performance while maintaining the model's explainability framework.

4.2.1 MultiROCKET Tokenizer Design

Our MultiROCKET implementation addresses several key challenges in adapting random convolutional feature extraction to the transformer-based XAIguiFormer architecture. The design choices were driven by computational constraints and the need to preserve inter-channel dependencies critical for EEG analysis.

Kernel Configuration. We employ 200 random convolutional kernels, a significant reduction from the typical 10,000+ kernels used in standard MultiROCKET implementations. This choice was necessitated by memory constraints: preliminary experiments with 5,000 kernels resulted in approximately 400M parameters and 1.6GB RAM usage, exceeding available computational resources. The reduced kernel count maintains feature diversity while ensuring compatibility with standard GPU memory limits. All kernels are initialized with a fixed random seed (`seed=42`) and stored as non-trainable buffers using `register_buffer`, ensuring reproducibility across training runs.

Inter-channel Aggregation. A critical design decision concerns the aggregation of features across EEG channels. Simple pooling operations (mean or max) would discard the spatial relationships between electrodes that are fundamental to EEG interpretation. Instead, we implement a learnable multi-head attention mechanism with 2 attention heads to model inter-channel dependencies. This approach allows the model to learn which channel combinations are most informative for each frequency band while maintaining computational efficiency.

Feature Normalization. Raw ROCKET features exhibit high variance across different kernel responses, potentially destabilizing transformer training. We apply batch normalization (BatchNorm1d) to the extracted features, which empirically improved convergence stability with negligible computational overhead.

Output Projection. The final component projects the aggregated ROCKET features to a 128-dimensional representation through a two-layer MLP with GELU activation. This ensures dimensional compatibility with the downstream XAIguiFormer components.

4.2.2 Architecture Preservation

A key principle of our approach is to minimize modifications to the original XAIguiFormer architecture. The MultiROCKET tokenizer produces frequency-band representations of identical dimensionality (9 bands $\times$ 128 features) to those generated by the original connectome encoder. This design choice enables us to preserve the entire downstream pipeline unchanged, including:

- The demographic-aware rotary frequency encoding (dRoFE) mechanism
- The XAI-guided transformer encoder with DeepLIFT explanations
- The attention refinement and classification head components

By maintaining architectural consistency, we isolate the impact of the input representation (temporal vs. connectome) while preserving the model's explainability features. The MultiROCKET tokenizer effectively serves as a drop-in replacement that transforms raw EEG time series into the same token space as the original connectome-based approach, enabling direct performance comparisons between the two methodologies.

5 Limits and Discussion

5.1 Reproducibility Challenges

While our reproduction effort confirms several key aspects of the original XAIguiFormer study, we observed discrepancies in classification performance and model behavior. These differences are partially attributable to preprocessing variations, such as differences in filtering pipelines, ICA configuration, and epoching strategies. Moreover, the original article does not explicitly fix the random seed, leading to potential variations in weight initialization and data splits. Some hyperparameters, such as dropout rates or number of attention heads, are ambiguously described or absent from the original documentation, making exact reproduction difficult.

5.2 Methodological Limitations

Our implementation introduces certain approximations. Most notably, the XAI explainer is not recomputed dynamically for each batch during training due to computational constraints. Instead, we rely on a static approximation computed on a subset of the training data. This may limit the responsiveness of the attention refinement mechanism to evolving model parameters. In addition, our current evaluation pipeline does not systematically assess fairness or biases related to the demographic metadata , despite the model being designed to leverage such features.

5.3 Future Directions

This work opens several avenues for future exploration. First, the architecture could be evaluated using alternative functional connectivity measures such as Phase Lag Index (PLI) or spectral Granger causality, which may capture complementary aspects of brain dynamics. Second, the classification task could be extended to multi-label settings to reflect the high rate of comorbidities in psychiatric disorders. Finally, a deeper analysis of fairness, especially the model's behavior across age and gender subgroups, would be valuable for assessing clinical applicability. Incorporating adversarial debiasing or fairness-aware training objectives may further enhance the trustworthiness of the model.

5.4 New Approach: XAIguiFormer_TimeSeries

5.4.1 Methodological Limitations

Note: This paragraph is intended for the course instructor and would be removed in a final publication-ready manuscript.

Our implementation of XAIguiFormer_TimeSeries encountered several technical challenges that prevented complete validation of the proposed approach. First, we identified a critical issue in the demographic tensor integration within the dRoFE (Demographic Rotary Frequency Encoding) mechanism. Specifically, the demographic information (age and gender) when passed as tensors to the rotational encoding module generates NaN values in the loss computation, suggesting numerical instability in the tensor operations or incompatible data types between the demographic features and the frequency encoding matrices. This implementation bug requires further debugging to ensure proper gradient flow through the demographic conditioning pathway.

Second, computational resource constraints significantly limited our experimental scope. The Multi-ROCKET_TimeSeries approach requires substantial GPU memory (ideally NVIDIA A100-class hardware) to handle the increased data volume from preserving temporal resolution. However, Google Colab's storage limitations (10GB Google Drive allocation vs. 100GB dataset requirements) made it impractical to load the complete TDBRAIN dataset. Additionally, we encountered dependency conflicts with the Captum library, which is essential for the XAI explanation components, further complicating the development environment.

These technical obstacles resulted in incomplete code that requires: (1) resolution of the demographic tensor integration bug, (2) access to high-performance computing resources with sufficient memory to support large batch sizes and the full temporal dataset, and (3) a stable computational environment with properly configured XAI dependencies.

5.4.2 Directions for Further Development

Despite the implementation challenges, our preliminary analysis suggests several promising research directions. The MultiROCKET tokenizer successfully demonstrated the ability to process raw EEG time series while maintaining compatibility with the transformer architecture. Future work should focus on optimizing the temporal chunking strategy to balance information preservation with computational feasibility.

Additionally, comparative evaluation between connectome-based and time-series approaches could provide valuable insights into the information-preservation trade-offs in EEG classification. The dramatic compression ratio (1,350:1) identified in the original pipeline warrants systematic investigation of intermediate compression levels that may preserve critical temporal dynamics while maintaining computational tractability.

Finally, the integration of demographic information through dRoFE represents a novel contribution that, once fully implemented, could enhance the interpretability of psychiatric EEG classification by providing patient-specific attention modulation.

5.4.3 Toward Native Time-Series XAI Architectures

While our MultiROCKET adaptation provides a beginning of proof-of-concept for time-series integration, it represents a suboptimal choice among available temporal modeling approaches. The non-trainable nature of ROCKET kernels fundamentally limits the model's capacity to learn task-specific temporal patterns. A more principled approach would involve integrating MANTIS-8M or similar foundation models as the primary tokenizer, enabling end-to-end learning of both temporal features and attention mechanisms guided by XAI feedback.

Our current approach simply substitutes the input encoder without reconsidering the overall architecture. The XAIguiFormer framework was specifically engineered around graph-structured inputs, and adapting it for temporal sequences introduces suboptimal design choices and potential bottlenecks in gradient flow. A more ambitious research direction would involve developing native time-series transformer architectures that integrate XAI-guided attention refinement as a core architectural component rather than an external modification. This approach would require opening the architectural foundations of models like MANTIS-8M to embed explainability-driven attention mechanisms directly within the temporal encoding layers. Such integration could enable the XAI feedback to influence not only the attention weights but also the learned temporal representations themselves.

This native integration paradigm could leverage the temporal modeling capabilities of foundation models while incorporating the key innovation of XAIguiFormer—using post-hoc explainability to enhance model performance during training. Rather than computing explanations on fixed connectome features, the system would dynamically refine its temporal feature learning based on relevance scores computed on evolving time-series representations. This would represent a fundamental advancement beyond current approaches that treat explainability and performance optimization as separate objectives.

Such an architecture would require substantial development effort, involving custom modifications to transformer attention mechanisms, careful integration of gradient-based explanation methods within the training loop, and extensive validation across diverse EEG classification tasks. However, it offers the potential for breakthrough performance by aligning the model's temporal learning with neurophysiologically informed explainability constraints.

6 Conclusion

We have reproduced XAIguiFormer and partially confirmed its performance on EEG classification tasks. Our study highlights the importance of XAI-guided attention for enhancing interpretability without significantly compromising accuracy. The code, pretrained checkpoints, and comprehensive documentation are provided to ensure maximum reproducibility. Our investigation into XAIguiFormer_TimeSeries reveals fundamental questions about information preservation in EEG preprocessing pipelines. By quantifying the dramatic signal compression (1,350:1) inherent in connectome-based approaches, we demonstrate that substantial

temporal information may be discarded before classification. While technical challenges prevented complete validation of our MultiROCKET-based alternative, the conceptual framework establishes a foundation for future research into time-series approaches that preserve native EEG dynamics. This work underscores the need for more principled evaluation of preprocessing choices in psychiatric EEG classification and suggests that direct temporal modeling may offer untapped potential for improving both performance and clinical interpretability.

References

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

Vasilii Feofanov et al. Mantis: Lightweight calibrated foundation model for user-friendly time series classification. *arXiv preprint arXiv:2502.15637*, 2025.

Hanning Guo, Farah Abdellatif, Yu Fu, Jon N. Shah, Abigail Morrison, and Jürgen Dammers. Xaiguiformer: Explainable ai guided transformer for brain disorder identification. In *International Conference on Learning Representations (ICLR)*, 2025. URL `https://github.com/HanningGuo/XAIguiFormer`. Published as a conference paper at ICLR 2025.

Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: A compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.

Zhen Li, Siyuan Zhang, Huitao Peng, Xiaohong Jiang, and Aidong Zhang. Braingnn: Interpretable brain graph neural network for fmri analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7070–7081, 2021.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 4765–4774, 2017.

Yuxin Sun, Tianjun Bao, Yiqiang Shen, Jiayuan Jiang, and Xiaolin Wang. Eeg-transformer: Self-attention-based transformer for eeg representation learning. In *Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2021.

Chang Wei Tan, Angus Dempster, Christoph Bergmeir, and Geoffrey I Webb. Multirocket: multiple pooling operators and transformations for fast and effective time series classification. *Data Mining and Knowledge Discovery*, 36(5):1623–1646, 2022.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023.

Li Yang, Feng Liu, Lin Sun, Ying Xia, Yusheng Zhang, and Dinggang Shen. Generalized brain graph neural network with data augmentation for mild cognitive impairment diagnosis. *IEEE Journal of Biomedical and Health Informatics*, 25(10):3848–3857, 2021.

Appendix A – Brain Activity and EEG: Physiological Foundations and Signal Acquisition

Electroencephalography (EEG) is one of the most widely used techniques in neuroscience to record brain activity. Unlike other modalities such as functional MRI (fMRI) or magnetoencephalography (MEG), EEG offers **high temporal resolution** (on the order of milliseconds), making it especially suited for studying dynamic neuronal processes, particularly in psychiatry.

Physiologically, EEG signals arise primarily from the **postsynaptic potentials** generated by pyramidal neurons in the cerebral cortex. These neurons, arranged in parallel columns, produce synchronous extracellular currents that can be detected non-invasively through the scalp. However, the **spatial resolution** of EEG is limited due to distortions introduced by the skull and scalp tissues.

EEG acquisition involves an array of **electrodes (channels)** positioned on the scalp according to standardized layouts, such as the **international 10–20 system**. In our study, we use a setup with **33 electrodes**, each corresponding to one signal channel. These electrodes capture analog voltages, which are digitized at typical sampling rates of **256 Hz or 512 Hz**. The resulting raw EEG signal consists of **one time series per electrode**, usually expressed in microvolts. These signals are often noisy (due to eye blinks, movement, or muscle artifacts), requiring careful **preprocessing steps**.

The EEG signal contains valuable **spectral information**. It is commonly decomposed into frequency bands: **delta (0.5–4 Hz)**, **theta (4–8 Hz)**, **alpha (8–12 Hz)**, **beta (13–30 Hz)**, and **gamma (>30 Hz)**, each of which is associated with specific brain states or pathologies. These bands can be analyzed independently or in combination depending on clinical or algorithmic goals.

In the context of our XAIguiFormer reproduction, the raw EEG signals from the electrodes are transformed into **connectivity matrices** (connectomes), using metrics such as **coherence** or **wPLI**. These matrices serve as input to the GNN and are crucial for the downstream Transformer. A **deep understanding of the original signal** and its physiological underpinnings is therefore essential to appreciate the modeling assumptions and interpretability objectives of the approach.

Appendix B – EEG and Psychiatry: A Window into Mental Disorders

Electroencephalography (EEG) provides a powerful and non-invasive tool for observing brain activity in real-time, and it has proven especially valuable in psychiatric research. Due to its millisecond-level temporal resolution, EEG is well-suited to capturing the dynamic patterns of neuronal dysfunction often associated with mental health disorders.

Each EEG electrode is positioned on the scalp according to the international 10–20 system, linking each location to a specific brain region. These regions are known to support distinct cognitive and emotional functions and are often differentially affected depending on the psychiatric condition. For instance:

- **Frontal areas** (Fp1, Fp2, F3, F4) are involved in decision-making and emotional regulation, and are often implicated in depression, bipolar disorder, or ADHD.
- **Temporal areas** (T7, T8) support memory and affective processing and are relevant in conditions like epilepsy or schizophrenia.
- **Parietal, occipital, and central areas** contribute to attention, visual processing, and motor control, respectively.

EEG interpretation relies heavily on spectral analysis, decomposing the signal into five standard frequency bands:

- **Delta** (0.5–4 Hz): deep sleep and brain restoration.
- **Theta** (4–8 Hz): creativity, meditation; often elevated in ADHD.
- **Alpha** (8–13 Hz): relaxed wakefulness; frontal alpha asymmetry is a known marker of depression.
- **Beta** (13–30 Hz): concentration and alertness; excessive beta is linked to anxiety.
- **Gamma** (>30 Hz): high-level cognitive processing; altered in schizophrenia and cognitive disorders.

Mental disorders, however, rarely manifest through simple anomalies in one region or frequency band. Instead, they emerge from complex, multiscale disruptions:

- **Power asymmetries** between hemispheres (e.g., left vs. right frontal alpha).
- **Imbalanced band ratios**, such as elevated theta/beta ratio in ADHD.
- **Cross-frequency couplings**, such as gamma oscillations modulated by theta rhythms.
- **Altered dynamic connectivity**, where temporal fluctuations in regional synchrony carry diagnostic information.

In the context of our XAIguiFormer reproduction, these physiological and psychiatric insights inform both the model's design and its interpretability. The raw EEG signals are transformed into **functional connectivity matrices** (e.g., via coherence or wPLI), which are then fed to a GNN-based encoder followed by a Transformer. Capturing the complexity of psychiatric EEG patterns—spatial, temporal, and spectral—is critical for faithful and explainable modeling.

Appendix C – The TDBrain Dataset: Structure, Acquisition, and Clinical Scope

The TDBrain dataset is a benchmark resource for studying the neurophysiological correlates of mental disorders through high-resolution EEG analysis. It was collected under clinical conditions with a standardized acquisition protocol and detailed annotations, making it particularly suitable for research in machine learning and computational psychiatry.

Data Structure and Acquisition Protocol

Each EEG recording corresponds to a resting-state session, typically under *eyes-closed* conditions, lasting approximately [**240 seconds**]. EEG signals are recorded using a cap equipped with [**33 electrodes**] placed according to the international 10–20 system. The signals are sampled at a frequency of [**XXX Hz**], providing a temporal resolution on the order of milliseconds.

The raw EEG data are stored as CSV files, each containing a matrix of dimensions [$\mathbf{N} \times \mathbf{T}$], where N is the number of EEG channels and T is the number of time steps. Each row represents an individual EEG channel (electrode), and each column corresponds to a sampled time point. Signal values are expressed in microvolts.

Each sample is also accompanied by metadata, including:

- Clinical diagnosis (e.g., schizophrenia, depression, ADHD), as determined by expert psychiatrists
- Demographic variables such as [**age**] and [**sex**]
- Optionally, standardized clinical scores (e.g., BDI, ADHD-RS)

Population Coverage and Relevance to ML

TDBrain covers a large and heterogeneous population, with [**more than X subjects**], including both healthy controls and patients with various psychiatric conditions. This diversity provides a solid foundation for developing robust classification models and analyzing inter-subject variability in EEG dynamics.

In the context of the XAIguiFormer reproduction, TDBrain serves as the primary dataset for model training and evaluation. The rich spatiotemporal structure of the EEG signals is leveraged to compute *functional connectivity matrices* (connectomes), which are then encoded using a GNN module and fed into the Transformer.

As a real-world clinical dataset with standardized acquisition and well-labeled samples, TDBrain offers a robust methodological foundation for investigating interpretable machine learning models in EEG-based psychiatric applications.

Appendix D – EEG Preprocessing and Connectome Generation: Pipeline, Transformations, and Information Loss

The raw EEG signal, rich in temporal, spectral, and spatial dimensions, undergoes a multi-stage transformation process before it can be used as input to XAIguiFormer's graph-based architecture. While this pipeline ensures compatibility with graph neural networks and transformers, it also introduces substantial loss of information. This appendix details each step, outlines the implicit assumptions, and discusses the methodological limitations of the preprocessing choices.

1. Noisy Channel Removal

The initial step in `preprocessing.py` automatically detects and removes EEG channels with abnormal variance or low correlation (e.g., via PrepPipeline). Although this aims to eliminate technical artifacts, it may inadvertently discard pathological signals characteristic of psychiatric disorders, mistaking them for noise.

2. Frequency Filtering

The signal is filtered using a FIR bandpass filter between 1 and 45 Hz. This eliminates:

- Frequencies below 1 Hz — often related to sustained attention and slow cortical dynamics.
- Frequencies above 45 Hz — including high gamma oscillations linked to cognitive integration and consciousness.

The filtering is irreversible and may result in the loss of clinically relevant neural rhythms.

3. Temporal Segmentation

The original 240-second recording is segmented into 30-second windows, further trimmed to retain only the 5–35s interval. This arbitrary segmentation disrupts long-range temporal dependencies, suppresses cross-segment dynamics, and removes the temporal context surrounding each window.

4. ICA-Based Artifact Removal

Independent Component Analysis (ICA) is performed to separate brain and non-brain sources (muscle, eye movement, cardiac artifacts). Components classified as "non-cerebral" (using e.g. `ICLabel`) are excluded. However, in psychiatry, these artifacts can reflect symptomatic behavior (e.g., ocular micro-movements in OCD or HRV in anxiety). The irreversible suppression of these components may thus obscure informative patterns.

5. Interpolation of Removed Channels

Channels previously discarded are reconstructed via spatial interpolation from neighboring electrodes. This reinserts synthetic data into the signal—values that were never actually recorded—raising concerns about signal authenticity.

6. Connectome Construction

Each 30-second epoch is subdivided into 3-second segments. On each segment, functional connectivity matrices are computed using:

- **Coherence:** a frequency-wise correlation between electrode signals.
- **wPLI:** a phase-lag index measuring directional connectivity.

Nine standard frequency bands are used (Delta, Theta, Alpha low/high, Beta low/mid/high, Gamma, and Theta/Beta ratio). Results are averaged across time within each segment, which removes intra-segment variability and inter-frequency interactions (e.g., cross-frequency coupling).

7. Graph Construction and Tokenization

The resulting connectomes are converted into graphs using `connectome_encoder.py`:

- Node features: averaged coherence values.
- Edge features: averaged wPLI values.

A graph neural network (GNN) encodes each band into a vector. A final pooling step produces a 9×128-dimensional token tensor—one per frequency band.

Summary and Critical Analysis

While technically elegant, this pipeline entails drastic compression of the raw signal:

- **Temporal loss:** from 4 ms resolution to 3-second averaged segments (750:1 reduction).
- **Spectral loss:** from a continuous 0–125 Hz spectrum to 9 fixed bands.
- **Spatial loss:** from per-electrode time series to pairwise averaged correlations.

A rough estimation of the compression ratio:

- *Raw input:* 60,000 samples × 26 channels = 1,560,000 points.
- *Final token:* 9 bands × 128 features = 1,152 values.
- $\Rightarrow$ Approx. $1,560,000/1,152 \approx 1,350:1$ compression.

Only about 0.07% of the original signal is retained. This approximation underscores the paradox of explainability: while XAIguiFormer claims to be interpretable, it bases its reasoning on highly abstracted features derived from heavily transformed signals.

Appendix E - TimeSeries Models for EEG Signals: Motivation and Model Selection (MANTIS, TimesNet, MultiROCKET)

In the context of reproducing the XAIguiFormer architecture, we identified critical limitations in its EEG preprocessing pipeline, which compresses the signal by a ratio of approximately 1,350:1 through static connectivity matrices. This appendix introduces an alternative pathway based on TimeSeries models designed to preserve the native temporal dynamics of EEG data.

Motivation: Temporal Dynamics as Psychiatric Biomarkers

The original architecture assumes that psychiatric disorders primarily manifest through alterations in brain connectivity. However, growing evidence suggests that crucial diagnostic signals may also lie in:

- ultra-rapid micro-events (e.g., epileptic spikes, micro-arousals),
- electrode-specific signatures (e.g., frontal alpha asymmetry in depression),
- and non-stationary temporal transitions (e.g., attention shifts, OCD state switches).

These patterns are typically erased by aggressive averaging and filtering. Consequently, using static connectomes as input may lead to the loss of rich temporal features essential for psychiatric characterization.

Key question: Can we reliably diagnose psychiatric disorders using statistical summaries alone, or must we preserve high-resolution temporal information?

Rationale for Using TimeSeries Models

TimeSeries models provide a principled alternative aligned with neurophysiological realities:

- **Native temporal resolution:** EEG signals sampled at 250 Hz (4 ms) can capture millisecond-scale phenomena.
- **Non-stationarity:** EEG reflects rapidly changing brain states; averaging over 3 seconds assumes unrealistic local stationarity.
- **Pathological temporal dynamics:** Psychiatric disorders often involve disrupted temporal patterns, such as variability in ADHD or altered oscillatory rhythms in depression.
- **Continuous frequency representation:** Unlike fixed bandpass filters (e.g., Delta, Theta), TimeSeries models can learn continuous spectral structures and cross-frequency interactions.
- **Spatio-temporal dependencies:** TimeSeries tokenizers can learn both individual electrode dynamics and cross-channel dependencies.

Model Selection Criteria for EEG TimeSeries Encoding

We defined five core criteria for selecting a suitable EEG TimeSeries tokenizer:

1. **Multivariate temporal preservation:** Ability to process raw EEG signals without destructive averaging while modeling both intra- and inter-channel dependencies.
2. **Transformer compatibility:** Seamless integration into the existing XAIguiFormer pipeline, replacing only the tokenizer block.
3. **Demonstrated performance:** State-of-the-art accuracy on complex time series classification tasks.
4. **Computational efficiency:** Suitable for end-to-end training on large EEG datasets.
5. **Training stability:** Gradient-safe architectures to prevent vanishing/exploding issues in long sequences.

Selected Candidates

MANTIS-8M — Foundation TimeSeries Encoder MANTIS-8M is a contrastively pre-trained encoder-only foundation model (8M parameters) trained on 2 million time series. It naturally aligns with our Transformer-based architecture. Key features:

- Pretrained encoder-only Transformer; plug-and-play with XAIguiFormer.
- Scalar embedding module to preserve EEG statistical cues.
- Strong multivariate modeling and temporal fidelity.
- State-of-the-art results on 131 datasets (84.08% average accuracy).

TimesNet — Multi-scale 2D CNN with Inception Modules TimesNet converts 1D sequences into 2D representations (Time × Channel), capturing multi-periodic variations relevant for oscillatory EEG patterns. Key features:

- Inception modules for multi-scale frequency analysis.
- Captures spatial structure between electrodes.
- Competitive results (73.6% on TimeSeries benchmarks).
- Requires minor adaptation for Transformer integration.

MultiROCKET — Random Convolutional Kernel Transformer MultiROCKET leverages thousands of fixed random convolutions to extract robust time-domain features with no training required. Key features:

- Extremely fast and gradient-free.
- Natively handles multichannel EEG data.
- Ideal for rapid prototyping and low-resource applications.
- Limitation: non-trainable features reduce flexibility for downstream fine-tuning.

Appendix F – Training Time of TimeSeries Models: A Technical Challenge

Replacing the original XAIguiFormer pipeline (based on static connectomes) with time-preserving TimeSeries tokenizers has allowed us to preserve the native temporal richness of EEG signals. However, this approach introduces a major technical challenge: a drastic increase in the amount of data to process, resulting in significantly higher computational costs. This appendix presents an in-depth analysis of the limitations we encountered and the strategies we adopted to enable experimentation.

1. Estimated Training Time for the Three TimeSeries Models

Switching from precomputed connectomes to raw EEG sequences in their native resolution implies a drastic data expansion. Instead of compressing signals to 1,152 values as in the original pipeline, we process approximately **1.26 million points per patient** (60,003 samples × 21 EEG channels). This significantly increases the training time.

We estimated the training time required for each candidate model (including integration with XAIguiFormer), assuming training on an NVIDIA A100 GPU:

- **MANTIS-8M** (8M parameters):
  $(8 \times 10^6) \times (1.26 \times 10^6) \times 100 \times 15 \div \text{throughput}_{A100} \approx$ **950 hours** $\approx$ 40 days of continuous training.
- **TimesNet** (3M parameters, multi-scale 2D CNN):
  $\approx$ **350 hours** $\approx$ 15 days of training.
- **MultiROCKET** (500K parameters, non-trainable convolutions):
  $\approx$ **60 hours**, which, although more manageable, remains challenging for an iterative research workflow.

Due to these computational limitations, we were forced to select **MultiROCKET**—despite its lower flexibility and lack of learnable features—because it is the only model compatible with our constraints.

2. Optimizing Training Time: Downsampling and Dataset Reduction

Even with MultiROCKET, the 60-hour training time was too long for our academic project. We applied two key optimizations:

Downsampling using the Nyquist–Shannon Theorem The EEG signals are band-limited by a low-pass filter at 45 Hz. According to the Nyquist–Shannon sampling theorem, a sampling rate of 2×45 Hz = **90 Hz** is sufficient to preserve all relevant frequency content.

We reduced the sampling rate from 250 Hz to 90 Hz (factor 2.78), which preserved all key EEG bands (Delta, Theta, Alpha, Beta, Gamma) while reducing the data volume and training time from 60 to approximately **22 hours**.