Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Explainable EEG Foundation Models for Schizophrenia Detection Across Independent Datasets

Pouya Tolou Kouroshi
Iran University of Science and Technology (IUST)
Sara Bagherzadeh
Islamic Azad University
Mohammadreza Norouzi
University of Tehran
Ahmad Shalbaf
shalbaf@sbmu.ac.ir

Shahid Beheshti University of Medical Sciences

Research Article

Keywords: Electroencephalogram (EEG), Explainable AI, EEGMamba, Foundation model, schizophrenia

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11291511/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Explainable EEG Foundation Models for Schizophrenia Detection Across Independent Datasets

Pouya Tolou Kouroshi[1], Sara Bagherzadeh[2], Mohammadreza Norouzi[3], Ahmad Shalbaf[4]

[1]Biomedical Engineering Department, School of Electrical Engineering, Iran University of Science and Technology (IUST), Tehran, Iran

[2]Department of Biomedical Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran

[3]Department of Biotechnology, College of Science, University of Tehran, Tehran, Iran

[4]Department of Biomedical Engineering and Medical Physics, Faculty of Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran

Corresponding author: Ahmad Shalbaf, Department of Biomedical Engineering and Medical Physics, School of Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran. Tel: (+98)9122163201, Email: shalbaf@sbmu.ac.ir, ORCID ID: https://orcid.org/0000-0002-1595-7281.

Abstract

Schizophrenia is a complex psychiatric disorder associated with altered brain electrophysiology. EEG foundation models provide promising approaches for learning transferable neural representations; however, their effectiveness and interpretability for schizophrenia detection remain insufficiently investigated. Three pretrained EEG foundation models, EEGMamba, EEGPT, and BIOT, were systematically evaluated for schizophrenia classification, with Simple Mamba included as an additional comparison. Two independent resting-state EEG datasets were analyzed: ASEEG, comprising 51 individuals with schizophrenia (SZ) and 50 healthy controls (HC), and Warsaw, comprising 14 SZ and 14 HC participants. Subject-wise cross-validation was employed. Integrated Gradients (IG) and channel occlusion were used to investigate spatial importance and explanation consistency. EEGMamba achieved the strongest performance on both datasets. On ASEEG, it obtained 79.18 ± 8.00% accuracy, 82.33 ± 12.54% sensitivity, an F1-score of 0.79 ± 0.10, and ROC-AUC of 0.83 ± 0.09. On Warsaw, EEGMamba achieved 78.33 ± 24.90% accuracy and 80.00 ± 14.21% sensitivity. XAI revealed model- and dataset-dependent patterns involving frontal/frontopolar, central, temporal, and parietal regions, with recurrent importance of C3, C4, Cz, Fp1/Fp2, and P3/P4. On Warsaw, BIOT demonstrated strong agreement between IG and occlusion (ρ=0.744, $p<0.001$). EEG foundation models show potential for subject-independent schizophrenia detection, with EEGMamba providing the strongest predictive performance. Complementary XAI analyses revealed spatially distributed discriminative information and highlighted the importance of jointly evaluating predictive performance and interpretability across independent datasets.

Keywords: Electroencephalogram (EEG); Explainable AI; EEGMamba, Foundation model; schizophrenia

1. Introduction

Schizophrenia is a severe and chronic neuropsychiatric disorder affecting approximately 20 million people worldwide [2]. It typically emerges during adolescence or early adulthood and is characterized by positive symptoms such as hallucinations and delusions, negative symptoms including social withdrawal and reduced motivation, and cognitive impairments involving attention, memory, and executive function [1,2]. Despite extensive research, its underlying neurobiological mechanisms remain incompletely understood, making early and reliable diagnosis challenging.

Current diagnosis primarily relies on psychiatric interviews, behavioral assessments, and standardized criteria established in the Diagnostic and Statistical Manual of Mental Disorders and the International Classification of Diseases [1,2]. Although these approaches are widely accepted in clinical practice, they remain partly subjective and can be influenced by clinician experience, particularly in early-stage or atypical cases. Consequently, there is growing interest in identifying objective physiological biomarkers that can complement conventional clinical assessment and improve diagnostic reliability.

Electroencephalography (EEG) is particularly promising because it provides a non-invasive, cost-effective, and high-temporal-resolution measure of brain electrical activity [3,4]. Previous studies have identified schizophrenia-related abnormalities in neural oscillations, functional connectivity, and cortical synchronization, suggesting that EEG contains valuable information for automated diagnosis [1,2,4]. However, EEG signals are nonlinear, non-stationary, multichannel, and highly susceptible to artifacts and inter-subject variability. Traditional machine-learning methods therefore often require extensive preprocessing and handcrafted feature engineering, which may limit their generalizability [2].

Deep learning has addressed some of these limitations by enabling automatic representation learning directly from EEG data. Architectures including convolutional neural networks (CNNs), recurrent neural networks (RNNs), long short-term memory (LSTM) networks, graph neural networks (GNNs), and Transformers have been applied to capture spatial, temporal, and connectivity-related characteristics of mental disorders [5] especially schizophrenia [2–4]. Transfer-learning approaches have further improved performance when EEG datasets are limited. For example, Shalbaf et al. employed pretrained convolutional networks for schizophrenia classification [9], while Aich et al. combined EEG image representations with pretrained deep models and feature-selection techniques [10]. However, these methods largely rely on models pretrained outside the EEG domain rather than on large-scale electrophysiological data.

More recently, EEG-specific foundation models such as EEGPT, EEGMamba, and BIOT have been developed to learn transferable representations from large and heterogeneous EEG datasets [11–14]. Through Transformer-based and state-space modeling mechanisms, these models are designed to capture complex temporal, spatial, and spectral EEG characteristics. Nevertheless, their effectiveness for schizophrenia detection remains insufficiently explored. Therefore, a systematic comparison of EEGPT, EEGMamba, and BIOT under a consistent subject-independent evaluation framework is needed to determine whether EEG foundation models can provide more robust and generalizable representations for automated schizophrenia detection.

1.1. Related studies

Automated schizophrenia detection using electroencephalography has advanced substantially over the past decade. Early studies relied on handcrafted time-, frequency-, time-frequency-, entropy-, and nonlinear-domain features combined with conventional machine learning classifiers. Although promising, these approaches required considerable domain expertise and often showed limited generalizability across datasets and recording protocols [2]. Deep learning reduced this dependence by enabling automatic feature extraction. Shoeibi et al. developed a hybrid convolutional neural network–long short-term memory architecture to capture spatial and temporal electroencephalography patterns, achieving 99.25% accuracy [3]. Bagherzadeh et al. extracted effective brain connectivity using transfer entropy and combined pretrained convolutional neural networks with long short-term memory networks, reporting 99.90% accuracy and highlighting the diagnostic value of functional connectivity [4]. Mostafavi et al. converted electroencephalography signals into continuous wavelet transform scalograms and applied transfer learning with self-distillation, demonstrating robust classification using single-channel recordings [6].

Recent studies have increasingly combined complementary electroencephalography representations. Hao et al. used gamma-band features with EEGNet to achieve high recognition performance with reduced model complexity [7], whereas Soltani-Nejad et al. integrated temporal and functional connectivity features within a hybrid deep neural network and reported accuracy above 99% [8].

Other studies have explored multimodal representations, graph-based learning, and interpretability. Zan et al. combined raw multichannel electroencephalography signals with spectrograms in a two-branch convolutional neural network, achieving accuracies of 98.5% and 99.4% on two datasets [15]. Gosala et al. represented electrode interactions as graphs and combined graph convolutional networks with long short-term memory layers, obtaining 99.25 ± 0.24% average accuracy for 8-second epochs [16]. Almadhor et al. emphasized interpretability by integrating bidirectional long short-term memory networks with attention mechanisms, SHapley Additive exPlanations (SHAP), and Local Interpretable Model-agnostic Explanations (LIME) [17].

Overall, the reviews by Teixeira et al. and Saha et al. identify convolutional neural networks, long short-term memory networks, graph-based models, attention mechanisms, and transformers as major approaches for electroencephalography-based schizophrenia detection. Nevertheless, small datasets, heterogeneous preprocessing, limited cross-dataset validation, and insufficient interpretability remain important challenges, motivating the development of more transferable learning frameworks [1,2].

1.2. The gap between DL studies

Despite high reported accuracies, most EEG-based schizophrenia studies rely on small, single-center datasets and task-specific models evaluated under closely matched conditions, limiting cross-subject, cross-dataset, and clinical generalization. Variations in participants, channel montage, recording protocols, or target tasks often require substantial retraining [1–8,15–17].

Many approaches also depend on extensive preprocessing or EEG transformation into spectrograms, scalograms, or connectivity images. Differences in filtering, artifact removal, segmentation, frequency-band selection, and channel configuration can reduce reproducibility. In addition, transfer learning commonly uses CNNs pretrained on natural-image datasets such as ImageNet, whose learned features are not specifically optimized for EEG temporal, spectral, and spatial characteristics [1,2,4,6,9,10].

Interpretability remains another challenge, as high classification performance does not reveal the neurophysiological information driving predictions. Although attention mechanisms and XAI methods have improved model transparency, consistent and clinically meaningful explanations remain limited [2,17]. These limitations motivate EEG-specific foundation models pretrained on large, heterogeneous neurophysiological datasets and adapted to downstream tasks with limited fine-tuning. Because such models remain largely unexplored for schizophrenia detection, their robustness, transferability, interpretability, and cross-subject generalization require systematic evaluation [1,2,6,9,10].

1.3. Contribution of the study

Building upon these limitations, this research investigates foundation models for automated schizophrenia detection from EEG signals. Instead of training a task-specific network entirely from scratch, the proposed framework adapts pretrained EEG representations to the schizophrenia classification task. The central hypothesis is that large-scale pretraining can provide more transferable features, improve cross-subject robustness, and reduce dependence on handcrafted feature engineering and dataset-specific preprocessing.

The main contributions of this study are summarized as follows:

1. **Comparative evaluation of EEG foundation models:** EEGMamba, EEGPT, and BIOT were systematically evaluated for schizophrenia detection under a common experimental framework, with Simple Mamba included as an additional sequence-modeling baseline.

2. **Systematic optimization of EEG input and training settings:** The effects of EEG window length, recording duration, window overlap, learning rate, weight decay, and fine-tuning learning rate were investigated through grid-search optimization.

3. **Multi-method explainability analysis:** Integrated Gradients and channel occlusion were jointly applied to characterize the spatial contribution of EEG channels, providing complementary gradient- and perturbation-based explanations.

4. **Assessment of explanation consistency:** Agreement between the two XAI approaches was quantitatively evaluated using channel-wise rankings and Spearman correlation, allowing predictive performance and explanation consistency to be examined as complementary properties of EEG foundation models.

2. Material and Methods

This study used two resting-state EEG datasets for automated classification of individuals with schizophrenia and healthy controls: the Atieh Schizophrenia EEG database and the Warsaw schizophrenia EEG database. Both datasets contain recordings obtained through 19 scalp electrodes positioned according to the international 10–20 system, allowing the signals to be organized using a consistent spatial configuration. However, the databases differ in sample size, recording conditions, sampling frequency, and participant characteristics. Therefore, each dataset was initially processed independently before being adapted to the input requirements of the proposed foundation-model framework. The principal characteristics of the two databases are described below.

2.1. SZ databases

The first database is Atieh Schizophrenia EEG (ASEEG) [18,19] which is a publicly available resting-state EEG dataset collected at Atieh Clinic, Tehran, Iran, between 2020 and 2025. This database has been recorded under the supervision of Shahid Beheshti University of Medical Sciences. It includes 101 participants, comprising 51 individuals with schizophrenia and 50 healthy controls. Schizophrenia diagnoses were independently confirmed by an experienced psychiatrist and psychologist according to DSM-5 criteria. The mean ages were 31.29 ± 13.63 years for the schizophrenia group and 29.40 ± 8.30 years for healthy controls. Participants with relevant psychiatric, neurological, or major brain-related conditions that could confuse the recordings were excluded.

EEG signals were acquired using a 21-channel Mitsar EEG system at a sampling rate of 500 Hz. Nineteen scalp electrodes were positioned according to the international 10–20 system, with A1 and A2 serving as reference and ground electrodes. Electrode impedance was maintained below 5 kΩ. Recordings were performed in an electrically shielded, sound-attenuated room under controlled conditions. Each participant underwent five-minute eyes-closed and five-minute eyes-open resting-state recordings, separated by a two-minute interval. Participants were instructed to remain relaxed and minimize eye, facial, and body movements, while vigilance and recording quality were monitored throughout data acquisition. In this study we used eyes closed part of data due to fair comparison with Warsaw database. Table 1 reports characteristics of used databases.

Table 1. Characteristics of the EEG datasets used in this study.

Characteristic	ASEEG dataset	Warsaw dataset
Participants	101	28
Schizophrenia (SZ)	51	14
Healthy controls (HC)	50	14
Diagnostic criteria	DSM-5	ICD-10 (F20.0, paranoid schizophrenia)
Mean age – SZ	31.29 ± 13.63 years	28.10 ± 3.73 years
Mean age – HC	29.40 ± 8.30 years	27.75 ± 3.30 years
Sex matching	Unequal sex distribution	Age- and sex-matched
Scalp EEG channels	19	19
Electrode system	International 10–20	International 10–20
Electrodes	Fp1, Fp2, F7, F3, Fz, F4, F8, T3, C3, Cz, C4, T4, T5, P3, Pz, P4, T6, O1, O2	Fp1, Fp2, F7, F3, Fz, F4, F8, T3, C3, Cz, C4, T4, T5, P3, Pz, P4, T6, O1, O2
Sampling frequency	128 Hz	250 Hz
Recording condition	Resting-state, eyes closed	Resting-state, eyes closed
Recording duration	5 min eyes closed	5 min eyes closed *

* due to compare with ASEEG data, we selected 5 minutes from the first of Warsaw database

The second dataset was the publicly available Warsaw schizophrenia EEG database (IBIB PAN/RepOD) [20], developed in Warsaw, Poland, and available through the RepOD repository (doi:10.18150/repod.0107441) [21]. The dataset includes 28 adults, comprising 14 individuals with paranoid schizophrenia and 14 healthy controls, matched for age and sex. The mean ages were 28.10 ± 3.73 years for the schizophrenia group and 27.75 ± 3.30 years for healthy controls. Patients were diagnosed according to ICD-10 criteria (F20.0), and the original recording protocol was approved by the Ethics Committee of the Institute of Psychiatry and Neurology in Warsaw [4,21].

EEG was recorded for approximately 15 minutes under resting-state eyes-closed conditions using 19 electrodes positioned according to the international 10–20 system (Fp1, Fp2, F7, F3, Fz, F4, F8, T3, C3, Cz, C4, T4, T5, P3, Pz, P4, T6, O1, and O2) at a sampling frequency of 250 Hz. The recordings were provided as participant-level EDF files. Thus, the present study used 28 recordings (14 schizophrenia and 14 healthy controls) as a secondary analysis of the anonymized, publicly available data.

2.2. Overview of the Proposed Pipeline

Figure 1 illustrates the overall framework used for schizophrenia classification and model interpretation. Resting-state EEG recordings from the ASEEG and Warsaw datasets were first subjected to preprocessing, including re-referencing, filtering, artifact removal, segmentation, and normalization. The resulting EEG segments were then analyzed using EEGMamba, EEGPT, BIOT, and Simple Mamba within a subject-wise cross-validation framework. Model performance was evaluated using accuracy, precision, specificity, sensitivity, F1-score, and ROC-AUC. Finally, the predictions of the foundation models were interpreted using Integrated Gradients (IG) and channel occlusion, from which spatial importance maps, channel rankings, and agreement between the two XAI methods were obtained.

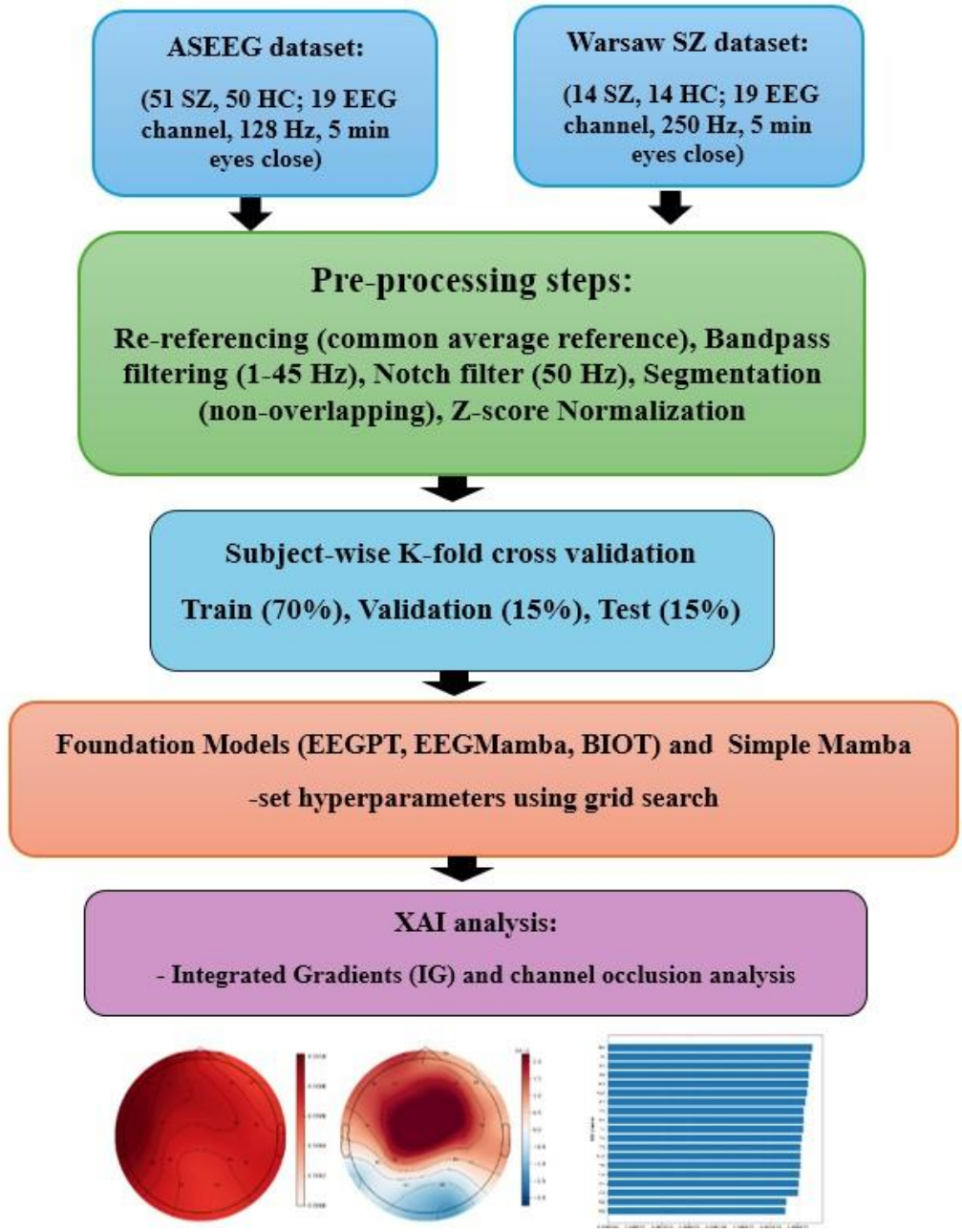


Figure 1. Overall framework for EEG-based schizophrenia classification and explainability using EEG foundation models.

2.3. EEG Foundation Models

To investigate the effectiveness of modern pretrained EEG foundation models for schizophrenia detection, three state-of-the-art architectures, namely EEGPT, EEGMamba, and BIOT, were evaluated under an identical experimental protocol. All models were initialized using publicly available pretrained weights and subsequently adapted to the schizophrenia classification task through supervised fine-tuning. To ensure a fair comparison, the

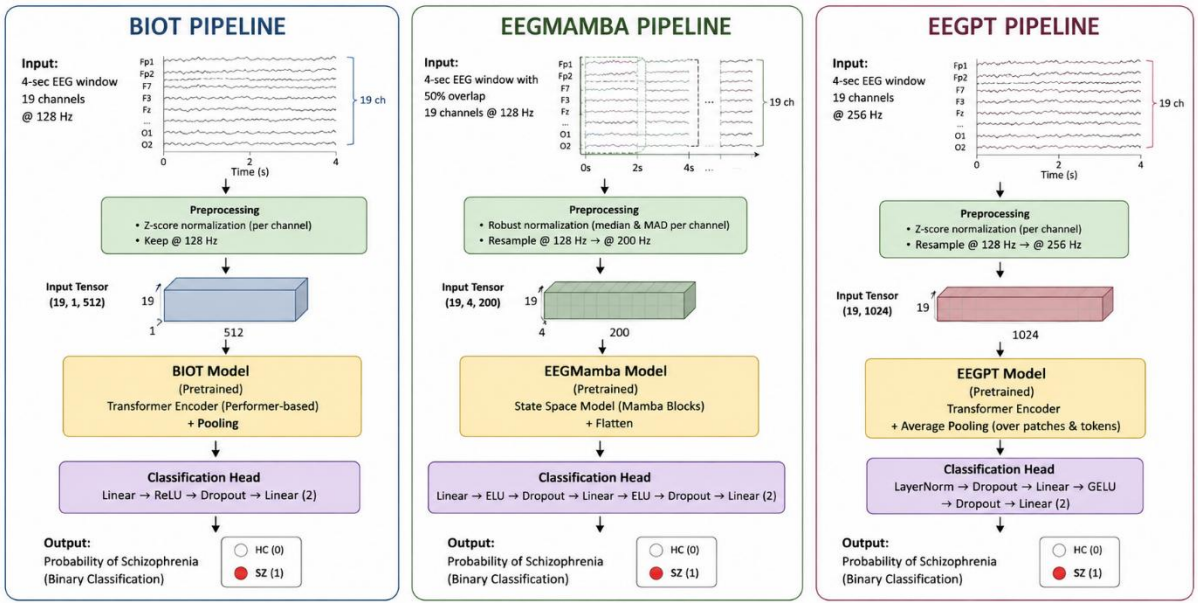
same preprocessing pipeline, subject-independent data partitioning, optimization strategy, and evaluation metrics were applied across all models.

The raw resting-state EEG recordings were first restricted to the initial 120 second of each recording. Signals were segmented into non-overlapping temporal windows, normalized independently within each window, and converted to the sampling frequency required by each pretrained model. Importantly, all preprocessing operations were performed independently for each subject without using information from other subjects, thereby preventing information leakage. Subject-independent stratified 10-fold cross-validation was employed, ensuring that EEG recordings from the same subject never appeared simultaneously in the training, validation, and testing subsets. Within each outer fold, a validation subset was further selected from the training subjects for model selection and early stopping. Final predictions for each subject were obtained by averaging the posterior probabilities of all windows belonging to that subject, followed by majority decision at the subject level.

Rather than training the models from scratch, transfer learning was adopted to exploit the rich EEG representations learned during large-scale self-supervised pretraining. Training was performed in two consecutive stages. During the first stage, the pretrained backbone remained frozen while only the newly added classification head was optimized. During the second stage, the classification head and the final portion of the pretrained encoder were jointly fine-tuned using a smaller learning rate, allowing adaptation to schizophrenia-specific EEG patterns while preserving the general representations learned during pretraining. Model parameters were optimized using the Adaptive Moment Estimation with Decoupled Weight Decay (AdamW) optimizer [22] with class-balanced loss, cosine annealing learning-rate scheduling, gradient clipping, mixed-precision training, and early stopping based on validation balanced accuracy. The best-performing model from each fold was subsequently evaluated on the corresponding unseen test subjects.

2.3.1. BIOT

BIOT (Biosignal Transformer) is a pretrained transformer architecture designed for biosignal representation learning such as EEG and ECG through spectral tokenization and linear-attention mechanisms [11,23]. Unlike EEGPT and EEGMamba, BIOT operates on short EEG segments transformed into frequency-domain representations using short-time Fourier analysis. Since the released pretrained model was trained using a predefined bipolar EEG montage, the recorded referential EEG channels were first converted to the corresponding bipolar derivations before resampling to the required sampling frequency. The pretrained encoder was then initialized using the official BIOT checkpoint and fine-tuned following the same two-stage transfer learning strategy employed for EEGPT and EEGMamba. Subject-level predictions were generated by averaging window-level probabilities, ensuring a consistent evaluation protocol across all three foundation models. Figure 2 (a) illustrates architecture of this model.



(a) (b) (c)

Figure 2. Architecture of (a) BIOT, (b) EEGMamba and (c) EEGPT models.

2.3.2. EEGMamba

EEGMamba is a state-space foundation model that replaces conventional self-attention with selective state-space modeling (Mamba) to efficiently capture long-range temporal dependencies while maintaining linear computational complexity with respect to sequence length [12,24]. The pretrained encoder provides generalized EEG representations that can be transferred to downstream neurological disorders. Similar to EEGPT, EEGMamba was initialized from publicly available pretrained weights, followed by two-stage supervised fine-tuning using the same optimization protocol and subject-independent evaluation framework. Figure 2 (b) illustrates architecture of this model.

2.3.3. EEGPT

EEGPT is a transformer-based EEG foundation model pretrained using large-scale self-supervised masked EEG representation learning [25]. The model learns contextual relationships among temporal EEG tokens through multi-head self-attention, enabling the extraction of robust and transferable neural representations. During downstream adaptation, a lightweight classification head was appended to the pretrained encoder, and the model was fine-tuned using the two-stage transfer learning strategy described above. Subject-level predictions were obtained by averaging window-level probabilities across all EEG segments belonging to each subject. Figure 2 (c) illustrates architecture of this model.

2.4. Simple Mamba

The Simple Mamba network was designed to classify schizophrenia from multichannel EEG while preserving temporal dependencies across relatively long EEG windows. Figure 3 illustrates the architecture of this simple Mamba model on SZ databases. Each input consisted of a 19-channel EEG segment, which was first projected into a 64-dimensional latent representation using a 1 ×1 convolutional stem. This operation transformed the input from (B,19,T) to (B, 64,T), where B denotes the batch size and T the number of temporal samples. After transposition to the sequence format required by the Mamba implementation, the latent sequence was processed by five consecutive residual Mamba blocks. Each block applied Layer Normalization followed by a Mamba state-space mixer configured with $d_{model} = 64$, $d_{state} = 16$, $d_{conv} = 4$, and an expansion factor of 2, together with dropout and a residual connection. This structure allows the model to capture both local temporal patterns and longer-range dependencies without relying on self-attention.

Following the Mamba blocks, a final Layer Normalization layer was applied, and the sequence representation was averaged across the temporal dimension to obtain a fixed 64-dimensional feature vector for each EEG window. This representation was then passed to a linear classification layer with two outputs corresponding to schizophrenia and healthy control classes. In the implemented configuration, EEG was segmented into 10-s windows with 50% overlap, corresponding to a 5-s hop, and each channel was standardized independently within each window using z-score normalization. Dropout rates of 0.10 and 0.20 were used after the convolutional stem and within the Mamba residual blocks, respectively, to reduce overfitting.

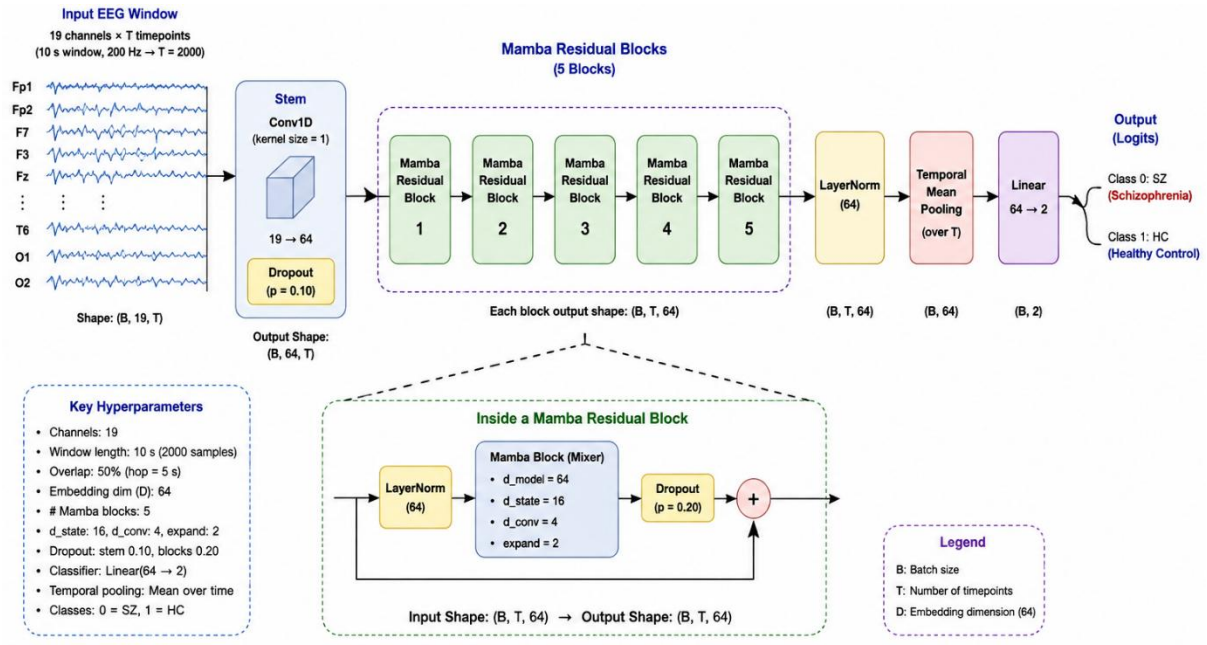


Figure 3. Architectural of simple Mamba for EEG signals

2.5. Model Training and Fine-Tuning

All foundation models were initialized using their publicly available pretrained weights and subsequently adapted to the schizophrenia classification task through supervised fine-tuning. To ensure a fair comparison, the same training strategy was adopted for EEGPT, EEGMamba, and BIOT. During the initial training stage, the pretrained backbone remained frozen while only the newly added classification head was optimized to adapt the generic EEG representations to the binary classification task. Subsequently, the classification head and the final layers of the pretrained backbone were jointly fine-tuned using a reduced learning rate, allowing the pretrained representations to be refined for schizophrenia classification while preserving the knowledge acquired during large-scale pretraining.

Model optimization was performed using the AdamW optimizer with a class-weighted cross-entropy loss function to compensate for the slight class imbalance. For a binary classification problem, the loss is defined as Eq (1) [22]:

$$L = -\frac{1}{N}\sum_{i=1}^{N} w_{y_i}\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right] \quad (1)$$

where Ndenotes the number of training samples, $y_i \in \{0,1\}$is the ground-truth class label, $\hat{y}_i$is the predicted probability for the positive class, and w_{y_i}represents the class-specific weighting factor. A cosine annealing learning-rate scheduler was employed throughout training, while gradient clipping and mixed-precision training were used to improve optimization stability and computational efficiency. Model selection was based on the highest balanced accuracy obtained on the validation set, and the corresponding checkpoint was retained for evaluation on the independent test subjects.

2.6. Explainable AI

To improve the interpretability of the pretrained EEG foundation models and identify the neurophysiological patterns contributing to schizophrenia classification, two complementary explainable artificial intelligence (XAI) techniques, namely Integrated Gradients (IG) [26] and channel occlusion analysis, were employed for EEGPT, EEGMamba, and BIOT. Both methods were applied to the best-performing model from each cross-validation fold, and explanations were generated exclusively for the independent test subjects to ensure unbiased interpretation. Window-level attribution maps were first computed for each EEG segment and subsequently averaged at the subject level, allowing every subject to contribute equally regardless of the number of analyzed windows. Finally, channel importance values were averaged across all subjects to obtain global channel importance rankings and scalp topographic visualizations. The EEGMamba implementation follows this workflow by computing Integrated Gradients, channel occlusion, subject-level aggregation, channel rankings, topographic maps, and agreement analysis between the two explanation methods.

Integrated Gradients estimates the contribution of each input feature by accumulating the gradients of the model output along a linear path from a baseline input to the original EEG signal. For an input sample x, baseline x', and model output $F(\cdot)$, the attribution assigned to the i-th feature is computed as Eq (2) [26]:

$$\mathrm{IG}_i(x) = (x_i - x_i')\int_0^1 \frac{\partial F(x'+\alpha(x-x'))}{\partial x_i}\, d\alpha \quad (2)$$

where x'denotes the baseline input, αrepresents the interpolation coefficient, and $F(x)$is the predicted probability of the schizophrenia class. The absolute Integrated Gradients quantify the overall importance of each EEG channel, whereas the signed Integrated Gradients preserve the direction of the contribution, indicating whether a feature increases or decreases the probability of schizophrenia.

To further validate the feature importance obtained from Integrated Gradients, channel occlusion analysis was performed. Each EEG channel was individually replaced by the baseline signal while keeping all remaining channels unchanged, and the resulting change in the predicted schizophrenia probability was measured. The importance of channel $c$was calculated as Eq (3):

$$O_c = \left| P_{\mathrm{SZ}}(x) - P_{\mathrm{SZ}}(x^{(\ -c)}) \right| \quad (3)$$

where $P_{\mathrm{SZ}}(x)$is the predicted schizophrenia probability for the original EEG sample and $P_{\mathrm{SZ}}(x^{(\ -c)})$denotes the prediction after occluding channel c. Larger values indicate that the corresponding channel has a greater influence on the model prediction.

For each subject, channel attribution scores obtained from all EEG windows were averaged to generate subject-level explanations as Eq (4):

$$\bar{A}_{s,c} = \frac{1}{N_s} \sum_{k=1}^{N_s} A_{s,c,k} \tag{4}$$

where $A_{s,c,k}$denotes the attribution score of channel $c$for the k-th EEG window of subject s, and $N_s$is the number of analyzed windows for that subject. Global channel importance was subsequently obtained by averaging the subject-level scores (Eq (5),

$$G_c = \frac{1}{S} \sum_{s=1}^{S} \bar{A}_{s,c} \tag{5}$$

where $S$is the total number of test subjects.

To assess the consistency between IG and channel occlusion, the normalized channel importance rankings produced by the two methods were compared using Spearman's rank correlation coefficient as Eq (6),

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{6}$$

where $d_i$is the difference between the two ranks assigned to channel i, and $n$denotes the total number of EEG channels. A high positive correlation indicates strong agreement between the gradient-based and perturbation-based explanation methods.

For BIOT, explanations were initially computed for the bipolar derivations used by the pretrained model. To facilitate comparison with EEGPT and EEGMamba, the bipolar attribution scores were subsequently projected onto their corresponding scalp electrodes, enabling visualization using standard EEG topographic maps while preserving the relative contribution of each bipolar derivation.

2.7. Evaluation Protocol

The performance of all models was evaluated using subject-independent stratified 10-fold cross-validation. Subjects were divided into ten mutually exclusive folds while preserving the class distribution across folds. In each iteration, one-fold was reserved for testing, whereas the remaining folds were used for training. A validation subset was selected exclusively from the training subjects for hyperparameter optimization, early stopping, and model selection. Importantly, all EEG recordings belonging to a given subject were assigned to only one subset within each fold, ensuring complete subject independence and preventing information leakage. After training, window-level posterior probabilities generated for each subject were averaged to obtain a single subject-level prediction. The predicted class corresponded to the highest averaged posterior probability, thereby reflecting the overall decision for the entire EEG recording rather than individual signal segments.

Model performance was evaluated at the subject level using Accuracy (ACC), Sensitivity (SEN), Specificity (SPE), Precision (PRE), F1-score (F1) and the Area Under the Receiver Operating Characteristic Curve (AUC). These metrics were computed from the numbers of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). For each metric, the mean and standard deviation across the ten subject-independent cross-validation folds were reported to assess both predictive performance and model robustness.

3. Results

To ensure a fair comparison among EEGPT, EEGMamba, BIOT and simple Mamba, hyperparameters were optimized using an identical grid-search strategy whenever compatible with the corresponding model architecture. Hyperparameter optimization was performed exclusively on the training and validation subjects within each cross-validation fold, while the independent test subjects remained completely unseen throughout model development. For each candidate configuration, model performance was evaluated based on subject-level validation balanced accuracy, and the configuration achieving the highest validation performance was selected for subsequent evaluation on the corresponding test fold.

All three models were trained using a cosine-annealing learning-rate scheduler and weighted cross-entropy loss to account for class imbalance. Mixed-precision training was employed to improve computational efficiency. Model optimization was monitored using subject-level validation balanced accuracy, which was also used as the criterion for early stopping and selection of the best-performing model. The experiments were conducted using subject-based stratified 10-fold cross-validation, with all EEG windows from a given participant restricted to the same data partition to prevent subject-level data leakage. Following model inference, the predicted probabilities of all EEG windows belonging to each participant were averaged to obtain a single subject-level probability. This

probability was subsequently thresholded to determine the final subject-level classification as healthy control or schizophrenia.

The grid search explored EEG window durations ranging from 4 to 15 s, input recording durations ranging from the initial 120 s to the complete EEG recording, and window overlap ratios ranging from 0% to 50%. The search additionally evaluated different initial learning rates for training the classification head, fine-tuning learning rates for the pretrained backbone, and weight-decay values. Model-specific architectural parameters, including the encoder structure, embedding dimension, number of layers, and attention or state-space configuration, were retained according to the official pretrained implementations. The final model-specific hyperparameters selected through this procedure are summarized in Table 2.

Table 2. Final hyperparameter configuration.

Hyperparameter	**EEGPT**	**EEGMamba**	**BIOT**
Pretrained checkpoint	Official EEGPT	Official EEGMamba	Official BIOT (PREST)
Backbone	Transformer	Mamba State Space Model	Transformer
Input duration	240 s	240 s	240 s
Window length	4 s	10 s (ASEEG), 6 s (Warsaw)	6 s (ASEEG), 10 s (Warsaw)
Window overlap	None	25% (ASEEG), 50% (Warsaw),	50% (ASEEG), 0 (Warsaw)
Batch size	4	4	4
Optimizer	AdamW	AdamW	AdamW
Initial learning rate (head)	1×10^{-3}	1×10^{-3}	1×10^{-3}
Fine-tuning learning rate	1×10^{-4}	1×10^{-4}	1×10^{-4}
Weight decay	1×10^{-2}	1×10^{-2}	1×10^{-2}
Gradient clipping	1.0	1.0	1.0

The classification performance of the evaluated EEG foundation models was assessed separately on the ASEEG and Warsaw datasets. Accuracy, precision, sensitivity, specificity, F1-score, and ROC-AUC were used as performance measures. In addition, IG and channel occlusion were applied to characterize the contribution of individual EEG channels to model predictions.

3.1. Classification results on ASEEG data

Table 3 presents the classification performance of EEGPT, EEGMamba, Simple Mamba, and BIOT on the ASEEG dataset. Among the evaluated models, EEGMamba achieved the highest overall performance, with an accuracy of 79.18 ± 8.00%, precision of 81.72 ± 10.09%, sensitivity of 82.33 ± 12.54%, F1-score of 0.79 ± 0.10, and ROC-AUC of 0.83 ± 0.09. The highest specificity was obtained by EEGMamba (76.00 ± 20.00%), while BIOT showed a comparable specificity of 74.00 ± 21.18%.

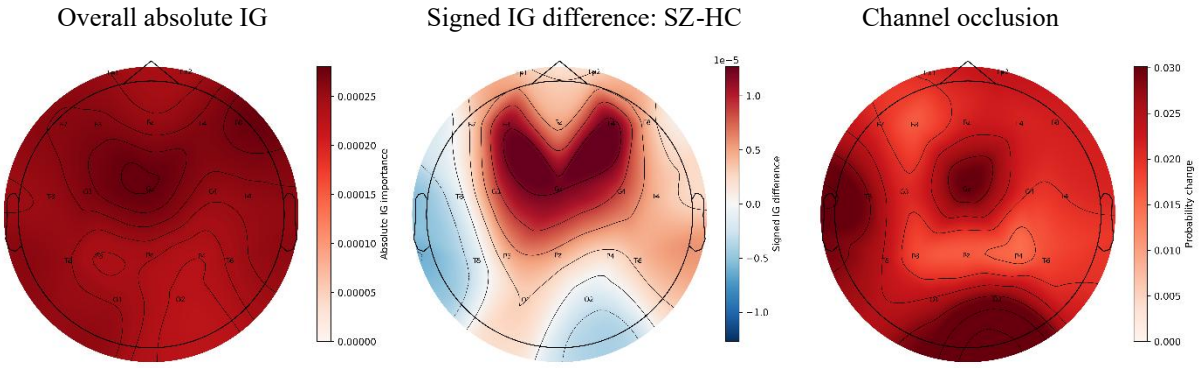
Table 3. Classification performance of the evaluated EEG foundation models for schizophrenia detection on the ASEEG dataset. Results are reported as mean ± standard deviation across the subject-based cross-validation folds. Accuracy, precision, sensitivity, and specificity are expressed as percentages, while F1-score and ROC-AUC are reported on a scale from 0 to 1. The input window configuration used for each model is indicated in parentheses.

Model	**Acc (%)**	**Pre (%)**	**Sen (%)**	**Spec (%)**	**F1-score**	**ROC-AUC**
EEGPT (4s)	74.25±10.54	73.58±15.45	76.47±15.08	72.00±21.05	0.75±0.12	0.76±0.20
EEGMamba (10s, 25% Overlap)	**79.18±8.00**	**81.72±10.09**	**82.33±12.54**	**76.00±20.00**	**0.79±0.10**	**0.83±0.09**
Simple Mamba	69.31±6.58	69.94±8.32	66.82±8.26	71.77±9.15	0.68±0.17	0.68±0.14
BIOT (6 s, 50% overlap)	76.27±14.19	74.67±11.25	78.33±17.58	74.00±20.18	0.76±0.19	0.77±0.21

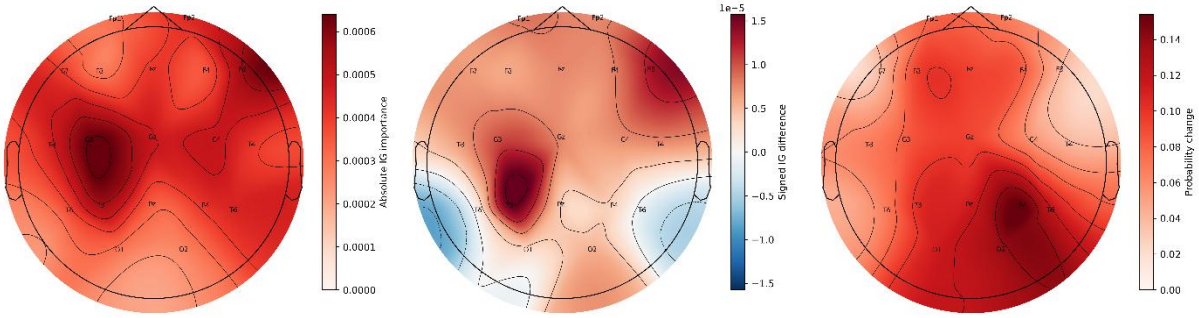
3.2. XAI results on ASEEG data

The spatial distributions of channel importance obtained using Integrated Gradients (IG) and channel occlusion are presented in Figure 4 (a-d) for EEGMamba, EEGPT, BIOT and simple Mamba. For EEGMamba (figure 4

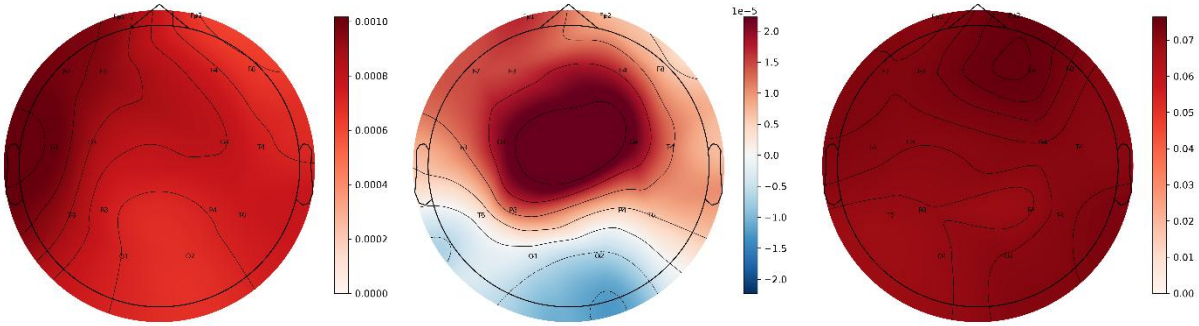
(a)), the absolute IG map showed broadly distributed contributions, with relatively greater attribution over the central and frontal regions. The signed IG map showed the strongest positive SZ–HC differences over the frontal-central region, whereas negative differences were mainly observed over the left lateral and posterior regions. The occlusion map showed greater prediction changes over central and posterior regions. For EEGPT (figure 4 (b)), absolute IG attribution was most prominent over the left-central region. The signed IG map showed a strong positive difference over the left-central area and another positive region over the right frontal area, while negative differences appeared over the left frontal/temporal and posterior regions. Its occlusion map showed the greatest importance over the right posterior region. For BIOT (figure 4 (c)), absolute IG attribution was comparatively uniform across the scalp, whereas the signed IG map showed a pronounced positive difference over the central and frontal-central regions and negative differences over the posterior scalp. The BIOT occlusion map was also relatively spatially uniform. For Simple Mamba (figure 4 (d)), the absolute IG map showed relatively widespread channel contributions, with higher attribution particularly over the lateral frontal and central regions. The signed IG difference map revealed a heterogeneous spatial pattern, with positive SZ–HC differences primarily over the anterior midline, right-central, and left posterior regions, whereas negative differences were more evident over the left frontal and right posterior areas. The channel-occlusion map showed the greatest probability changes over the central-to-right-central region, while comparatively smaller effects were observed across most peripheral electrodes. Overall, the IG and occlusion analyses suggest that Simple Mamba relies on spatially distributed EEG information, with particularly relevant contributions from central and frontal regions for schizophrenia classification.



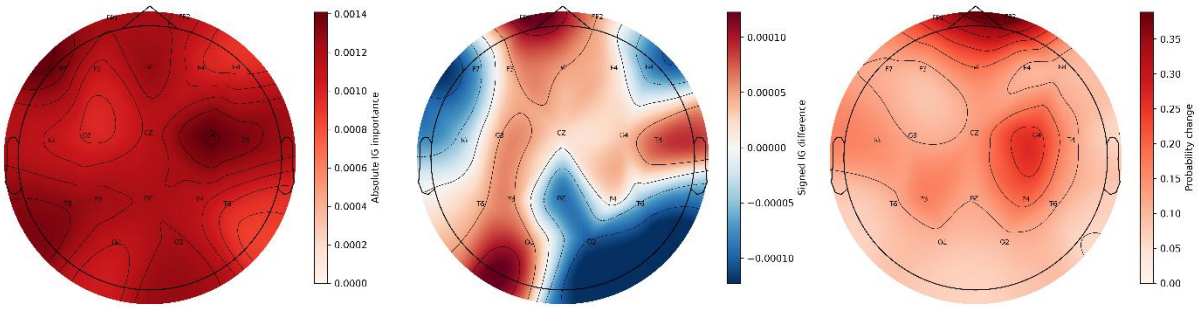
(a) EEGMamba



(b) EEGPT



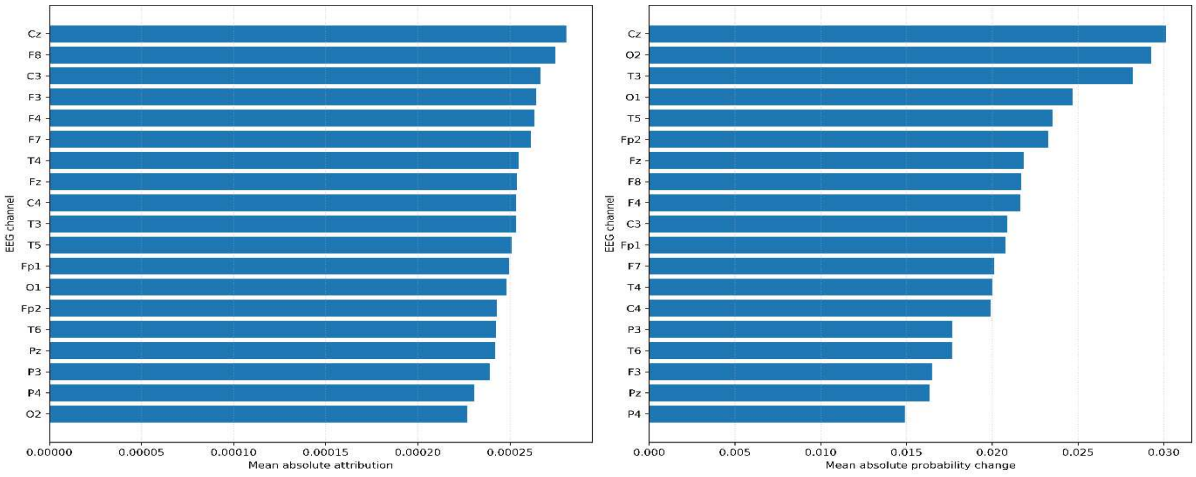
(c) BIOT



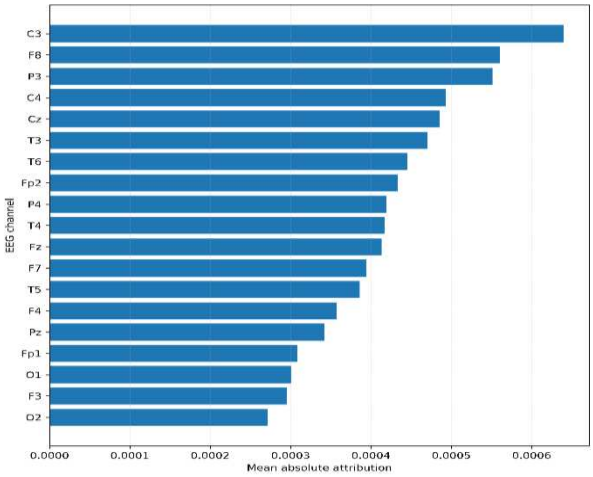
(d) simple Mamba

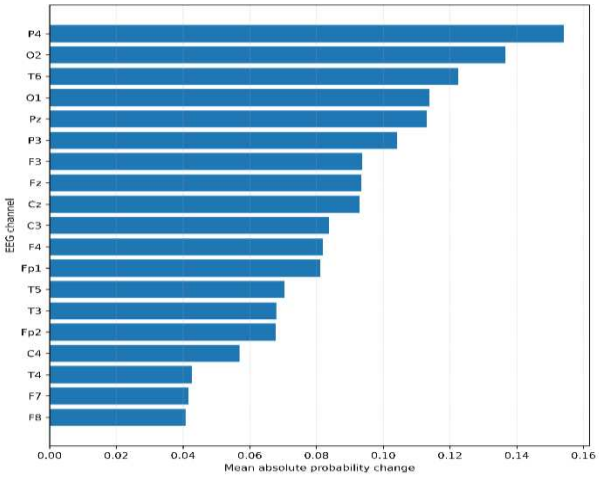
Figure 4. Spatial distribution of EEG channel importance for (a) EEGMamba, (b) EEGPT, (c) BIOT and (d) simple Mamba on the ASEEG dataset. Each model is represented by three scalp maps, from left to right: overall absolute Integrated Gradients attribution averaged across the out-of-fold test data; signed Integrated Gradients difference between the SZ and HC groups, where positive/red values indicate a greater contribution toward the schizophrenia output and negative/blue values indicate a greater HC-associated or schizophrenia-opposing contribution; and channel-occlusion importance, quantified as the mean absolute change in predicted schizophrenia probability after masking each EEG channel.

The quantitative channel rankings derived from IG and occlusion are shown in Figure 5. For EEGMamba (figure 5 (a & b)), the IG ranking placed Cz as the highest-ranked channel, while C4 showed the largest occlusion importance. For EEGPT (figure 5 (c & d)), C3 had the highest IG attribution, whereas P4 showed the greatest change under channel occlusion. For BIOT (figure 5 (e & f)), the highest IG importance was observed for Fp2–F8, while the occlusion ranking was headed by Fp2–F8, T8–P8, and P4–O2. For Simple Mamba (Figure 5(g & h)), the IG-based ranking identified C4 as the most influential channel, followed by T4, F7, T5, and Fz, indicating relatively strong contributions from central, temporal, and frontal electrodes. In contrast, the channel-occlusion analysis showed a more distinct ranking, with Fp2 producing the largest decrease in prediction probability when occluded, followed by Fp1 and C4. Thus, while IG indicated a relatively distributed pattern of channel attribution, the occlusion analysis highlighted the importance of the prefrontal channels, particularly Fp2 and Fp1, together with C4, for Simple Mamba's classification decisions. These rankings provide channel-wise ordering obtained independently from the gradient- and perturbation-based analyses.

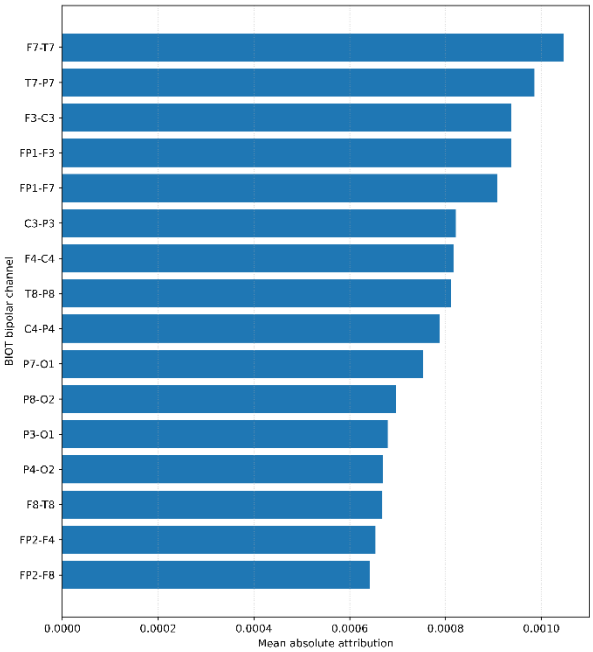


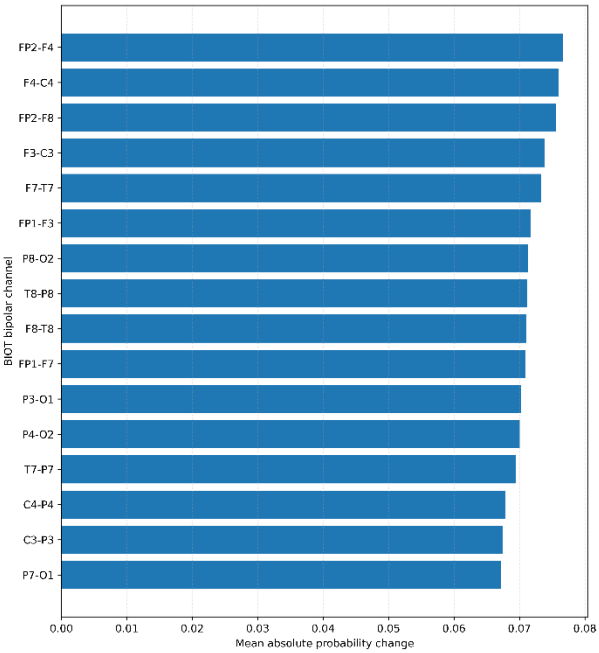
(a) EEGMamba (b)



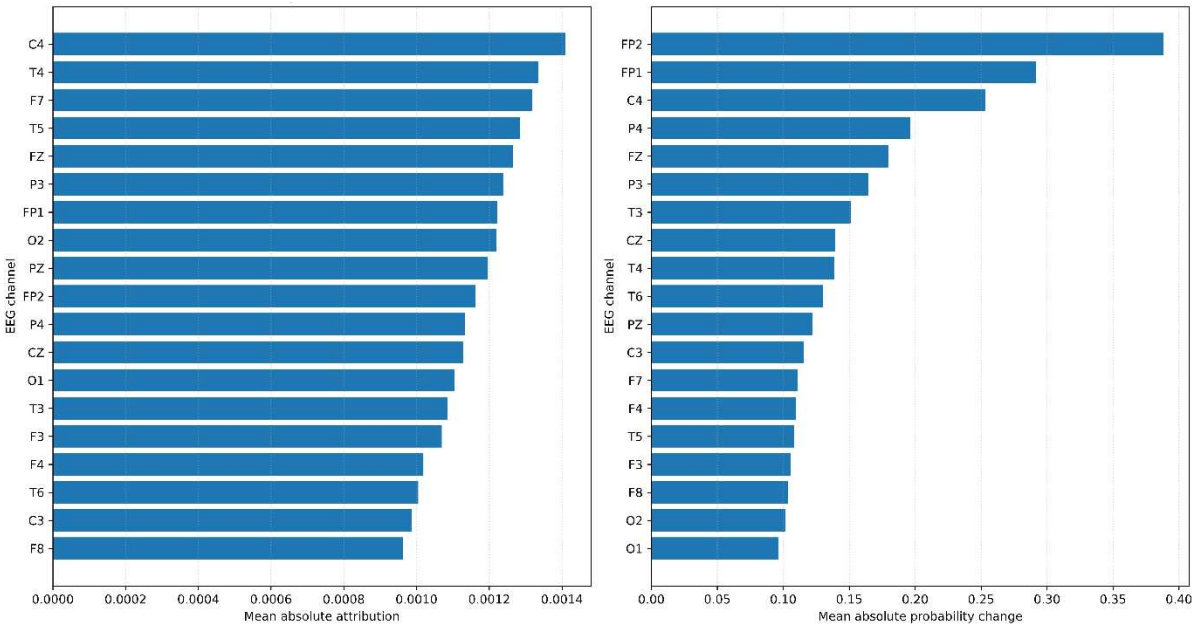


(c) EEGPT (d)





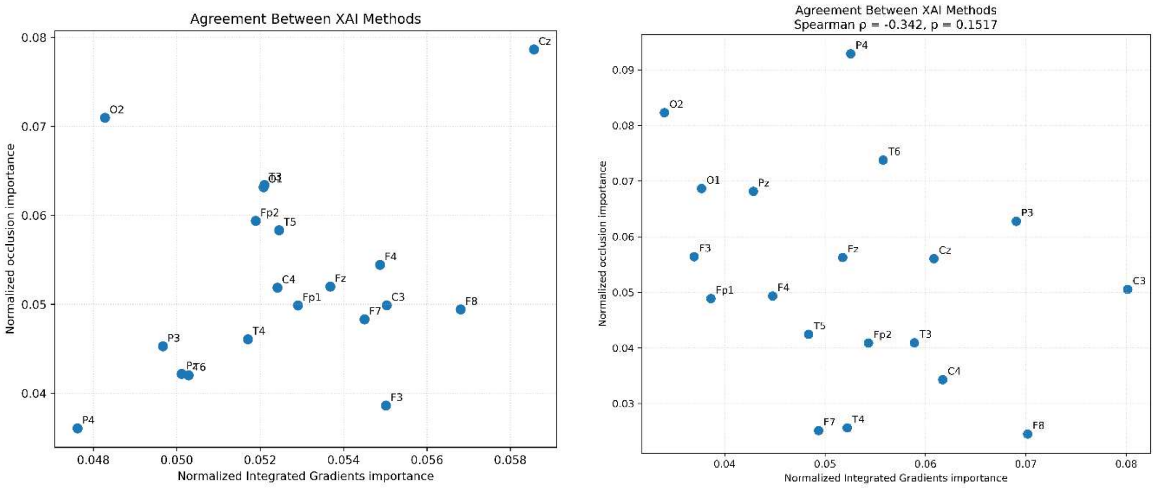
(e) BIOT (f)



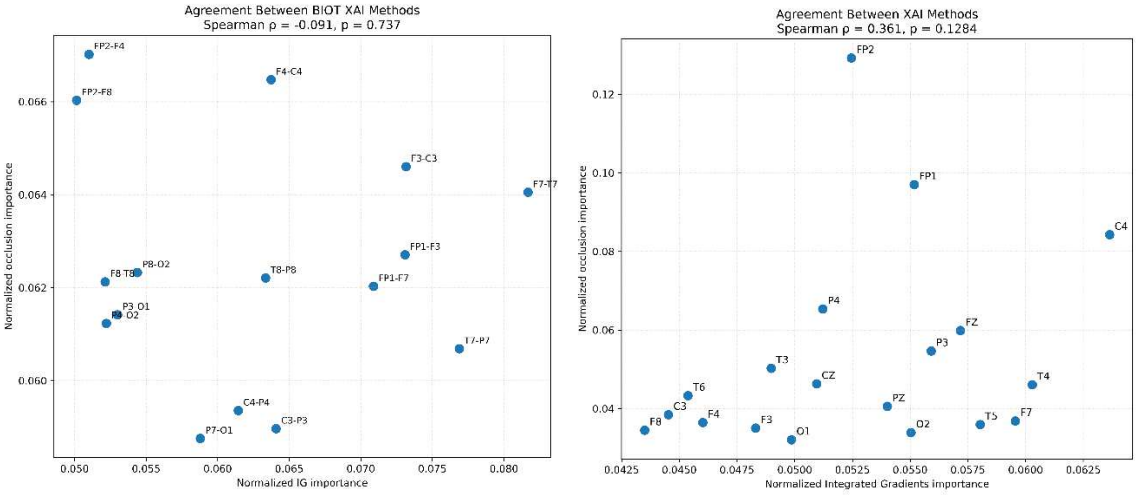
(g) simple Mamba (h)

Figure 5. Ranking of EEG channel importance for the ASEEG dataset based on IG and channel occlusion. Results are shown for EEGMamba [(a) IG and (b) channel occlusion], EEGPT [(c) IG and (d) channel occlusion], BIOT [(e) IG and (f) channel occlusion] and simple Mamba [(g) IG and (h) channel occlusion]. Channels are ordered according to their mean importance across the evaluated test data. IG represents mean absolute attribution, while channel occlusion represents the mean absolute change in predicted schizophrenia probability after masking the corresponding channel.

Figure 6(a–d) compares the normalized IG and occlusion importance scores to assess the agreement between the two XAI methods. EEGMamba showed the clearest positive relationship between the two measures among the evaluated models (Figure 6(a)). For EEGPT, the Spearman correlation was ρ = 0.342 (p = 0.151) (Figure 6(b)), while BIOT showed a weaker association (ρ = 0.091, p = 0.723) (Figure 6(c)). For Simple Mamba (Figure 6(d)), a moderate positive correlation was observed between IG and occlusion importance (ρ=0.361, p=0.1284). Although channels such as Fp2 and Fp1 exhibited relatively high occlusion importance, their IG importance was less pronounced, indicating that the two explanation methods captured partially different aspects of channel relevance. Overall, the correspondence between IG- and occlusion-derived channel importance varied across the four models, and the correlations were not statistically significant.



(a) EEGMamba (b) EEGPT



(c) BIOT (d) simple Mamba

Figure 6. Agreement between Integrated Gradients and channel-occlusion estimates of EEG channel importance on the ASEEG dataset for (a) EEGMamba, (b) EEGPT, (c) BIOT and (d) simple Mamba. Each point represents an individual EEG channel, with normalized IG importance on the horizontal axis and normalized channel-occlusion importance on the vertical axis. Spearman's rank correlation coefficient and its corresponding *p*-value are displayed for each model.

3.3. Classification results on Warsaw database

Table 4 summarizes the classification results obtained on the Warsaw dataset. EEGMamba achieved the highest classification accuracy (78.33 ± 20.90%), sensitivity (80.00 ± 14.21%), precision (68.33±20.04) and ROC-AUC was 0.70 among the evaluated models. For specificity, EEGPT and BIOT both reached 74.28% mean specificity, with the reported standard deviations shown in Table 4.

Table 4. Classification performance of the evaluated EEG foundation models for schizophrenia detection on the Warsaw dataset. Results are reported as mean ± standard deviation across the subject-based cross-validation folds. Accuracy, precision, sensitivity, and specificity are expressed as percentages, while F1-score and ROC-AUC are reported on a scale from 0 to 1. The input window length and overlap configuration, where applicable, are provided in parentheses for each model.

Model	ACC (%)	PRE (%)	SEN (%)	SPE (%)	F1-score	ROC-AUC
EEGPT (4 s)	68.57±25.43	69.54±26.65	67.85±15.56	74.28±19.34	0.68±0.65	0.65±0.26
Simple Mamba	71.42±12.49	66.66±17.26	71.42±19.33	71.42±39.33	0.66±0.34	0.67±0.21
EEGMamba (6, 50% overlap)	**78.33±20.90**	**68.33±20.04**	**80.00±14.21**	**65.00±14.75**	**72.66±0.23**	**0.70±0.13**
BIOT (10 s, 0% overlap)	67.14±18.89	55.23±36.91	60.00±40.82	74.28±37.79	55.23±0.23	0.67±0.24

3.4. XAI results of Warsaw database

Figure 7 (a-d) presents the spatial channel-importance patterns for the Warsaw dataset. For EEGMamba (figure 7 (a)), absolute IG attribution was broadly distributed, while the signed IG map showed positive SZ–HC differences predominantly over the central and posterior-central regions and negative differences over the frontal and lateral regions. The occlusion map was comparatively distributed across the scalp. For EEGPT (figure 7 (b)), the absolute IG map showed the greatest attribution over the left-central region. In contrast to ASEEG, its signed IG map was predominantly negative across the scalp, with a small positive region around the central area. The EEGPT (figure 7 (b)) occlusion map showed the greatest prediction sensitivity over the posterior-right region. For BIOT (figure 7 (c)), absolute IG attribution was broadly distributed, while the signed IG map showed a prominent positive region over the central-right scalp and negative differences over the frontal and left lateral regions. The occlusion map showed greater importance toward the right posterior scalp. For Simple Mamba (Figure 7(d)), the absolute IG map showed relatively distributed attribution, with greater importance over the lateral regions, particularly toward the right-central and left posterior scalp. The signed IG difference map revealed a pronounced positive

SZ–HC difference over the central region, accompanied by additional positive differences over the left lateral and right posterior regions, whereas negative differences were mainly observed over the frontal and posterior-central areas. The channel-occlusion map showed its strongest prediction sensitivity over the right-central region, with comparatively lower effects across the remaining scalp areas.

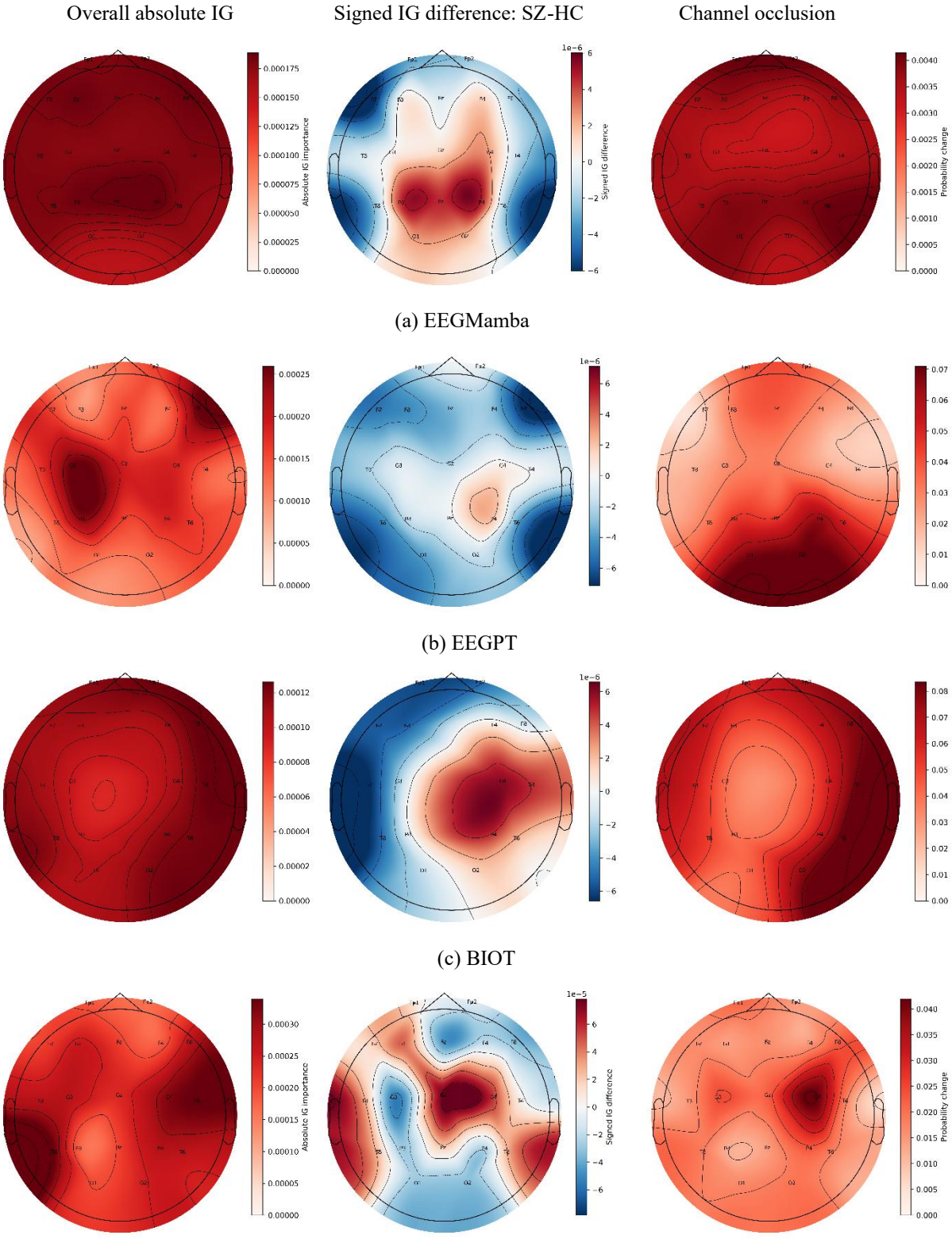
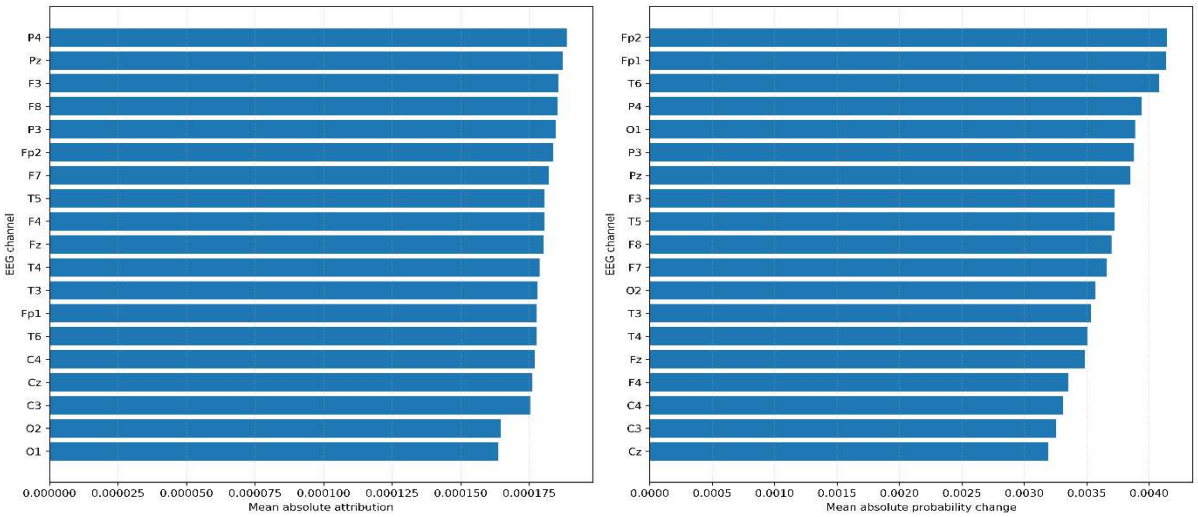


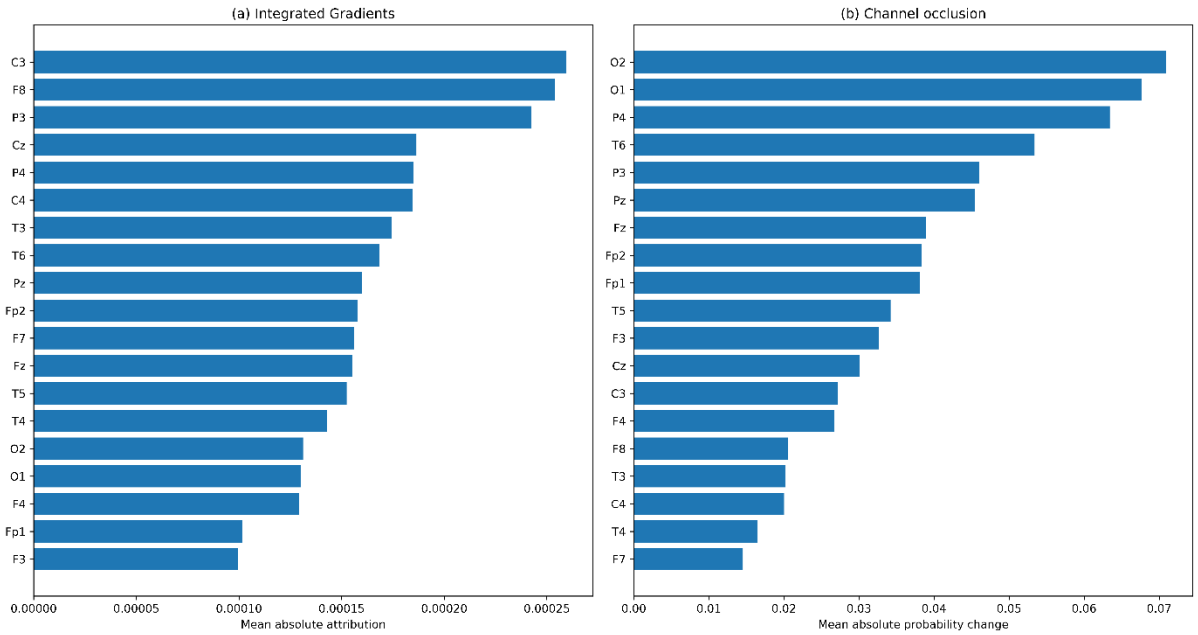
Figure 7. Spatial distribution of EEG channel importance for (a) EEGMamba, (b) EEGPT, (c) BIOT and (d) simple Mamba on the Warsaw dataset. For each model, the three scalp maps from left to right show overall

absolute IG attribution, signed IG difference between SZ and HC subjects, and channel-occlusion importance. Absolute IG represents the magnitude of attribution averaged across the evaluated test data. In the signed difference maps, positive/red values represent relatively greater schizophrenia-supporting contributions and negative/blue values represent relatively greater HC-associated or schizophrenia-opposing contributions. Channel-occlusion importance represents the mean absolute change in schizophrenia probability after masking each EEG channel.

The channel-wise rankings for the Warsaw dataset are shown in Figure 8 (a-h). For EEGMamba, P4 had the highest IG attribution, while Fp2 ranked highest according to channel occlusion (figure 8 (a & b)). For EEGPT, C3 was the highest-ranked channel according to IG, whereas P4 showed the greatest occlusion importance (figure 8 (c & d)). For BIOT, the IG ranking was headed by Fp2–F8, followed closely by T8–P8 and P4–O2, while T8–P8 showed the highest occlusion importance (figure 8 (e & f)). For Simple Mamba, T4 showed the highest IG attribution, followed by T5 and C4 (Figure 8(g)), indicating prominent contributions from temporal and central channels. In contrast, the occlusion analysis identified C4 as the most influential channel by a clear margin, followed by C3 and P4 (Figure 8(h)), highlighting the sensitivity of Simple Mamba predictions to central EEG activity.



(a) EEGMamba (b)



(c) EEGPT (d)

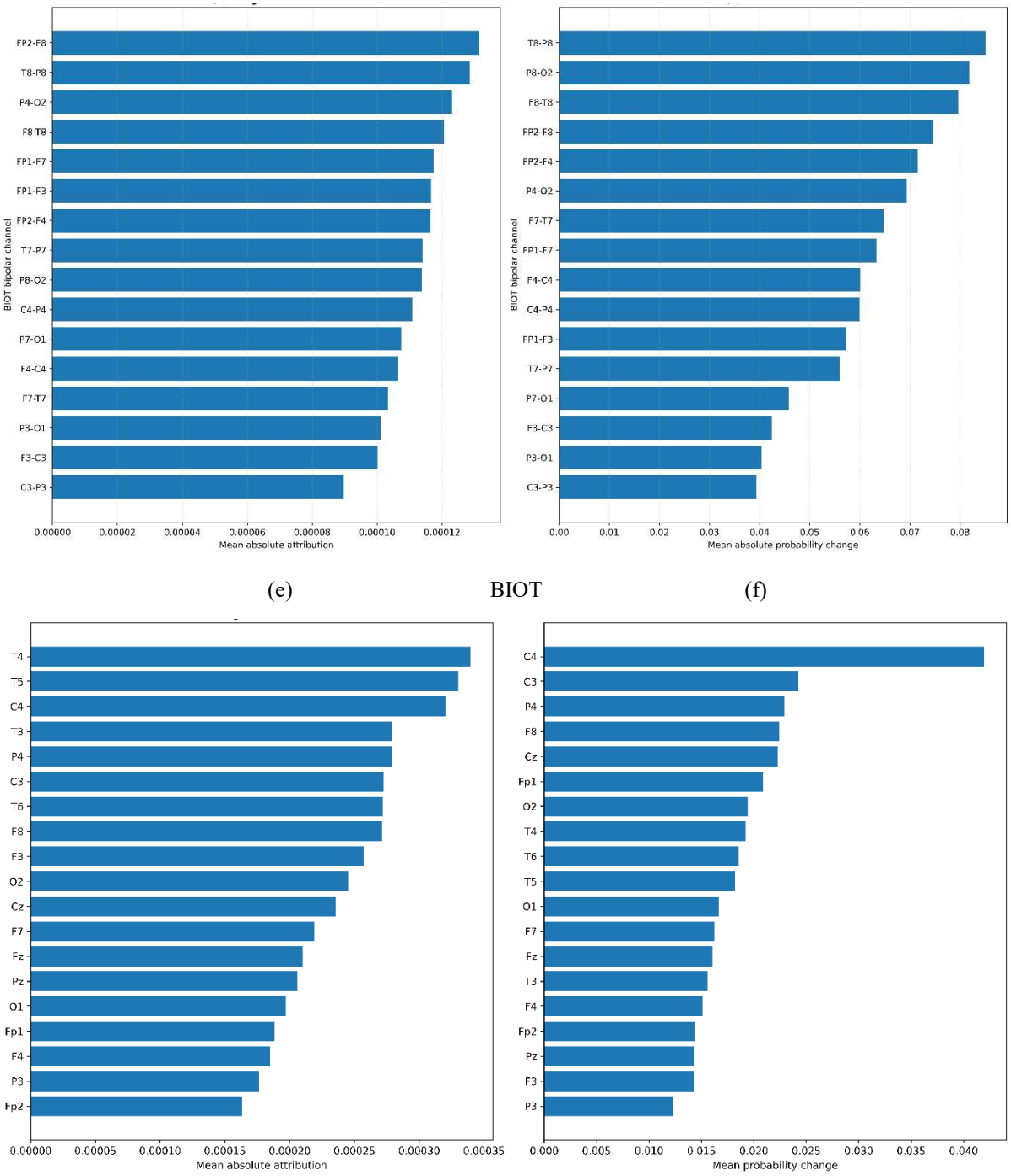
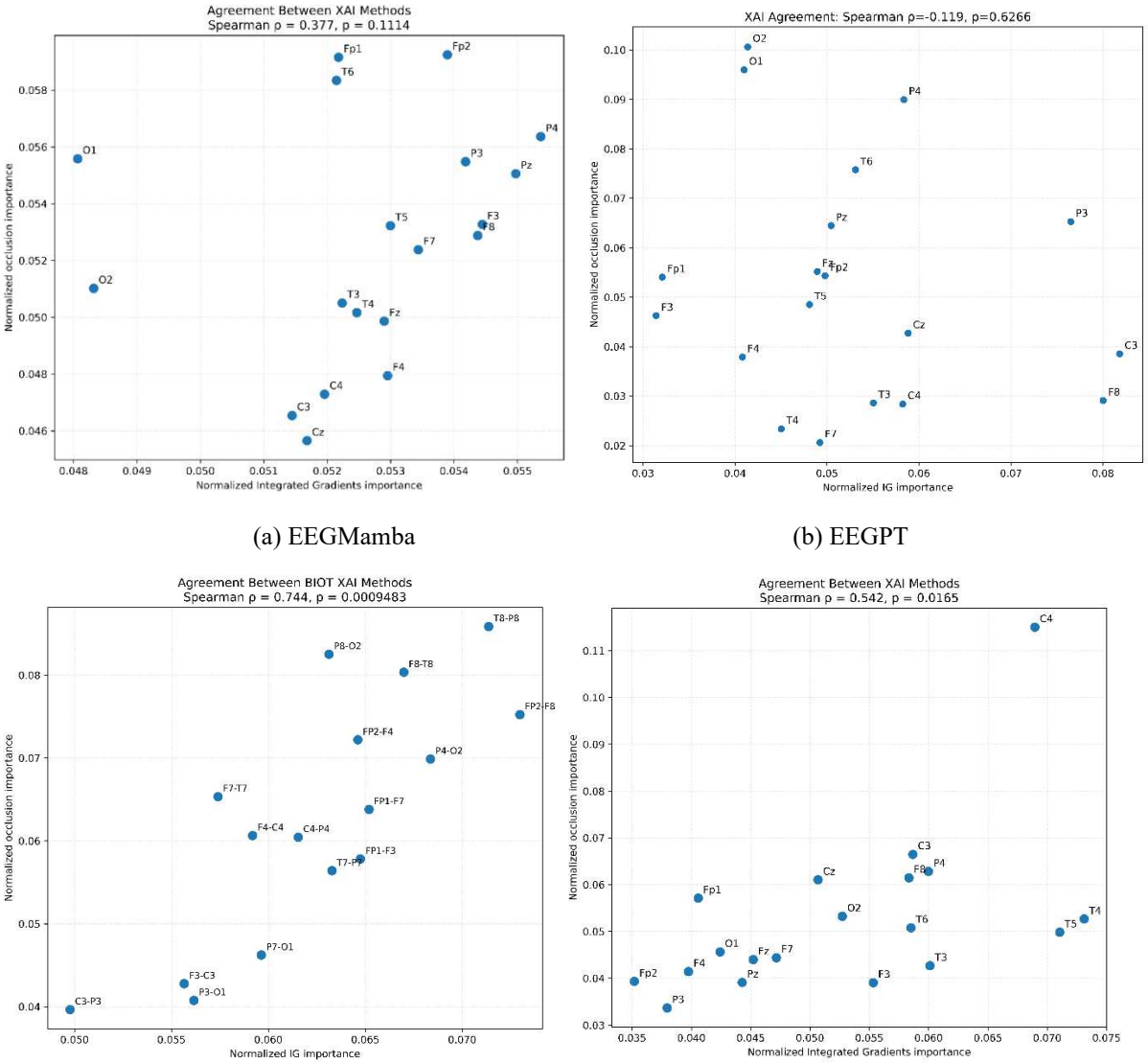


Figure 8. Ranking of EEG channel importance for the Warsaw dataset based on Integrated Gradients and channel occlusion. Results are shown for EEGMamba [(a) IG and (b) channel occlusion], EEGPT [(c) IG and (d) channel occlusion], BIOT [(e) IG and (f) channel occlusion] and simple Mamba [(g) IG and (h) channel occlusion]. Channels are ordered according to their mean importance across the evaluated test data. IG represents mean absolute attribution, while channel occlusion represents the mean absolute change in predicted schizophrenia probability after masking the corresponding channel.

Figure 9 (a-d) shows the agreement between the two XAI approaches on the Warsaw dataset. The reported Spearman correlations were $\rho = 0.377$ ($p = 0.1114$) for EEGMamba (Figure 9 (a)), $\rho = 0.119$ ($p = 0.6266$) for EEGPT (Figure 9 (b)), $\rho = 0.744$ ($p = 0.0009483$) for BIOT (Figure 9 (c)) and $\rho = 0.542$ ($p = 0.0165$) for simple Mamba (Figure 9 (d)). Among the three models, BIOT therefore exhibited the highest correlation between the IG and channel-occlusion rankings.



(a) EEGMamba (b) EEGPT

(c) BIOT (d) simple Mamba

Figure 9. Agreement between IG and channel-occlusion estimates of EEG channel importance on the Warsaw dataset for (a) EEGMamba, (b) EEGPT, (c) BIOT and (d) simple Mamba. Each point corresponds to one EEG channel. The horizontal axis represents normalized IG importance and the vertical axis represents normalized channel-occlusion importance. Spearman's rank correlation coefficient and the corresponding *p*-value quantify the association between channel rankings produced by the two XAI approaches.

4. Discussion

This study evaluated the classification performance and interpretability of three pretrained EEG foundation models—EEGPT, EEGMamba, and BIOT—for schizophrenia detection across the ASEEG and Warsaw datasets, with Simple Mamba included as a non-foundation Mamba-based comparison. Performance varied according to architecture and dataset, while EEGMamba consistently achieved the highest overall accuracy. On ASEEG, EEGMamba obtained the highest accuracy (79.18 ± 8.00%), precision (81.72 ± 10.09%), sensitivity (82.33 ± 12.54%), F1-score (0.79 ± 0.10), and ROC-AUC (0.83 ± 0.09), with a specificity of 76.00 ± 20.00%. BIOT achieved the second-highest accuracy (76.27 ± 14.19%), with sensitivity of 78.33 ± 17.58% and ROC-AUC of 0.77 ± 0.21, followed by EEGPT with 74.25 ± 10.54% accuracy. Simple Mamba showed the lowest ASEEG accuracy (69.31 ± 6.58%) and ROC-AUC (0.68 ± 0.17). These findings indicate that pretrained EEG representations contain useful discriminatory information for schizophrenia detection, with EEGMamba being particularly effective on ASEEG.

A similar pattern was observed on Warsaw, where EEGMamba again achieved the highest accuracy (78.33 ± 20.90%) and sensitivity (80.00 ± 14.21%), with a mean ROC-AUC of 0.70 ± 0.83. However, EEGPT achieved

the highest precision (69.54 ± 26.65%), while EEGPT and BIOT both reached the highest specificity (74.28%), compared with 65.00% for EEGMamba. Simple Mamba achieved the second-highest accuracy (71.42 ± 12.49%) and sensitivity (71.42 ± 19.33%), whereas EEGPT and BIOT obtained accuracies of 68.57 ± 25.43% and 67.14 ± 18.89%, respectively. Thus, although EEGMamba maintained the highest mean accuracy and sensitivity across both datasets, the ranking of the other models changed, indicating that no representation-learning strategy was uniformly superior across all metrics.

The cross-dataset results highlight the sensitivity of EEG foundation models to downstream clinical datasets. Although EEGMamba achieved similar mean accuracies on ASEEG and Warsaw (79.18% and 78.33%), its accuracy variability increased from 8.00% to 20.90%. Similar variability occurred for several EEGPT and BIOT metrics. The Warsaw findings therefore indicate lower stability across subject-based folds rather than simply reduced performance. Differences in cohort composition, acquisition systems, recording protocols, signal characteristics, preprocessing, sample size, or correspondence with pretraining EEG distributions may contribute to this variability. Although their individual effects cannot be isolated here, these findings emphasize the importance of evaluating EEG foundation models on independent datasets.

Comparison between EEGMamba and Simple Mamba provides further insight into EEG-specific pretraining. On ASEEG, EEGMamba exceeded Simple Mamba by approximately 9.9 percentage points in accuracy (79.18% vs. 69.31%) and also achieved higher sensitivity, precision, F1-score, and ROC-AUC. On Warsaw, EEGMamba similarly achieved higher accuracy (78.33% vs. 71.42%) and sensitivity (80.00% vs. 71.42%). Nevertheless, the gap was smaller on Warsaw, where Simple Mamba outperformed EEGPT and BIOT in mean accuracy. These results suggest that Mamba sequence modeling can capture informative temporal EEG dependencies, while EEG-specific pretraining may provide an additional advantage. However, this comparison is not a controlled ablation of pretraining because the models also differ in architecture, representation-learning strategy, and optimization characteristics. Therefore, the differences cannot be attributed exclusively to pretraining.

The XAI analyses provided insight into the spatial EEG information used for classification. On ASEEG, EEGMamba exhibited broadly distributed absolute IG attribution with greater central and frontal contributions, EEGPT showed a more pronounced left-central pattern, BIOT displayed a relatively uniform distribution, and Simple Mamba emphasized lateral-frontal and central regions. Signed IG maps also demonstrated differences in attribution direction between SZ and HC subjects, particularly over frontal-central and central regions. Channel occlusion complemented these findings by identifying electrodes whose removal produced the largest changes in predicted schizophrenia probability.

The prominence of frontal and central regions is broadly consistent with previous electrophysiological findings. Winterer et al. [27] developed a schizophrenia classifier using frequency-domain and evoked-potential measures from frontal and central electrodes and reported frontally pronounced delta abnormalities. Electrophysiological abnormalities have also been investigated extensively as potential diagnostic and pathophysiological markers of schizophrenia [28]. Meta-analytic evidence indicates altered resting-state oscillatory activity, including regional effects involving frontal and temporal cortices, although these effects are heterogeneous [29]. Thus, the frontal-central emphasis in our XAI results is compatible with previously reported electrophysiological alterations but should not be interpreted as direct localization of pathological activity.

Quantitative channel rankings provided further correspondence with previous studies. On ASEEG, EEGMamba ranked Cz highest by IG and C4 by occlusion. EEGPT identified C3 as the highest IG channel and P4 by occlusion, while Simple Mamba ranked C4 highest by IG and highlighted Fp2, Fp1, and C4 through occlusion. Ellis et al. [30], using the Warsaw dataset, identified T3, C4, F3, P3, Fz, Fp2, F7, and Cz as significant in more than half of their evaluated models. P3 showed the highest mean relevance, followed by Fz and C4, while C4 was significant across essentially all training-validation folds and most test folds. The recurrence of C4, Cz, Fp2, Fz, and related frontal-central electrodes provides external support for the spatial plausibility of our explanations.

On Warsaw, EEGMamba identified P4 as the highest-ranked IG channel and Fp2 as most influential by occlusion. EEGPT ranked C3 highest by IG and P4 by occlusion, whereas Simple Mamba highlighted T4, T5, and C4 by IG and C4, C3, and P4 by occlusion. These findings partly overlap with the reproducibly relevant C4, P3, Fz, Fp2, Cz, T3, F3, and F7 electrodes previously identified on Warsaw [30]. Although rankings differ between models, both analyses emphasize distributed central, frontal/frontopolar, temporal, and parietal regions rather than a single scalp location.

The involvement of parietal electrodes is also compatible with evidence that schizophrenia involves distributed rather than exclusively frontal dysfunction. Reviews have implicated frontal-temporal and occipital-parietal systems in altered information integration [31], while meta-analytic neuroimaging studies have reported frontal and parietal abnormalities, particularly during cognitive processes such as working memory [32]. The importance

of P3/P4 may therefore reflect distributed fronto-central-parietal information. However, scalp electrodes do not provide one-to-one localization of cortical generators, and anatomical interpretation should remain cautious.

Differences between IG and occlusion rankings are not necessarily contradictory. IG estimates feature contributions relative to a baseline through integrated gradients, whereas occlusion measures the prediction change produced by removing an electrode. A channel may therefore have high IG attribution but a smaller occlusion effect when its information is redundant with neighboring electrodes. Conversely, moderate gradient attribution can accompany a larger occlusion effect when the model relies on information uniquely represented by that channel. This distinction is especially relevant for multichannel EEG because volume conduction and spatial correlations create substantial redundancy.

The agreement analyses support this interpretation. On ASEEG, normalized IG and occlusion importance showed non-significant relationships for EEGPT (ρ=0.342, p=0.151), BIOT (ρ=0.091, p=0.723), and Simple Mamba (ρ=0.361, p=0.1284). Thus, gradient attribution and perturbation-based importance should not be considered interchangeable. On Warsaw, BIOT showed a strong significant relationship (ρ=0.744, p=0.0009483), and Simple Mamba showed a significant moderate association (ρ=0.542, p=0.0165). Relationships were weaker and non-significant for EEGMamba (ρ=0.377, p=0.1114) and EEGPT (ρ=0.119, p=0.6266). Therefore, XAI agreement depended on both architecture and dataset. Importantly, BIOT's stronger XAI agreement did not correspond to the highest Warsaw classification performance, demonstrating that explanatory consistency and predictive discrimination are distinct model properties.

Overall, the recurrence of central (C3, C4, Cz), frontal/frontopolar (Fp1, Fp2, Fz and lateral frontal), temporal, and parietal (P3/P4) electrodes is noteworthy. Several were independently reported as relevant to schizophrenia classification [27,30,31]. This convergence supports the neurophysiological plausibility of the learned representations but does not establish these electrodes as diagnostic biomarkers. IG and occlusion identify information influencing model decisions; they do not demonstrate disease causality, cortical source localization, or biological specificity. Future studies integrating source localization, connectivity analysis, frequency-specific attribution, and external cohorts could determine whether these recurrent spatial patterns represent robust schizophrenia-related signatures.

Table 5 compares the present framework with previous deep-learning studies on the Warsaw dataset. Earlier approaches reported accuracies of 85.04–99.90%, sensitivities reaching 99.65%, and F1-scores exceeding 90% in several studies [4,33–41]. EEGMamba achieved 78.33% accuracy, 80.00% sensitivity, and a 76.22% F1-score. Although lower numerically, direct comparison requires caution because previous studies used different validation protocols, including conventional 5-fold and 10-fold cross-validation, LOSO, hold-out validation, and subject-based cross-validation.

An important distinction is that our 10-fold cross-validation was performed at the subject level, ensuring that all EEG windows from each test participant were excluded from training and validation. This evaluates generalization to unseen individuals and prevents subject-level information leakage. Therefore, differences from previous studies may partly reflect validation design and prediction level rather than architecture alone. Segment-level partitioning is not directly comparable with strictly subject-independent evaluation because segments from the same participant may share subject-specific characteristics. Nevertheless, unless the partitioning procedures of previous studies are explicitly verified, their higher performance should not automatically be attributed to data leakage.

Most previous Warsaw approaches also employed task-specific pipelines using handcrafted or transformed representations, including CWT/DWT, time-frequency encoding, dDTF connectivity, microstates, saliency maps, and their combinations with CNNs or other architectures [4,33–41]. In contrast, this study evaluates pretrained EEG foundation models within a subject-independent framework, examining whether large-scale pretrained representations can transfer to schizophrenia detection and remain interpretable across independent datasets. Consequently, the lower numerical performance should be considered in the context of the stringent evaluation and distinct transfer-learning paradigm.

Table 5. Comparison with State-of-the-art DL studies on same database (Warsaw)

ref	Method	Evaluation	Acc	Sen	F1-score
[33]	CWT, AlexNet-SVM, ResNet-18-SVM, VGG-19-SVM, Inception-v3-SVM	10-fold CV	98.60%	99.65%	-
[4]	TE, ResNet50V2, VGG16, DenseNet121, InceptionV3, EfficientNetB0, EfficientNetB0-LSTM	10-fold CV	99.90%	-	99.93
[34]	TE, AlexNet, Resnet-50, Shufflenet, Inception, and Xception	10-fold CV	98.51%	98.51%	98.51%
[35]	spatial–temporal feature mapping and LeViT	14-fold CV	85.04%	-	-
[36]	HSTG-Net, CRFM, Sym-T	5-fold CV	92.34%	91.78 %,	92.00
[37]	Saliency PSD topomap, MBSzEEGNet	5-fold CV subject base	85.71%	76.19%	73.19
[38]	Microstate, CNN	LOSO	93.00%	-	-
[39]	PDC, dDTF, TE, ResNet-50, Inception-v3, DenseNet-201, EfficientNetB0, and NasNet-Mobile	LOSO	96.67%	96.82%	-
[40]	DWT, dDTF, ConvNext-Tiny, MobileVit-V2, RegNetx, MNasNet	Hold out	96.00%	94.10%	94.10
[41]	DWT, MEMD, PCA, CCA, FFNN, custom CNN, VGG16, ResNet50, DenseNet121, EfficientNetB0	Hold out	90.64%	90.09%	91.70
Our (2026)	**EEGMamba,** EEGPT, BIOT	10-fold CV (subject wise)	**78.33%**	**80.00%**	**76.22%**

Table 6 compares the foundation models with the previous ASEEG deep-learning approach [42], which combined recurrence plot (RP) and wavelet synchrosqueezed transform (WSST) images with a hybrid ResNet-18–Vision Transformer and achieved 82.14% accuracy, 82.54% recall, and an F1-score of 0.83 using subject-wise 10-fold cross-validation. EEGMamba achieved comparable values of 79.18% accuracy, 82.33% recall, and an F1-score of 0.79. Its recall was particularly close to the previous method (82.33% vs. 82.54%), with differences of 2.96 percentage points in accuracy and 0.04 in F1-score.

Despite slightly lower performance, the approaches differ substantially. The method in [42] requires explicit RP and WSST transformations and a task-specific CNN–Transformer architecture, whereas EEGMamba transfers representations acquired through large-scale EEG pretraining. Its competitive performance therefore suggests that pretrained EEG representations can be adapted to ASEEG without the image-transformation and feature-fusion pipeline of [42]. The present study additionally compares multiple foundation architectures and investigates their decisions using complementary XAI techniques.

Table 6. Comparison with State-of-the-art DL studies on same database (ASEEG)

ref	Method	Evaluation	Acc	Recall	F1-score
[42]	Merge RP and WSST, fusion of ResNet-18 and ViT	10-fold CV (subject wise)	82.14%	82.54%	0.83
Our (2026)	**EEGMamba,** EEGPT, BIOT	10-fold CV (subject wise)	79.18%	82.33%	0.79

Several limitations should be considered. First, despite using two independent datasets, their sample sizes remain relatively limited for evaluating deep foundation models. Differences in recording systems, sampling rates, acquisition protocols, participant characteristics, and data quality may contribute to performance variation. Second, although subject-wise cross-validation prevents subject overlap between training and testing, evaluation on a completely independent external cohort would provide stronger evidence of generalizability. The relatively large standard deviations, particularly on Warsaw, also indicate fold-level variability.

The XAI analyses have additional limitations. IG and channel occlusion characterize model behavior rather than directly identifying schizophrenia biomarkers, and their results depend on the trained model, input representation, attribution baseline, and perturbation strategy. Their variable agreement further demonstrates that channel importance depends on the explanation method. Moreover, the present analyses primarily address spatial importance without explicitly identifying the temporal or frequency-specific EEG information underlying predictions.

A major advantage of this study is the systematic comparison of EEGMamba, EEGPT, and BIOT under a common subject-independent framework. Their different representation-learning architectures enable direct evaluation under comparable conditions, while Simple Mamba provides a reference for assessing EEG-specific representation learning relative to a general state-space approach. Evaluation on two independent datasets strengthens the analysis beyond a single cohort. Furthermore, combining IG with channel occlusion provides complementary gradient-based and perturbation-based perspectives on model behavior.

Nevertheless, foundation models are generally more computationally demanding and architecturally complex than conventional or lightweight approaches. Their effectiveness may depend on correspondence between pretraining and downstream data, including channel configuration, sampling characteristics, and recording paradigm. Performance variability between ASEEG and Warsaw illustrates this challenge. Spatial explanations also differed across models, and IG–occlusion agreement was not consistently significant. Consequently, strong predictive performance should not automatically be interpreted as producing stable or physiologically equivalent explanations across architectures.

5. Conclusion

This study investigated the applicability of EEG foundation models to subject-independent schizophrenia detection using two independent resting-state EEG datasets. Among the evaluated approaches, EEGMamba demonstrated the strongest overall classification performance, achieving the highest accuracy on both ASEEG and Warsaw, while EEGPT and BIOT also showed competitive performance across several evaluation measures. The lower and more variable performance observed on the Warsaw dataset indicates that the effectiveness of foundation-model representations can vary considerably across EEG datasets.

The XAI analyses further demonstrated that the models relied on distinct spatial EEG patterns. Integrated Gradients and channel occlusion identified central, frontal, and posterior regions as important to varying degrees across models and datasets, while the level of agreement between the two explanation approaches was model dependent. Notably, BIOT exhibited strong and statistically significant agreement between IG and occlusion on the Warsaw dataset, despite not achieving the highest classification accuracy. These findings emphasize that predictive performance and explanation consistency represent complementary aspects of model evaluation. Overall, the results support the potential of pretrained EEG foundation models for schizophrenia classification while highlighting the importance of subject-independent validation, evaluation across independent datasets, and complementary explainability analyses before such models can be considered for broader clinical applications.

Data availability

In this study we have used Warsaw SZ dataset and ASEEG database which can be accessed through bellow links:

Warsaw SZ data: https://repod.icm.edu.pl/dataset.xhtml?persistentId=doi:10.18150/repod.0107441,

ASEEG database: https://zenodo.org/uploads/18029536.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of generative AI use

During the preparation of this work the authors used ChatGPT to check grammatical errors and text consistency due to our non-native nature in English. After using this tool/service, the authors reviewed and edited the content as needed and took full responsibility for the content of the published article.

Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author Contributions

Pouya Tolou Kouroshi: Conceptualization, Methodology, Validation, Writing—original draft, Writing—review & editing. **Sara Bagherzadeh**: Conceptualization, Methodology, Software, Investigation, Validation, Writing—original draft, Writing—review & editing. **Mohammadreza Norouzi:** Conceptualization, Methodology, Investigation, Validation, Writing—original draft. **Ahmad Shalbaf[4]:** Conceptualization, Methodology, Validation, Supervision, Writing—review & editing.

References

1. Saha A, Park S, Geem ZW, Singh PK. Schizophrenia detection and classification: A systematic review of the last decade. Diagnostics. 2024 Nov 29;14(23):2698.

2. Teixeira FL, Costa MR, Abreu JP, Cabral M, Soares SP, Teixeira JP. A narrative review of speech and EEG features for schizophrenia detection: Progress and challenges. Bioengineering (Basel). 2023;10(4):493. doi:10.3390/bioengineering10040493.

3. Shoeibi A, Sadeghi D, Moridian P, Ghassemi N, Heras J, Alizadehsani R, Khadem A, Kong Y, Nahavandi S, Zhang YD, Gorriz JM. Automatic diagnosis of schizophrenia in EEG signals using CNN-LSTM models. Frontiers in neuroinformatics. 2021 Nov 25;15:777977.

4. Bagherzadeh S, Shahabi MS, Shalbaf A. Detection of schizophrenia using hybrid of deep learning and brain effective connectivity image from electroencephalogram signal. Computers in Biology and Medicine. 2022 Jul 1;146:105570.

5. Farhadi Sedehi J, Bagherzadeh S. Personalized prediction of response to selective serotonin reuptake inhibitors (SSRIs) treatment for major depression disorder using deep learning networks from connectivity maps of EEG. Scientific Reports. 2026. 10.1038/s41598-026-74701-2

6. Mostafavi M, Ko SB, Shokouhi SB, Ayatollahi A. Transfer learning and self-distillation for automated detection of schizophrenia using single-channel EEG and scalogram images. Physical and Engineering Sciences in Medicine. 2025 Mar;48(1):3-18.

7. Hao S, Zhang ZJ, Wang X, Zhang P, Ren HP, Ren WC. Enhancing schizophrenia diagnosis efficiency with EEGNet: A simplified recognition model based on γ band features. Psychiatry Res Neuroimaging. 2025;352:112047.

8. Soltani-Nejad M, Salar-Pour F, Rakhshan SA, Nezamabadi-Pour H. Enhanced hybrid deep neural network for EEG-based schizophrenia diagnosis using functional and temporal features. Scientific Reports. 2025 Nov 28;15(1):42592.

9. Shalbaf A, Bagherzadeh S, Maghsoudi A. Transfer learning with deep convolutional neural network for automated detection of schizophrenia from EEG signals. Physical and Engineering Sciences in Medicine. 2020 Dec;43(4):1229-39.

10. Aich U, Saha A, Woźniak M, Ijaz MF, Singh PK. Schizophrenia detection from electroencephalogram signals using image encoding and wrapper-based deep feature selection approach. Scientific Reports. 2025 Jul 1;15(1):21390.

11. Yang C, Westover M, Sun J. Biot: Biosignal transformer for cross-data learning in the wild. Advances in Neural Information Processing Systems. 2023 Dec 15;36:78240-60.

12. Wang J, Zhao S, Luo Z, Zhou Y, Li S, Pan G. Eegmamba: An eeg foundation model with mamba. Neural Networks. 2025 Jul 16:107816.

13. Liu D, Chen Y, Chen Z, Cui Z, Wen Y, An J, Luo J, Wu D. EEG-FM-Compass: Progress, Benchmarking, and Future Directions for EEG Foundation Models. National Science Review. 2026 Aug 3:nwag466.

14. Shi S, Song Y, Wang Y, Xiao J, Lin X, Wang H, Zhao Z, Wang P, Lu Z, Li X, Chen H. Toward general-purpose foundation models for electroencephalography: A unified data registry. Brain-X. 2026 Mar;4(1):e70046.

15. Zan H. Enhancing Schizophrenia Diagnosis Through Multi-View EEG Analysis: Integrating Raw Signals and Spectrograms in a Deep Learning Framework. Clinical EEG and Neuroscience. 2025 Nov;56(6):507-17

16. Gosala B, Singh AR, Tiwari H, Gupta M. GCN-LSTM: A hybrid graph convolutional network model for schizophrenia classification. Biomedical Signal Processing and Control. 2025 Jul 1;105:107657.

17. Almadhor A, Ojo S, Nathaniel TI, Alsubai S, Alharthi A, Hejaili AA, Sampedro GA. An interpretable XAI deep EEG model for schizophrenia diagnosis using feature selection and attention mechanisms. Frontiers in Oncology. 2025 Jul 22;15:1630291.

18. Bagherzadeh S, Norouzi M, Farokhshad F, Tolou Kouroshi P, Ghasri A, Kazemi R, Rostami R, Shalbaf A. Atieh Schizophrenia EEG, a novel high-quality dataset designed to advance biomarker and machine learning research. Scientific Data. 2026 Jul 31. https://doi.org/10.1038/s41597-026-07968-9

19. Bagherzadeh, S., Norouzi, M., Farokhshad, F., Tolou Koroushi, P., Ghasri, A., Bahri Hampa, S., Kazemi, R., Rostami, R., & Shalbaf, A. (2025). ASEEG: Atieh Schizophrenia EEG, a novel high-quality dataset designed to advance biomarker research (Version First version) [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.18029536

20. Olejarczyk E, Jernajczyk W. Graph-based analysis of brain connectivity in schizophrenia. PloS one. 2017 Nov 30;12(11):e0188629.

21. Olejarczyk E, Jernajczyk W. EEG in schizophrenia. RepOD. 2017 Sep;1. [Dataset]. https://doi.org/10.18150/repod.0107441

22. Loshchilov I, Hutter F. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101. 2017 Nov 14.

23. Kuruppu G, Wagh N, Kremen V, Varatharajah Y. EEG foundation models: a critical review of current progress and future directions. Journal of neural engineering. 2026 Apr 1;23(2):021001.

24. Gui Y, Chen M, Su Y, Luo G, Yang Y. EEGMamba: Bidirectional state space model with mixture of experts for EEG multi-task classification. arXiv preprint arXiv:2407.20254. 2024 Jul 20.

25. Wang G, Liu W, He Y, Xu C, Ma L, Li H. EEGPT: Pretrained transformer for universal and reliable representation of eeg signals. Advances in Neural Information Processing Systems. 2024 Dec 16;37:39249-80.

26. Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic Attribution for Deep Networks. Proceedings of the 34th International Conference on Machine Learning (ICML*)*, Proceedings of Machine Learning Research, **70**, 3319–3328.

27. Winterer G, Ziller M, Dorn H, Frick K, Mulert C, Wuebben Y, Herrmann WM. Frontal dysfunction in schizophrenia–a new electrophysiological classifier for research and clinical applications. European archives of psychiatry and clinical neuroscience. 2000 Aug;250(4):207-14.

28. Galderisi S, Mucci A, Volpe U, Boutros N. Evidence-based medicine and electrophysiology in schizophrenia. Clinical EEG and Neuroscience. 2009 Apr;40(2):62-77.

29. Liu Y, Xu P, Hu S. Resting-state gamma power in schizophrenia: a systematic review and meta-analysis. Frontiers in Psychiatry. 2026 Jan 19;16:1731645.

30. Sancho ML, Ellis CA, Miller RL, Calhoun VD. Identifying reproducibly important EEG markers of schizophrenia with an explainable multi-model deep learning approach. In2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) 2024 Jul 15 (pp. 1-4). IEEE.

31. Gröhn C, Norgren E, Eriksson L. A systematic review of the neural correlates of multisensory integration in schizophrenia. Schizophrenia Research: Cognition. 2022 Mar 1;27:100219.

32. Ding X, Fu H, Luo L, Wu J. Working memory deficit in schizophrenia: a systematic review and meta-analysis of fMRI studies examining frontal and parietal brain activity. Folia Neuropathologica. 2024 Jan 1;62(2):136-46.

33. Shalbaf A, Bagherzadeh S, Maghsoudi A. Transfer learning with deep convolutional neural network for automated detection of schizophrenia from EEG signals. Phys Eng Sci Med. 2020;43:1229-1239. doi:10.1007/s13246-020-00925-9

34. Bagherzadeh S, Maghsoudi A, Shalbaf A. Detection of schizophrenia patients using convolutional neural networks from brain effective connectivity maps of electroencephalogram signals. TEHRAN UNIVERSITY MEDICAL JOURNAL (TUMJ)[Internet]. 2022;79(10):754-763. Available from: https://sid.ir/paper/1060712/en

35. Li B, Wang J, Guo Z, Li Y. Automatic detection of schizophrenia based on spatial–temporal feature mapping and LeViT with EEG signals. Expert Syst Appl. 2023;229:119969. doi:10.1016/j.eswa.2023.119969
36. Alsharabi N, Mahendran RK, Maher M, Alharbi S, Aljohani AF, Alogaili MAH. EEG-SymNet: multi-channel EEG signal-based schizophrenia diagnosis using channel recalibration and symmetric transformer. Eng Sci Technol Int J. 2025;69:102116. doi:10.1016/j.jestch.2025.102116
37. Huang W, Yan G, Chang W, Zhang Y, Yuan Y. EEG-based classification combining Bayesian convolutional neural networks with recurrence plot for motor movement/imagery. Pattern Recognition. 2023 Dec 1;144:109838.

38. Lillo E, Mora M, Lucero B. Automated diagnosis of schizophrenia using EEG microstates and Deep Convolutional Neural Network. Expert Syst Appl. 2022;210:118236. doi:10.1016/j.eswa.2022.118236
39. Bagherzadeh S, Shalbaf A. EEG-based schizophrenia detection using fusion of effective connectivity maps and convolutional neural networks with transfer learning. Cogn Neurodyn. 2024;18:2767-2778. doi:10.1007/s11571-024-10121-0

40. Liu T, Zhang J, Dong X, Li Z, Shi X, Tong Y, Yang R, Wu J, Wang C, Yan T. Occipital alpha connectivity during resting-state electroencephalography in patients with ultra-high risk for psychosis and schizophrenia. Frontiers in psychiatry. 2019 Aug 16;10:553.

41. Bhadra S, Kumar CJ, Bhattacharyya DK. Multiview EEG signal analysis for diagnosis of schizophrenia: an optimized deep learning approach. Multimed Tools Appl. 2025;84:27577-27608. doi:10.1007/s11042-024-20205-y
42. Bagherzadeh S, Norouzi MR, Tolou Kouroshi P, Farokhshad F, Ghasri AH, Kazemi R, Rostami R, Shalbaf A. Explainable Schizophrenia Detection Using Fusion of Convolutional neural network and Vision Transformer from Electroencephalogram. Scientific reports, 2026, in press