Scientific Reports

https://doi.org/10.1038/s41598-026-59478-8

Article in Press

Aggregating XAI-based explanations to identify spectral–spatial patterns in CNN-based resting-state EEG classification

Received: 27 April 2026

Accepted: 22 June 2026

Published online: 25 June 2026

Cite this article as: Rejer I., Gago I. & Marozas V. Aggregating XAI-based explanations to identify spectral–spatial patterns in CNN-based resting-state EEG classification. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-59478-8

Izabela Rejer, Izabela Gago & Vaidotas Marozas

Aggregating XAI-Based Explanations to Identify Spectral–Spatial Patterns in CNN-Based Resting-State EEG Classification

Izabela Rejer[1], **Izabela Gago**[1,*], **and Vaidotas Marozas**[2]

[1]West Pomeranian University of Technology in Szczecin, Żołnierska 49, 71-210 Szczecin, Poland
[2]Kaunas University of Technology, K. Donelaičio g. 73, LT-44249 Kaunas, Lithuania
[*]izabela.gago@zut.edu.pl

ABSTRACT

Convolutional neural networks (CNNs) achieve high performance in electroencephalographic (EEG) classification tasks; however, their decision-making mechanisms remain difficult to interpret. Explainable artificial intelligence (XAI) methods are typically applied to provide insight into individual model decisions, yet such explanations do not reveal the overall structure of the patterns learned by the network. In this study, we hypothesize that, in EEG analysis, XAI can serve a deeper role: when appropriately applied, it can expose the general patterns learned by a trained CNN, thereby transforming it from a purely predictive model into a framework capable of revealing candidate discriminative structures that may form the basis for future neuroscientific hypotheses. This shift is enabled by structured aggregation of local explanations, through which instance-level insights are consolidated into cross-subject patterns. To implement this strategy, we employ averagedLIME, an extension of Local Interpretable Model-agnostic Explanations, which aggregates sample-level explanations into global class-level saliency maps. In trial-based EEG paradigms, such aggregation reinforces consistent patterns across samples analogously to event-related potentials averaging. In this study, we examine whether this strategy remains effective in a temporally unaligned resting-state setting, taking subject-independent classification of Alzheimer's disease and cognitively normal controls as a case study. Four neural architectures are compared, with spectral CNN yielding the most robust performance (95.81% test accuracy). The best-performing model is subsequently analyzed using averagedLIME, while SHapley Additive exPlanations (SHAP) and Grad-CAM are employed as complementary explanation techniques. Quantitative analyses demonstrated that the averagedLIME patterns were stable across cross-validation folds, robust to perturbation parameter selection, reproducible on unseen test data, and strongly supported by independent SHAP explanations. When interpreted in conjunction with the averaged input representations, the resulting saliency maps reproduced established EEG slowing patterns in Alzheimer's disease, while also highlighting localized spatial–spectral structures that were less apparent in the averaged EEG representations alone. These findings demonstrate that structured aggregation of CNN explanations enables extraction of stable cross-subject patterns from resting-state EEG and may reveal candidate discriminative structures that can motivate future neuroscientific hypotheses.

Keywords: Explainable Artificial Intelligence (XAI), averagedLIME, Alzheimer's Disease, Convolutional Neural Networks (CNN), Electroencephalography (EEG), Resting-State EEG, Model Interpretability, Feature Attribution, Spectral CNN

Introduction

In recent years, convolutional neural networks (CNNs) have demonstrated strong performance in the analysis of electroencephalographic (EEG) data [1], automatically extracting complex spatial and spectral features relevant to cognitive and pathological conditions. Despite their high effectiveness, CNNs are typically treated as black-box classifiers, whose internal representations remain difficult to interpret. Explainable artificial intelligence (XAI) methods are therefore increasingly employed to provide insight into the decision-making mechanisms of these models.

In EEG research, XAI is most often used as a post-hoc validation tool, verifying whether a trained network focuses on physiologically plausible regions during classification [2, 3]. In this work, we pursue a different objective. We investigate whether stable cross-subject patterns can be extracted from CNN explanations through structured aggregation of local attribution maps. Rather than inspecting individual explanations, we construct class-level saliency maps that summarize the spectral–spatial patterns learned by the network across multiple samples. In this setting, XAI serves not only as a validation mechanism but also as a framework for identifying candidate discriminative patterns encoded in the learned model representations.

The key methodological component enabling this transformation is averagedLIME, introduced in our previous work [4].

AveragedLIME extends the standard Local Interpretable Model-agnostic Explanations (LIME) framework by aggregating local saliency maps across all samples belonging to a given class. In trial-based EEG paradigms with temporally aligned epochs, such aggregation acts analogously to event-related potential (ERP) averaging: consistent discriminative patterns are reinforced, whereas sample-specific noise is attenuated. Importantly, such aggregation is generally not meaningful in classical computer vision tasks, where averaging saliency maps across heterogeneous images would obscure rather than reveal relevant structures. In EEG, however, consistent spatial organization across subjects allows global class-level patterns to emerge.

In our earlier study, this strategy was validated in stimulus-locked paradigms with temporally aligned trials. Here, we investigate a more demanding scenario, asking whether stable global patterns can still be extracted in the absence of temporal alignment. To address this question, we apply our approach to resting-state EEG in the context of subject-independent classification of Alzheimer's disease (AD) patients and cognitively normal controls (CN). In contrast to stimulus-locked paradigms, resting-state data lack temporal synchronization, and discriminative information is primarily expressed through distributed spectral organization rather than time-locked responses. This makes resting-state EEG a particularly challenging context for stabilizing and interpreting CNN-derived patterns.

At the same time, Alzheimer's disease provides a well-established neurophysiological reference for interpretation. Resting-state EEG alterations in AD are commonly described as global slowing, reflected in increased delta and theta power and reduced alpha and beta activity, particularly over frontal and parieto-occipital regions [5, 6]. These effects are often summarized using measures such as the theta/alpha ratio or shifts in peak alpha frequency toward lower values [7, 8, 9]. While these canonical findings offer a reference point, the precise spatial–spectral organization of discriminative information remains incompletely characterized, particularly at the level of specific electrode–frequency interactions.

The experimental evaluation was conducted on a cohort of 63 subjects (35 AD, 28 CN) from the dataset *A dataset of EEG recordings from: Alzheimer's disease, Frontotemporal dementia and Healthy subjects* [10]. To ensure subject-independent generalization and minimize bias, subjects were divided into seven non-overlapping groups and evaluated using a leave-some-subjects-out scheme. To establish a reliable classification model prior to interpretability analysis, we compared four neural architectures trained on different EEG representations: (i) a traditional artificial neural network (ANN) operating on hand-crafted features, (ii) a temporal CNN trained on raw channel × time signals, (iii) a spectral CNN trained on channel × frequency representations, and (iv) the compact EEGNet architecture. Among the evaluated architectures, the spectral CNN achieved the highest accuracy and most stable generalization performance.

The best-performing spectral CNN model (test accuracy: 95.81%; balanced precision and recall across classes) was subsequently analyzed using averagedLIME to generate global spatial–spectral saliency maps. SHapley Additive exPlanations (SHAP) and Grad-CAM techniques were additionally employed to verify the consistency of the identified patterns. When interpreted in conjunction with the averaged input data, the resulting global class-level saliency maps highlighted spectral–spatial regions consistent with established electrophysiological findings in AD, such as posterior alpha attenuation, while also revealing more localized spatial–spectral structures, including asymmetric beta reductions and symmetric theta enhancements across temporal regions. These results indicate that systematic aggregation of local explanations enabled identification of not only canonical patterns of EEG slowing but also more specific spatial–spectral relationships that became visible only at the population level. In this sense, the CNN functions not merely as a predictive classifier, but also as a source of candidate discriminative patterns whose aggregation may reveal stable cross-subject structures that would be difficult to identify from individual explanations alone.

The contributions of this study are fourfold. First, we demonstrate that structured aggregation of local XAI explanations enables the extraction of stable class-level patterns from CNN models trained on resting-state EEG in the context of AD classification. Second, we provide a comparative evaluation of neural architectures for AD classification, identifying spectral representation as particularly effective for capturing disease-related resting-state alterations. Third, we perform a comprehensive validation of the averagedLIME explanation framework, evaluating its stability across cross-validation folds, robustness to perturbation parameters selection, generalization to unseen test data, agreement with alternative explanation methods, and perturbation-based faithfulness. Finally, we show that, when interpreted in conjunction with averaged input representations, aggregated CNN explanations highlight spectral–spatial regions consistent with established electrophysiological findings in AD while also revealing more specific spatial–spectral discriminative structures that are not readily apparent from individual explanations.

Related Work

The use of CNNs for the detection of AD based on EEG signals has been the subject of increasing interest in recent years. Numerous studies have investigated the effectiveness of deep learning approaches in distinguishing AD patients from cognitively healthy individuals, with particular focus on signal preprocessing strategies and CNN architecture design.

One of the the first examples is the work by Morabito et al. [11], who developed and evaluated a CNN-based classification system designed to distinguish between CN, mild cognitive impairment (MCI), and AD using EEG recordings from 19

channels. In their study, the EEG signals were band-pass filtered between 0.1 and 30 Hz prior to feature extraction and model training. Next, 12 statistical features were extracted from each channel: the average, standard deviation, and skewness of the energy distribution for the entire signal and for three frequency bands. The resulting feature vector, composed of 228 features (19 channels x 12 parameters), was fed into a CNN. The network architecture consisted of three blocks. The first included a single convolutional layer with 19 filters of 12 elements (one per channel), followed by sigmoid, and max-pooling. The second block was an autoencoder that compressed the 19 outputs of the convolutional layer into 10 latent features. The final block was a three-layer classifier with 10, 7, and 2 neurons, respectively. The study involved 142 participants (63 AD, 56 MCI, 23 CN) and reached an accuracy of 82% for the three-class classification and 85% for the binary classification of AD vs. CN.

Ieracitano et al. [12] further addressed this problem using a CNN trained on 19-channel EEG data from 189 balanced AD, MCI, and CN participants. Their preprocessing pipeline included band-pass filtering between 0.5 and 32 Hz, downsampling from 1024 to 256 Hz, segmentation into non-overlapping 5-s epochs (1280 samples × 19 channels), and computation of the power spectral density (PSD) using a 1280-point FFT, yielding 159 features within the 0.5–32 Hz range. The resulting 19×159 PSD matrices were normalized, converted to grayscale images, and used as inputs to two CNN models (CNN1 and CNN2). CNN1 consisted of a single convolutional layer (16 filters, 3×3), followed by ReLU activation and max-pooling (3×3, stride 2), a single fully connected layer with 300 neurons, and an output softmax layer. CNN2 extended this architecture with an additional convolutional layer (32 filters, 3×3) and a second max-pooling layer (2×2, stride 2). The CNN1 architecture yielded superior performance, achieving 83.33% accuracy in the three-class classification task and 92.95% in the binary classification of AD vs. CN.

Huggins et al. [13] addressed the same three-class classification task by adapting the AlexNet architecture for EEG analysis. Their study used EEG data from 141 participants (52 AD, 37 MCI, 52 HA). Before being fed into the network, the signals underwent band-pass filtering (1–60 Hz), ICA (Independent Component Analysis)-based artifact removal, additional notch filtering at 21 and 42 Hz, and segmentation into epochs. Next, CWT was computed for each EEG channel, and the resulting time–frequency representations were arranged into a single topography-consistent image reflecting the scalp layout. These composite images were then resized and converted into 224 × 224 × 3 pseudo-images used as input to the DL model. The adapted AlexNet architecture consisted of five convolutional layers followed by two fully connected layers with 4096 units each and a final softmax layer. The convolutional layers employed 96 filters of size 11×11 in the first layer, 256 filters of size 5×5 in the second, 384 filters of size 3×3 in the third and fourth, and 256 filters of size 3×3 in the fifth. The model achieved a three-class classification accuracy of 98.9%.

Another example is the study by Xia et al. [14], who proposed a novel CNN architecture — Deep Pyramid CNN (DPCNN) — for distinguishing between the same three classes. The CNN architecture featured a pyramidal design, in which the number of filters decreased progressively with network depth. The network comprised four convolutional layers with 24, 16, 8, and 4 filters, respectively. Each layer was followed by batch normalization and the Leaky ReLU activation function. The classification block consists of two fully connected layers, with the first followed by a Leaky ReLU activation and dropout, and the second serving as the final softmax layer. To evaluate the proposed architecture, the authors used a dataset comprised recordings from 100 participants (49 AD, 37 MCI, 14 CN), acquired from 19 EEG electrodes and band-pass filtered within 0.5–48 Hz. Each EEG channel was represented by 16 Fourier-based frequency-domain features. These were then concatenated across the 19 channels, resulting in a one-dimensional feature vector of length 304 (19 × 16 features). The model achieved a three-class classification accuracy of 97.10%.

Yet another example is the work by Kachare et al. [15], who proposed a lightweight CNN architecture named LCADNet for AD detection. Their CNN model included two convolutional layers (with 20 and 10 kernels of size 3×3, respectively), one max-pooling layer (2x21) placed between the convolutional layers, a flattening layer, two fully connected layers (with 50 and 80 units, respectively), and a softmax output layer with two units. Prior to being fed into the network, the EEG signals were filtered using a bandpass filter (0.5–100 Hz) and segmented into 5-seconds epochs of size 16x1280 (16 EEG channels x 1280 time points; sampling rate: 256 Hz). The dataset included recordings from 23 participants (12 diagnosed with AD and 11 CN). The authors report classification accuracy of 98.5% in the cross-validation scheme and the mean accuracy of 93.26% when trained and tested on independent datasets.

To provide a more comprehensive view of the existing literature, Table 1 compiles both the studies discussed above and additional works addressing EEG-based AD recognition. The table is organized into two consecutive sections. The first section includes CNN-based studies, together with works that benchmark CNN architectures against alternative machine-learning and deep-learning methods. The second section complements this perspective by presenting studies that rely exclusively on non-CNN deep learning or classical machine-learning techniques.

Across the studies summarized in Table 1, both CNN-based approaches and alternative deep-learning or classical machine-learning methods frequently report very high classification accuracies, often exceeding 95% (e.g., [14], [25], [16]) and reaching up to 98.9% [13]. However, most of these top-performing results is based on cross-validation performed within a single dataset rather than on independent test cohorts. Such evaluation schemes — used, for example, in [13], [14], [16], or [20]

Table 1. Related papers on EEG-Based AD recognition; *ASR* - Artifact Subspace Reconstruction, *BD* – Box Dimension, *BPF* – Band-pass Filtering, *CWT* – Continuous Wavelet Transform, *DWT* – Discrete Wavelet Transform, *EPNN* – Enhanced Probabilistic Neural Network, *FT* – Fourier Transform, *HE* – Hurst Exponent, *HFD* – Higuchi's Fractal Dimension, *KFD* – Katz Fractal Dimension, *kNN* – k-Nearest Neighbors, *LDA* – Linear Discriminant Analysis, *LightGBM* - Light Gradient-Boosting Machine, *LPF* – Low-pass Filtering, *LPSD* – Lomb-Scargle periodogram PSD, *MA rejection* – Manual Artefact Rejection, *MLP* – Multi-Layer Perceptron, *MSPCA* – Multiscale Principal Component Analysis, *NF* – Notch Filtering, *RBP* - Relative Band Power, *RF* - Random Forest, *SVM* – Support Vector Machine, *TD-PSD* – Time-Dependent PSD, *WPSD* – Welch's PSD.

Classifiers	Subjects	Ch.	Preprocessing	Features / Input images	Accuracy	Ref.
CNN	63 AD, 56 MCI, 23 CN	19	BPF (0.1-30 Hz)	12 statistical features per channel	85% (AD vs. CN)	[11]
CNN1, CNN2	63 AD, 63 MCI, 63 CN	19	BPF (0.5-32 Hz)	159 PSD features per channel	92.95% (AD vs. CN)	[12]
AlexNet	52 AD, 37 MCI, 52 CN	21	BPF (1-60 H z), ICA, NF (21 and 42 Hz)	21 CWT scalograms combined to a single image according to the scalp topography	98.9% (AD vs. MCI vs. CN)	[13]
DPCNN	49 AD, 37 MCI, 14 CN	19	BPF (0.5-48 Hz)	16 FT coefficients per channel	97.10% (AD vs. MCI vs. CN)	[14]
LCADNet, VGG, ResNet, and others	12 AD, 11 CN	16	BPF (0.5-100 Hz), NF (60 Hz)	-	93.26% (AD vs. CN)	[15]
CNN Ensemble	Set1: 24 AD, 24 CN, Set2: 80 AD, 12 CN	19	BPF (0.5-45 Hz), MA rejection, decomposition into 5 frequency bands	-	97.90% (AD vs. CN)	[16]
CNN, EEGNet, MLP, and others	36 AD, 23 FTD, 29 CN	19	BPF (0.5-40 Hz), ASR, ICA	1185 FT coefficients per channel	79.45% (AD vs. CN)	[17]
ResNet, SVM, kNN, and others	24 AD, 24 CN	19	MSPCA	13 statistical features per each of the 4 DWT components obtained for each channel	97.83% (AD vs. CN)	[18]
CNN, kNN, SVM, LDA	64 AD, 64 MCI, 64 CN	19	-	7 TD-PSD features	82.30% (AD vs. MCI vs. CN)	[19]
ANN	57 AD, 7 MCI, 102 CN	21	BPF (0.5-60 Hz), NF (50 Hz)	band power in 5 frequency bands; TBR; sample, Shannon, dispersion and multiscale entropy	93.74% (mild AD vs. moderate AD vs. severe AD vs. MCI vs. CN)	[20]
LSTM, SVM	51 AD, 35 CN	20	LPF, Wavelet-ICA	400 features: mean, median, LPSD, WPSD for 5 frequency bands and each channel	90.28% (for all channels and rythms)	[21]
EPNN	37 AD, 37 MCI	19	BPF (0.5-32 Hz), NF (50 Hz), MA rejection	BD, HFD, KFD, HE features calculated for each channel and each frequency band	90.3% (AD vs. MCI)	[22]
SVM, kNN	50 AD, 50 CN	8	LPF, MA rejection, ICA, wavelet-based denoising	power-based, wavelet-based, and complexity-based features	96% (AD vs CN)	[23]
SVM	15 AD, 9 MCI, 17 CN	19	BPF (0.5-30 Hz), NF, MA rejection	RP, sample entropy, and permutation entropy on 5-level DWT	95.12% for AD, 97.56% for MCI, 92.68% for CN	[24]
RF, SVM, kNN, MLP, LightGBM	36 AD, 23 FTD, 29 CN	19	BPF (0.5-45 Hz), ASR, ICA	5 RBP features per channel	77.01% (AD vs. CN, with RF)	[10]

— may allow information from the training folds to influence the test folds. As a consequence, the reported accuracies may overestimate the true generalization capability of the models [17]. These results therefore provide useful within-dataset benchmarks but should be interpreted with caution when assessing broader clinical applicability.

The increasing number of studies applying CNN-based models to EEG-based AD recognition has highlighted the need for greater model interpretability to strengthen confidence in these approaches. Yet, despite this growing demand, XAI techniques remain used only sporadically in this field, leaving model interpretability an underexplored problem.

One of the few studies addressing this issue is the work by Sidulova et al. [26], which aimed to develop an automated EEG classification system for the early detection of AD and MCI, as well as to provide both quantitative and visual explanations of the diagnostic decisions made by the model. The study utilized EEG recordings from 440 participants, including 284 with AD, 56 with MCI, and 100 CN. The classification was performed using CNNs trained on two-dimensional topographic maps (headplot spectrograms) representing mean power values computed from 19 EEG channels within one of five standard frequency bands. Each spectrogram had a resolution of 488×435 pixels, and a separate CNN model was trained for each frequency band. To enhance predictive performance, an ensemble approach was implemented by averaging the outputs of the five CNNs corresponding to different frequency bands. The CNN ensemble achieved a validation accuracy of 96%, evaluated using a 10-fold cross-validation strategy. To interpret the model's decisions, the LIME method was applied to 40 randomly selected topographic maps, with 400 perturbations generated for each map. The regions highlighted by LIME were predominantly located over the temporal and frontal lobes, as well as the motor cortex—areas. However, the analysis did not reveal any clear or consistent correspondence between the spatial patterns identified and the predicted diagnostic outcomes.

Another example is the study by Lopes et al. [27], who proposed a CNN-based framework combined with gradient-derived saliency maps to identify EEG biomarkers of AD. Their dataset included 54 individuals — 20 cognitively normal participants, 19 with mild AD (AD1), and 15 with moderate-to-severe AD (AD2). The CNN was trained on 45×45 pseudo-image power-modulation spectrograms (1 Hz resolution) derived from eyes-closed resting-state 20-channel EEG. Since for each channel separate spectrogram was generated, the CNN input had a dimension of 45x45x20. The network architecture consisted of two convolutional layers followed by three fully connected layers. The CNN was trained on five diagnostic tasks (CN vs. AD1 vs. AD2; CN vs. AD; CN vs. AD1; AD1 vs. AD2; CN vs. AD2), yielding accuracies of 90.5%, 87.3%, 91.4%, 92.5%, and 93.0%, respectively. After training, for each power-modulation spectrogram from the training set, saliency maps were generated by computing the gradient of the output category with respect to the last dense layer. The individual saliency maps were then averaged over all training samples (separately for each task) to obtain the average final map of dimension 45x45. The map was segmented via thresholding and k-means clustering and the most discriminative regions ("patches") were identified. The patches included both previously known regions and new biomarkers. Significant areas were detected in the delta and low theta ranges (<5 Hz), as well as in the beta and gamma bands, which are usually discarded due to artifacts but here provided valuable diagnostic information. The importance of cross-range modulation, such as beta–m–alpha and gamma–m–alpha, was also observed. The most relevant features originated from the temporal, parietal, and occipital regions, with occipital contributions observed at higher frequencies than previously reported.

Another contribution in this field is the work of Khan et al. [28], who proposed a novel deep learning architecture combining Temporal Convolutional Network (TCN) and LSTM for the automatic discrimination of AD, frontotemporal dementia (FTD), and CN. The dataset consisted of 19-channel resting-state EEG recordings (eyes closed) collected from 88 participants (36 with AD, 23 with FTD, and 29 CN). As input, the authors used a feature matrix built from six RBP features computed for each EEG channel across six frequency bands (delta, theta, alpha, zaeta, beta, gamma), which were subsequently averaged across channels. The model achieved an average accuracy of 99.7% for the binary discrimination tasks, while the three-class classification reached 80.34%. To improve interpretability, SHAP was applied to estimate the relative importance of each frequency band in the model's predictions. For identifying healthy participants, the zaeta band (6-24 Hz) exhibited the highest relevance, followed by beta (24-30 Hz) and theta (4-8 Hz), whereas alpha (8-16 Hz), gamma (30-45 Hz), and delta (0.5-4 Hz) contributed minimally. For AD classification, beta emerged as the most influential band, with zaeta and theta also playing notable roles. A comparable pattern was observed for the FTD class, where beta again had the strongest impact and zaeta ranked second. As the analysis relied on power values averaged across all EEG channels, the resulting explanations reflected only global feature importance and did not reveal how specific electrodes or spatially organized neurophysiological patterns contributed to the model's decisions.

Collectively, these studies demonstrate that XAI techniques can provide valuable insight into CNN-based EEG classification. However, the resulting explanations remain focused primarily on individual predictions or on aggregated relevance estimates derived from predefined features. Methods such as LIME, SHAP, Grad-CAM, and saliency-based approaches are most commonly applied to explain individual model decisions, whereas broader analyses typically summarize relevance values at the level of channels, frequency bands, or handcrafted features rather than preserving the full spatial–spectral structure learned by the network [29, 30, 31].

More generally, the XAI literature has proposed a variety of global explanation approaches aimed at characterizing model

Table 2. Dataset characteristics.

Characteristic	Description
Dataset	A Dataset of Scalp EEG Recordings of Alzheimer's Disease, Frontotemporal Dementia and Healthy Subjects from Routine EEG
Subjects	Total: 88 (AD: 36, FTD: 23, CN: 29)
Recording condition	Eyes-closed resting-state EEG
EEG system	Nihon Kohden 2100
EEG channels	Fp1, Fp2, F7, F3, Fz, F4, F8, T3, C3, Cz, C4, T4, T5, P3, Pz, P4, T6, O1, O2
Electrode placement	International 10–20 system
Sampling frequency	500 Hz
Average recording duration	AD: 13.5 min FTD: 12 min, CN: 13.8 min
Available clinical scores	MMSE
Preprocessing	Butterworth band-pass filter (0.5–45 Hz)
	A1–A2 re-reference
	Artifact Subspace Reconstruction (ASR)
	ICA (RunICA) + ICLabel artifact removal

behavior at the dataset level. These include methods based on feature-importance aggregation, global surrogate models, rule extraction, and accumulation of local explanations [32, 33, 34]. Such approaches have proven effective for tabular and feature-based data, where model behavior can be summarized through a fixed set of interpretable variables. In contrast, relatively little attention has been devoted to approaches capable of identifying cross-subject explanation patterns while preserving the full spectral–spatial structure of EEG representations (with some exceptions, e.g., [27]). Addressing this gap constitutes the primary motivation of the present study.

Materials and Methods

Alzheimer Dataset and Data Partitioning

In this study, we utilized data from the publicly available dataset *A Dataset of Scalp EEG Recordings of Alzheimer's Disease, Frontotemporal Dementia and Healthy Subjects from Routine EEG* [10], hosted on the OpenNeuro repository. The dataset includes resting-state EEG recordings, acquired with eyes closed, from participants diagnosed with Alzheimer's disease (AD; n = 36), frontotemporal dementia (FTD; n = 23), and cognitively normal controls (CN; n = 29).

The EEG recordings available within this dataset were acquired from 19 scalp electrodes, placed according to the international 10–20 system [35] (sampling rate 500Hz). Preprocessing was performed in EEGLAB and included Butterworth band-pass filtering (0.5–45Hz), re-referencing to A1–A2, artifact correction with Artifact Subspace Reconstruction (ASR), and Independent Component Analysis (RunICA algorithm) for automatic removal of ocular and muscular artifacts. In addition, the dataset provides Mini-Mental State Examination (MMSE) scores for all participants, offering a standardized measure of cognitive function. Further details on participant demographics and preprocessing procedures are reported in [10]. The detailed description of the dataset is provided in Table 2.

For the purposes of this study, only the AD and CN cohorts were considered, as the primary objective was to identify EEG patterns that distinguish between AD and CN participants. To enable fair validation of subsequent classification models, we further excluded two participants — one from the AD group (sub003) and one from the CN group (sub060). Exclusion was based on the shortest effective recording length, which varied substantially across participants due to preprocessing procedures (including artifact removal). After this adjustment, the dataset comprised 35 AD and 28 CN participants, which enabled partitioning into seven balanced subsets, each comprising nine participants (4 CN and 5 AD).

Prior to trial generation, all EEG recordings were band-pass filtered to the 4–30 Hz range. Although delta-band activity (0.5–4 Hz) is frequently reported in Alzheimer's disease EEG studies, preliminary experiments performed during development of the analysis pipeline revealed that inclusion of this frequency range resulted in strong dominance of low-frequency power within the generated spectral representations, despite logarithmic compression of the power spectral density estimates. Consequently, higher-frequency structures became substantially less prominent and the resulting representations were dominated by the delta band. To obtain a more balanced representation of the EEG spectrum and facilitate analysis of patterns distributed

across multiple frequency bands, the frequency range was restricted to 4–30 Hz. This choice should not be interpreted as evidence that delta activity is irrelevant for Alzheimer's disease classification, but rather as a methodological decision intended to prevent dominance of a single frequency range within the analyzed representations.

To ensure consistency and equal contribution of data from all participants, recordings were truncated to the shortest available duration across subjects, resulting in a standardized 554-second segment for each participant. These segments were then divided into two types of trials: (i) 30-second trials with 20-second overlap and (ii) 10-second trials without overlap. In the first case, 52 trials were obtained per participant, each represented as a tensor of size $1 \times 19 \times 15000$ (one signal component, 19 EEG channels, and 15,000 time points). In the second case, 55 trials were generated per participant, each represented as a tensor of size $1 \times 19 \times 5000$. This dual segmentation strategy was motivated by empirical observations that time-domain models tended to perform better on shorter trials, whereas frequency-domain approaches benefited from longer segments.

Image Representations and Network Architectures

Four modeling approaches were examined, differing in both the neural network architecture and the form of the input EEG representation. The first two models operated directly on the raw time-series signals (*timeCNN* and *EEGNet*), the third utilized spectral representation derived from the frequency domain (*fftCNN*), and the last relied on time–frequency features extracted from the raw signals (*ANN*).

Each approach is briefly outlined below, together with the final set of architectural parameters employed. Additionally, the whole experimental pipeline is presented in Fig. 1. Model parameters were optimized using a grid-search procedure, with either a three- or five-level grid, depending on the number of tunable hyperparameters in each architecture. This strategy ensured a systematic exploration of the parameter space while maintaining computational feasibility.

The evaluated approaches were compared as complete EEG analysis frameworks rather than as different architectures operating on standardized input representations. While the underlying EEG preprocessing pipeline remained the same, input dimensionality and signal representation were adapted to the requirements of each model. This strategy preserved the intended operating characteristics of the compared approaches and avoided introducing biases associated with enforcing a common input format across fundamentally different architectures.

The timeCNN approach employed a standard CNN architecture applied directly to 10-second EEG trials. The CNN architecture consisted of four convolutional layers with kernel sizes of 3x3, comprising 16, 32, 64, and 128 filters, respectively. Each convolutional layer was followed by batch normalization, a ReLU activation function, and max pooling for dimensionality reduction. Due to the large disproportion between the spatial and temporal dimensions of the original input (19×5000), pooling was applied non-uniformly across the two axes. Specifically, the pooling sizes for successive layers were set to 2×5, 1×5, 1×5, and 1×5. The convolutional block was followed by three fully connected layers with 50, 40, and 30 units, respectively, each using ReLU activation. Each layer, both convolutional and fully connected, was followed by a dropout operation with a dropout rate of 0.5. The fully connected block was terminated with a softmax output layer for classification.

The EEGNet approach also utilized 10-second trials, but their temporal dimension was reduced via tenfold resampling, resulting in trials of size 19×500. Unlike the timeCNN model, this method employed EEGNet — a compact CNN developed for EEG signal analysis [36]. In our study, we used a modified version of this architecture, following the strategy proposed by Liu et al. [37], which introduces multiple parallel branches with temporal kernels of varying lengths, each designed to extract features corresponding to distinct EEG frequency bands.

Specifically, our model processed EEG trials of size $1 \times 19 \times 500$ through three parallel convolutional branches, each using 16 kernels of sizes 1×3, 1×5, and 1×9, respectively. Since the data were resampled from $f_s = 500$Hz to $f_s = 50$Hz, these kernel lengths corresponded to temporal windows covering the following oscillatory ranges. The shortest kernel (1×3) spanned 60ms and was thus sensitive to faster beta rhythms (approximately 13–30Hz), the medium kernel (1×5) covered 100ms, aligning with the alpha band (around 8–13Hz), while the longest kernel (1×9) extended over 180ms, capturing slower theta oscillations (approximately 4–8Hz). In each branch the temporal convolution was followed by batch normalization and depthwise convolution across the spatial domain using 32 depthwise filters of size (19,1). The filtered data was then passed once again through batch normalization, an exponential linear unit (ELU) activation function, average pooling (1,4), and dropout ($p = 0.5$).

The outputs from the three parallel branches were concatenated along the feature dimension, yielding 24 feature maps and passed to the second block of the EEGNet architecture. In this block, we followed the original EEGNet approach by applying a single separable convolution layer, implemented as two consecutive operations: i) a depthwise convolution with 96 filters of size (1×16), applied independently to each feature map to capture temporal dynamics within each spatial location; ii) a pointwise convolution with 96 filters of size (1×1), which linearly combined the outputs across feature maps, enabling integration across spatial filters and frequency bands. This convolutional stage was followed by batch normalization, ELU activation, average pooling with a kernel size of (1×8), and dropout with a rate of 0.5. Finally, the output was flattened and passed to a fully connected softmax layer with two units, corresponding to the two target classes.

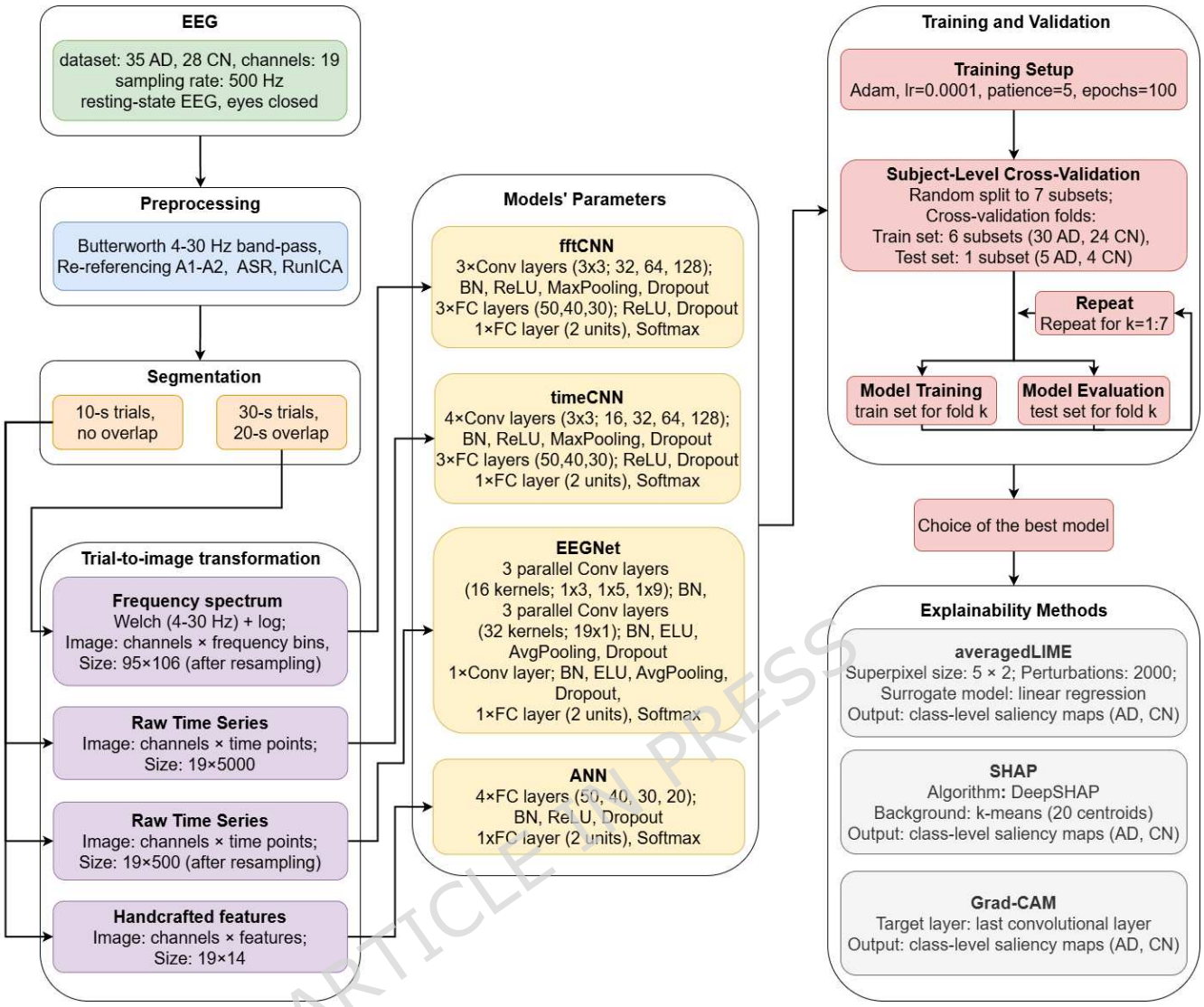

Figure 1. Overview of the experimental pipeline, including dataset partitioning, EEG-image representations, and key deep-learning model parameters.

In the fftCNN approach, spectral representations were constructed by estimating channel-wise power spectral density (PSD) using Welch's method, which computes averaged FFT-based periodograms over overlapping signal segments. A 2-second Hamming window with 50% overlap was employed, and the resulting PSD estimates were subsequently logarithmically transformed to stabilize variance. From the resulting spectra, the frequency range of 4–30Hz was extracted, which yielded 53 frequency bins with a resolution of 0.5Hz. For each trial, this representation was computed independently for all 19 EEG channels, resulting in a two-dimensional matrix of size 19×53. To obtain an image-like input better suited for convolutional processing, the representation was resampled by a factor of five along the channel dimension and by a factor of two along the frequency dimension, producing final matrices of size 95×106. These resampled images were subsequently used as inputs to a convolutional neural network. The network architecture was largely similar to that employed in the timeCNN approach, with two main modifications. First, only three convolutional layers were used, comprising 32, 64, and 128 filters, respectively. Second, max pooling was applied uniformly after each convolutional layer, with a consistent pooling size of 2×2.

In the ANN approach, handcrafted features were extracted from EEG epochs. For each channel, a set of 14 features was computed: spectral power in five frequency bands (4–8Hz, 8–13Hz, 13–15Hz, 15-18Hz, and 18-30Hz), four statistical descriptors (mean, standard deviation, skewness, kurtosis), three Hjorth parameters (activity, mobility, and complexity), the Higuchi fractal dimension (HFD), and Shannon entropy. This procedure yielded 266 features per epoch (19×14). The

features were provided as input to a fully connected ANN consisting of four hidden layers with 50, 40, 30, and 20 neurons, respectively. Each hidden layer was followed by batch normalization, ReLU activation, and dropout with a rate of 0.5. The network terminated with a softmax output layer for classification.

Training and Validation

A subject-level cross-validation protocol was employed to generate training and test splits for model development and evaluation. To this end, both cohorts (AD and CN) were randomly partitioned into seven mutually exclusive subsets. Each AD subset was then paired with a randomly selected CN subset, creating seven combined subsets, each comprising 5 AD and 4 CN participants. These subsets were subsequently organized into seven distinct training–testing pairs for cross-validation. In each pair, six subsets served as the training set, while the remaining subset was used exclusively as the test set. Consequently, each training set contained 54 participants (9 participants per subset $\times$ 6 subsets). Considering that each participant contributed 52 or 55 trials (depending on the EEG signal segmentation scheme), the training set comprised 2808 (2970) trials, with 1560 (1650) AD trials and 1248 (1320) CN trials. On the other hand, each test set included 9 participants from one subset, yielding 468 (495) trials, with 260 (275) AD and 208 (220) CN trials. This subject-level cross-validation strategy ensured strictly subject-independent evaluation, preventing any overlap between training and testing participants. Before model training, z-score normalization was applied to all input representations. To avoid data leakage, normalization parameters were computed exclusively from the training data of each fold and subsequently applied unchanged to the corresponding validation and test subsets.

For each cross-validation fold, a separate model was trained using the Adam optimizer with a learning rate of 0.0001. Model performance was monitored on a validation subset comprising 15% of the training data, and the weights corresponding to the lowest validation loss were retained. Training was conducted for a maximum of 100 epochs, with early stopping based on validation loss using a patience of five epochs. The batch size was set to 64, and categorical cross-entropy was used as the optimization objective. Prior to each epoch, the training data were randomly shuffled to reduce sensitivity to trial ordering. All models were implemented in PyTorch and trained on an NVIDIA GPU.

Model performance was evaluated on the corresponding test subset using overall accuracy (ACC), balanced accuracy (BA), and class-specific precision, recall, and F1-score metrics. Precision and recall were computed as:

$$\text{Precision} = \frac{TP}{TP+FP}, \tag{1}$$

$$\text{Recall} = \frac{TP}{TP+FN}, \tag{2}$$

where TP, FP, and FN denote true positives, false positives, and false negatives, respectively. The F1-score was calculated as:

$$F1 = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{3}$$

Overall accuracy was defined as:

$$\text{ACC} = \frac{TP+TN}{TP+TN+FP+FN}, \tag{4}$$

where TN denotes the number of true negatives. Balanced accuracy was calculated as the mean recall obtained for the two classes:

$$\text{BA} = \frac{\text{Recall}_{AD} + \text{Recall}_{CN}}{2}. \tag{5}$$

In our experiments, the AD class was treated as the positive class and the CN class as the negative class. Except for overall accuracy and balanced accuracy, all performance metrics were calculated separately for each class, enabling detailed assessment of the model's ability to correctly identify both AD and CN participants.

LIME for Averaged Saliency Maps

To gain insight into the decision-making process of the trained networks, we employed the *averagedLIME* method. LIME itself is a post-hoc explainability technique designed to provide human-interpretable insights into the predictions of black-box machine-learning models. The method operates under the assumption that while complex models may be globally non-linear and difficult to interpret, they can often be approximated locally by simpler, interpretable models, such as linear classifiers or decision trees [38]. Although applicable to a wide range of data types, LIME initially gained popularity for explaining predictions made by models trained on tabular data [39]. Over time, however, it has been increasingly adopted for interpreting the decisions of models processing image inputs, such as CNNs [40]. This shift in application has led to adaptations of the LIME framework that are tailored to the specific requirements of image data and convolutional models, enabling localized explanations for individual predictions based on visual inputs.

In essence, LIME explains a CNN model's decision for a chosen image by examining how the model responds to controlled variations of that image. To generate an explanation for a specific prediction, LIME perturbs the input image in its local neighborhood by generating a set of similar samples, typically through random perturbations. These perturbations are guided by binary masks that selectively occlude or retain different image segments, enabling the creation of new instances that differ slightly from the original image. The CNN model is then queried to obtain predictions for these perturbed samples. Using the resulting dataset of perturbed images and their corresponding model outputs, LIME fits a simple, interpretable model that mimics the behavior of the complex model in the neighborhood of the instance being explained.

The weights of this surrogate model are then interpreted as importance scores, reflecting the influence of each image segment on the model's prediction. By mapping these weights back onto the corresponding regions of the original image, LIME constructs a saliency map that visually highlights the areas most influential for the classification decision, thereby offering an interpretable explanation of the CNN's prediction.

Although the saliency maps generated by LIME provide local explanations of individual model decisions, in this study we extend their use toward identifying class-level patterns learned by the network. To transition from individual explanations to a more general interpretation, saliency maps are aggregated separately for each class following the averagedLIME approach proposed in [4]. This aggregation yields class-specific maps highlighting regions of the input representation that consistently influence the model's decisions, thereby providing an intuitive summary of the discriminative patterns encoded by the CNN.

The aggregation procedure employed in averagedLIME can be formally described as follows. Let $L_i(u,v)$ denote the local LIME explanation generated for sample i, and let p_i denote the probability assigned by the CNN to the class represented by that sample. For a given class c, the aggregated class-level saliency map is computed as a confidence-weighted average:

$$G_c(u,v) = \frac{\sum_{i=1}^{N_c} p_i L_i(u,v)}{\sum_{i=1}^{N_c} p_i}, \tag{6}$$

where N_c denotes the number of samples belonging to class c, $G_c(u,v)$ represents the resulting class-level saliency map, and u and v denote spatial coordinates within the feature map. The confidence weighting scheme reduces the influence of uncertain predictions and assigns greater importance to explanations associated with samples for which the classifier expressed higher confidence.

Aggregation of local explanations reduces sample-specific variability. While individual saliency maps may exhibit substantial fluctuations, averaging suppresses stochastic effects and enhances structures that recur consistently across samples. Consequently, the resulting class-level maps provide a more stable representation of the discriminative patterns learned by the model than individual explanations.

Complementary Explanation Methods

In addition to the averagedLIME framework, two complementary explanation methods, namely SHAP [41] and Grad-CAM [42], were employed to provide independent perspectives on CNN decision-making and to facilitate comparison of the resulting explanation patterns.

SHAP uses the concept of Shapley values from cooperative game theory to quantify the contribution of individual input features to the final model prediction. In this framework, the prediction is interpreted as a payoff distributed among features according to their marginal contributions, yielding attribution scores that reflect the importance of individual features for the model output. Formally, for a model f with input vector $\mathbf{x} = (x_1, x_2, \ldots, x_M)$ consisting of M features, the Shapley value ϕ_i for the i-th feature is given by:

$$\phi_i(f,\mathbf{x}) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(M-|S|-1)!}{M!} \left(f(S \cup \{i\}) - f(S)\right), \tag{7}$$

where $N = \{1, 2, \ldots, M\}$ is the set of all features and S denotes a subset of features excluding feature i. The term $f(S)$ represents the expected model output when only the features in subset S are known (and the others are marginalized). The weighting factor ensures that all possible coalitions of features are treated fairly.

In the context of CNNs, the input features correspond to individual pixels (or superpixels, if segmentation is applied) of the input image. The SHAP algorithm estimates the contribution of each pixel/superpixel to the model's output, allowing visualization of which image regions are most relevant for a given prediction. These contributions can be arranged into a saliency map $\Phi(u, v)$, whose elements correspond to attribution scores assigned to individual pixels/superpixels. The contributions are computed relative to a reference or background distribution, which represents the expected model output in the absence of informative input. This background can be defined in several ways, including: (i) using a random subset of training samples, (ii) computing the mean or median across the dataset, or (iii) generating representative centroids via clustering methods such as k-means.

In practice, the exact computation of Shapley values is combinatorially expensive, as it requires evaluating all possible subsets of features. SHAP addresses this challenge by introducing approximation strategies that are model-specific. For deep learning models, the DeepSHAP algorithm leverages backpropagation and linearization techniques to efficiently approximate Shapley values while preserving key theoretical properties. These approximations make SHAP feasible for high-dimensional data, such as EEG signals or images.

Grad-CAM [42] generates class-specific saliency maps by computing gradients of the target class score with respect to the feature maps of the final convolutional layer. These gradients quantify the importance of individual feature maps for a given prediction and are subsequently aggregated to obtain feature-map importance weights. For a target class c, the weights are computed as:

$$\alpha_k^c = \frac{1}{Z} \sum_u \sum_v \frac{\partial y^c}{\partial A^k(u, v)}, \tag{8}$$

where y^c denotes the prediction score for class c, A^k is the k-th feature map of the final convolutional layer, and Z represents the number of spatial locations within the feature map. The final Grad-CAM saliency map is then obtained as:

$$M_c(u, v) = \mathrm{ReLU}\left(\sum_k \alpha_k^c A^k(u, v) \right). \tag{9}$$

The ReLU operation suppresses negative contributions and retains only features that support the prediction of class c. Consequently, the resulting saliency map highlights regions whose activation contributes positively to the model decision.

To facilitate comparison with averagedLIME and SHAP, Grad-CAM maps were generated for all EEG trials and subsequently averaged separately within each diagnostic class, yielding class-level saliency maps.

Evaluation of Explanation Quality and Reliability

Because the proposed framework aims to identify stable cross-subject discriminative patterns rather than explain individual predictions only, additional analyses were performed to evaluate the quality, reliability, robustness, and interpretability of the obtained explanations. The evaluation comprised several complementary analyses, including assessment of explanation reproducibility across independently trained models, consistency between different explanation methods, generalizability to previously unseen data, sensitivity to methodological choices, faithfulness to the underlying model decisions, and relationships between attribution patterns and physiological signal characteristics.

The first stage of the analysis examined reproducibility of the extracted patterns across the seven subject-independent cross-validation folds. Class-level explanation maps generated from independently trained models were quantitatively compared to determine whether similar spatial–spectral structures emerged despite differences in training data composition. Similarity between explanation maps was quantified using Pearson correlation, Spearman correlation, and Intersection over Union (IoU), capturing complementary aspects of explanation consistency. Pearson correlation evaluates agreement in attribution magnitudes between corresponding regions of two explanation maps:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}, \tag{10}$$

where x_i and y_i denote corresponding saliency values from the compared maps.

Spearman correlation evaluates whether regions are assigned similar relative importance, regardless of the exact attribution values:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \tag{11}$$

where d_i denotes the difference between ranks assigned to the i-th pair of corresponding regions in the two explanation maps, and n denotes the total number of regions included in the comparison.

Finally, to assess agreement of the most strongly highlighted regions, explanation maps were transformed into binary masks and compared using Intersection over Union (IoU):

$$IoU = \frac{|A \cap B|}{|A \cup B|}, \tag{12}$$

where A and B denote binary masks derived from the compared explanation maps. In all analysis IoU was computed using binary masks formed from the top 10% attribution values. Together, these metrics provide complementary information regarding agreement in attribution magnitude, relative importance ranking, and localization of salient regions.

The second stage of evaluation examined consistency of the identified patterns across different explanation techniques. Agreement between averagedLIME, SHAP, and Grad-CAM was assessed to determine whether similar spectral–spatial structures were identified by methods based on substantially different explanatory principles. High agreement would suggest that the observed patterns are not specific to a particular explanation algorithm but reflect more general characteristics of the learned representation.

Next, explanation generalizability was evaluated through train–test consistency analysis. Class-level explanation maps generated from training samples were compared with corresponding maps obtained from previously unseen test data in order to assess whether the identified patterns remained detectable beyond the data used to construct the explanations.

Robustness of the averagedLIME framework with respect to methodological choices was additionally evaluated through a parameter sensitivity analysis. Saliency maps obtained under alternative perturbation settings were compared with those generated using the reference configuration in order to determine whether the identified patterns depended strongly on a particular choice of explanation parameters.

The subsequent stage of evaluation focused on faithfulness. Whereas the preceding analyses assessed the consistency, generalizability, and robustness of explanations, they did not determine whether highly attributed regions were genuinely involved in the model decision process. To address this question, a perturbation-based faithfulness analysis was performed. For each explanation map, a binary mask corresponding to the top 10% highest attribution values was generated. The identified regions were subsequently removed from the input representation by setting the corresponding pixels to zero, and the modified samples were re-evaluated by the CNN. If the highlighted regions were genuinely important for the prediction, their removal should lead to a reduction in model confidence. The resulting confidence change was computed as:

$$\Delta p = p_{\text{before}} - p_{\text{after}}, \tag{13}$$

where p_{before} and p_{after} denote the predicted probability of the true class before and after perturbation, respectively.

Finally, the relationship between model explanations and underlying signal characteristics was investigated by comparing attribution patterns with group-level spectral differences between AD and cognitively normal participants. For this purpose, average spectral representations were first computed separately for the two groups and subsequently subtracted to obtain a spectral power difference map:

$$\Delta P = P_{\text{AD}} - P_{\text{CN}}, \tag{14}$$

where P_{AD} and P_{CN} denote the average spectral representations of the Alzheimer's disease and cognitively normal groups, respectively. The resulting map highlights frequency–channel regions exhibiting the largest physiological differences between the two groups. Correlation analysis was then performed between these spectral difference maps and the corresponding attribution difference maps in order to evaluate the extent to which discriminative model patterns were related to group-level spectral power alterations.

For all analyses involving comparison of explanation maps, similarity was quantified using the Pearson correlation, Spearman correlation, and IoU measures described above.

In addition to explanation analysis, statistical comparisons between the evaluated CNN architectures were performed using non-parametric methods. A Friedman test was first applied to assess overall differences between models. When significant effects were observed, pairwise Wilcoxon signed-rank tests with Bonferroni correction were subsequently used to identify specific model pairs responsible for the observed differences.

Results

Models' Performance

We began the verification of our approach by constructing a set of 28 deep learning models designed to discriminate between AD and CN participants, comprising seven models for each of the four tested architectures: timeCNN, EEGNet, fftCNN, and ANN, following the data partitioning procedure described in the "Alzheimer Dataset and Data Partitioning" subsection. The resulting classification performance metrics, reported separately for both classes, are summarized in Table 3.

As can be seen in Table 3, the individual models achieved overall accuracies ranging from approximately 71% to 98%, with notable differences in stability and class-specific performance. Among the compared approaches, the CNN operating on frequency-domain representation (fftCNN) yielded the highest mean classification accuracy of 85.59%, outperforming both time-domain models, timeCNN (82.42%) and EEGNet (80.75%), and the ANN model based on engineered features (80.41%). The fftCNN also exhibited balanced performance across both diagnostic categories, reaching 87.57% precision and 86.79% recall for AD, and 83.94% precision with 84.10% recall for CN. This pattern indicates that the spectral CNN generalizes well across subjects and does not exhibit strong bias toward either class. The best individual fftCNN (Model 4) achieved an accuracy of 95.81%.

Although the average accuracy of timeCNN (82.42%) was close to that of fftCNN, class-specific metrics revealed that it did not maintain a consistent bias across runs: some models achieved higher accuracy (recall) for AD (up to 97.45%), while others favored CN, indicating that the discriminative temporal features captured by the network varied substantially between training folds. This variability suggests that temporal representation may be less robust in capturing stable disease-related patterns compared to spectral ones.

EEGNet, although achieving only a slightly lower average accuracy (80.75%) than timeCNN (82.42%), exhibited a considerably higher asymmetry between classes, with mean recall values of 74.60% for AD and 88.44% for CN. However, unlike timeCNN, which alternated between favoring AD and CN across repetitions, EEGNet consistently maintained a recall bias toward the CN class across all trained models. This stability suggests that, although EEGNet operated directly on time-domain signals, its architectural design enabled it to capture patterns relevant to AD. In particular, the use of temporal convolutional kernels tuned to physiologically meaningful EEG rhythms allowed the network to extract a subset of spectral–spatial patterns associated with AD-related activity.

Finally, the ANN model, trained on hand-crafted features, not only achieved the lowest average accuracy (80.41%) but also exhibited the highest variability across repetitions (71.43–98.73%). While one model (Model 4) reached near-perfect performance (98.73%), others performed substantially lower (approximately 70–75%), indicating the strong dependence of this architecture on subject partitioning and random initialization. The standard deviations of 14.49% and 16.64% for recall in the AD and CN classes, respectively, further illustrate the model's instability.

Overall, fftCNN achieved the highest average classification accuracy and the most balanced class-wise performance within the studied dataset, while maintaining relatively low variability across repetitions. These findings suggest that, for this cohort, the most discriminative information for differentiating AD from CN in resting-state EEG is primarily expressed in the frequency domain.

The superior performance of the spectral CNN representation may be partially explained by its closer correspondence to electrophysiological alterations commonly associated with Alzheimer's disease. Previous studies have consistently reported characteristic spectral changes in AD [43, 44, 45], including a slowing of neural activity manifested by increased power in lower-frequency bands together with reductions in alpha- and beta-band activity. Such alterations are naturally represented in the frequency domain and may therefore provide a more informative basis for classification than temporal EEG signals alone.

In addition to the physiological relevance of spectral features, the employed representation itself may facilitate the learning process. The spectral images preserve frequency-domain information for each EEG channel and arrange channel-specific spectra in a consistent order within a two-dimensional input space. Consequently, convolutional filters can learn patterns that jointly involve frequency content and channel-dependent activity, enabling the network to capture spectral differences associated with Alzheimer's disease across multiple recording sites. Such a representation may provide a more informative basis for classification than purely temporal signal representations. Consequently, the observed performance differences should be interpreted as effects of the complete analysis pipelines, encompassing both the adopted EEG representation and the classifier architecture, rather than as consequences of architecture alone.

To further assess the clinical relevance of the obtained results, an additional subject-level evaluation was performed. For each participant, class probabilities predicted for all EEG segments were averaged and a single diagnostic label was assigned

Table 3. Performance metrics obtained for each of the 7 trained models for each of the 4 tested architectures. The results are reported separately for both classes. All metric values are expressed as percentages.

Model	Accuracy	Precision$_{AD}$	Recall$_{AD}$	F1$_{AD}$	Precision$_{CN}$	Recall$_{CN}$	F1$_{CN}$
			timeCNN				
Model 1	84.44	**93.04**	77.82	84.75	76.98	**92.73**	84.12
Model 2	72.73	78.23	70.55	74.19	67.21	75.45	71.09
Model 3	90.71	92.25	90.91	91.58	88.84	90.45	89.64
Model 4	**92.12**	89.33	**97.45**	**93.22**	**96.41**	85.45	**90.60**
Model 5	75.76	85.71	67.64	75.61	67.99	85.91	75.90
Model 6	79.39	87.45	73.45	79.84	72.35	86.82	78.93
Model 7	81.82	78.46	92.73	85.00	88.24	68.18	76.92
Mean	**82.42**	**86.35**	**81.51**	**83.45**	**79.72**	**83.57**	**81.03**
Std	6.71	5.59	11.08	6.82	10.63	8.05	6.78
			EEGNet				
Model 1	81.21	85.27	80.00	82.55	76.79	82.73	79.65
Model 2	76.57	90.36	64.73	75.42	67.45	91.36	77.61
Model 3	88.48	91.92	**86.91**	89.35	**84.68**	90.45	87.47
Model 4	**89.09**	93.33	86.55	**89.81**	84.58	92.27	**88.26**
Model 5	77.98	**97.16**	62.18	75.83	67.40	**97.73**	79.78
Model 6	76.36	82.64	72.73	77.37	70.36	80.91	75.26
Model 7	75.56	84.07	69.09	75.85	68.40	83.64	75.26
Mean	**80.75**	**89.25**	**74.60**	**80.88**	**74.24**	**88.44**	**80.47**
Std	5.36	4.98	9.33	5.94	7.22	5.68	4.98
			fftCNN				
Model 1	84.07	80.98	93.21	86.67	89.53	72.64	80.21
Model 2	86.37	86.23	89.81	87.99	86.57	82.08	84.26
Model 3	75.05	78.08	76.60	77.33	71.43	73.11	72.26
Model 4	**95.81**	95.54	**96.98**	**96.25**	**96.15**	94.34	**95.24**
Model 5	87.84	**96.83**	80.75	88.07	80.08	**96.70**	87.61
Model 6	88.89	92.06	87.55	89.75	85.33	90.57	87.87
Model 7	81.13	83.27	82.64	82.95	78.50	79.25	78.87
Mean	**85.59**	**87.57**	**86.79**	**87.00**	**83.94**	**84.10**	**83.76**
Std	6.03	6.80	6.68	5.41	7.47	9.14	6.89
			ANN				
Model 1	72.70	94.95	53.71	68.61	62.50	96.43	75.84
Model 2	73.33	76.00	76.00	76.00	70.00	70.00	70.00
Model 3	71.43	68.24	90.86	77.94	80.49	47.14	59.46
Model 4	**98.73**	**98.31**	**99.43**	**98.86**	**99.28**	**97.86**	**98.56**
Model 5	80.32	85.99	77.14	81.33	74.68	84.29	79.19
Model 6	75.24	75.66	81.71	78.57	74.60	67.14	70.68
Model 7	91.11	88.08	97.14	92.39	95.90	83.57	89.31
Mean	**80.41**	**83.89**	**82.29**	**81.96**	**79.64**	**78.06**	**77.58**
Std	9.76	10.19	14.49	9.53	12.46	16.64	12.05

Table 4. Subject-level classification performance reported as mean ± standard deviation across the seven subject-independent test folds.

Model	Accuracy	Balanced Accuracy	Recall_{AD}	Recall_{CN}
timeCNN	87.30±14.95	87.50±14.79	85.71±13.98	89.29±13.34
EEGNet	84.13±10.84	85.00±10.70	77.14±13.80	92.86±12.20
fftCNN	87.30±10.00	86.79±10.58	91.43±10.69	82.14±18.90
ANN	82.54±15.53	82.14±15.97	85.71±15.12	78.57±22.49

according to the resulting mean probability vector. Subject-level performance metrics (Table 4) were then calculated using these aggregated predictions, with each participant represented by one predicted label and one reference diagnosis.

The subject-level analysis revealed that all neural-network models benefited from aggregation across EEG segments. The largest improvement was observed for timeCNN, whose accuracy increased from 82.42% at the segment level to 87.30% at the subject level. This result suggests that, although individual segment predictions were occasionally incorrect, the model produced highly consistent probability estimates across segments belonging to the same subject, allowing aggregation to recover the correct diagnosis. In contrast, fftCNN exhibited only a modest increase in accuracy (from 85.59% to 87.30%). Rather than indicating inferior performance, this result suggests that the spectral model already produced highly reliable segment-level predictions, leaving less room for improvement through aggregation. Importantly, both timeCNN and fftCNN achieved the same subject-level accuracy of 87.30%, despite the substantially higher segment-level accuracy of fftCNN. This observation suggests that the spectral representation provides more discriminative information at the level of individual EEG segments, whereas the time-domain model relies more strongly on averaging across multiple segments from the same subject.

To evaluate the statistical significance of performance differences between models, a Friedman test was conducted using subject-level balanced accuracy values obtained in the seven cross-validation folds. The test did not reveal significant differences between the evaluated models ($\chi^2 = 1.129$, $p = 0.772$). Consequently, pairwise Wilcoxon signed-rank tests were performed for completeness and likewise showed no statistically significant differences between any pair of models (all $p > 0.29$).

These findings should be interpreted in the context of the relatively small number of independent observations ($n = 7$ folds). Although fftCNN achieved the highest mean classification accuracy and one of the highest balanced accuracy values, the available sample size does not provide sufficient statistical power to detect significant differences between models. Nevertheless, fftCNN consistently exhibited strong performance across folds, achieved the highest segment-level accuracy, maintained balanced performance between diagnostic classes, and showed comparatively low variability across repetitions, which collectively motivated its selection for the subsequent explainability analyses.

AveragedLIME explanation analysis

To extract the most meaningful patterns distinguishing the AD and CN cohorts, the explainability analysis focused on the best-performing model, namely Model 4 of the fftCNN architecture. AveragedLIME was employed to identify the frequency–channel regions that contributed most strongly to the model's predictions and to reveal the characteristic EEG patterns associated with each diagnostic group.

To obtain class-level explanations, LIME was applied to all EEG trials from the training dataset. For each trial, a saliency map was generated using 2000 perturbations of rectangular superpixels of size 5×2, corresponding to one EEG channel and 0.5 Hz along the frequency axis. The number of perturbations was selected as a compromise between explanation stability and computational cost. The resulting saliency maps were subsequently averaged separately within the AD and CN classes, yielding two global saliency maps that summarize the spectral–channel regions most influential for the network's decisions.

Figure 2 presents representative local LIME explanations for five randomly selected EEG trials from each class together with the corresponding class-averaged saliency maps. For each trial, the upper row shows the input spectral image, whereas the lower row presents the associated LIME explanation. The last column contains the average input image and the corresponding averagedLIME saliency map computed across all training trials belonging to the given class.

As can be observed, substantial variability exists both in the input representations and in the corresponding local saliency maps. This variability reflects differences in spectral characteristics between individual EEG segments and results in explanation patterns that are often highly localized and trial-specific. While individual explanations provide insight into the network's decisions for particular samples, they do not readily reveal the global characteristics distinguishing the AD and CN groups. In contrast, averaging explanations across a large number of trials suppresses sample-specific fluctuations and highlights the spectral–channel regions that consistently contribute to the model's predictions. Consequently, the averagedLIME maps provide a more stable and interpretable representation of the decision patterns learned by the network.

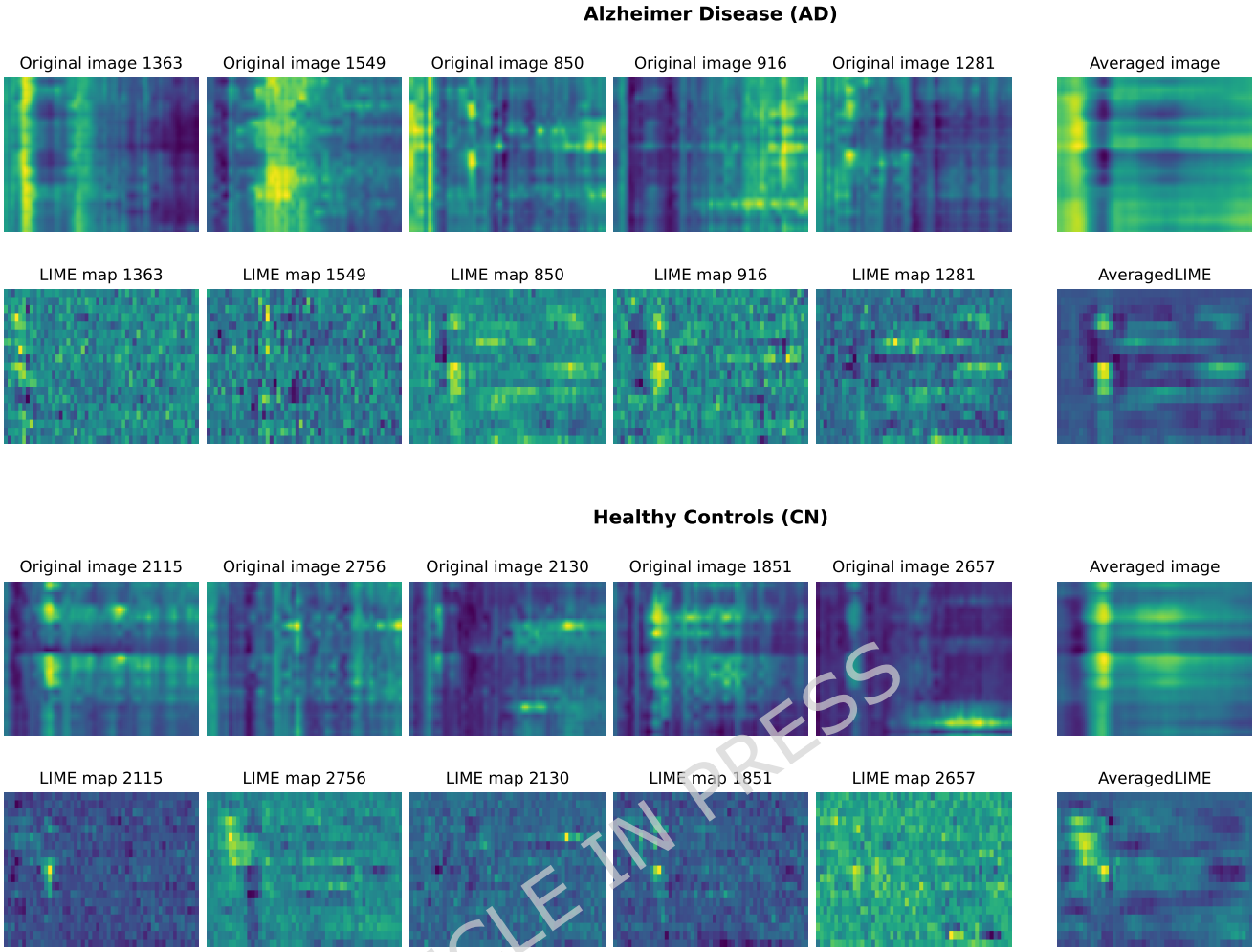

Figure 2. Representative local LIME explanations and corresponding averagedLIME saliency maps for the AD and CN classes. For each class, five randomly selected EEG trials are shown together with their original spectral representations (top row) and corresponding LIME saliency maps (bottom row). The last column presents the class-averaged input image and the averagedLIME saliency map.

The visual comparison presented in Fig. 2 suggests that averaging explanations substantially reduces the variability observed in individual LIME maps and reveals more coherent class-level patterns. However, the resulting averagedLIME maps could still be influenced by the particular composition of the training data used to generate them. To evaluate whether the identified spectral–channel structures represent reproducible characteristics learned by the model rather than artifacts of a specific train–test split, pairwise similarity between class-level saliency maps obtained independently across cross-validation folds was quantified using Pearson correlation, Spearman correlation, and Intersection over Union (IoU).

For the Alzheimer's disease class, averagedLIME explanations demonstrated substantial consistency across folds, achieving mean Pearson and Spearman correlations of (0.704 ± 0.068) and (0.605 ± 0.084), respectively. The overlap of the most salient regions reached an average IoU of (0.472 ± 0.073). Lower but still moderate consistency was observed for the normal control class, with Pearson and Spearman correlations of (0.520 ± 0.099) and (0.482 ± 0.085), respectively, and an average IoU of (0.298 ± 0.069).

These results indicate that the major spectral–channel structures identified by averagedLIME were recurrent across independently trained models and were not solely determined by a particular fold composition. The higher stability observed for the AD class suggests that the disease-related patterns learned by the network were more consistent across subjects and cross-validation splits. In contrast, the lower agreement observed for the CN class is likely associated with greater physiological heterogeneity among healthy individuals. Together with the visual inspection presented in Fig. 2, these findings support the central assumption of averagedLIME: although individual explanations may be highly variable, aggregation reveals robust

explanatory patterns that persist across independently trained models.

While inter-fold stability assesses whether explanations remain consistent across independently trained models, it does not address another important aspect of explanation reliability: sensitivity to the perturbation parameters used by the explanation algorithm itself. This issue is particularly relevant for LIME, whose explanations are generated by fitting a local surrogate model to randomly perturbed versions of the analyzed image. Consequently, the resulting attribution maps may potentially depend on the number of generated perturbations or on the size of the perturbed regions.

To evaluate the robustness of averagedLIME explanations with respect to these design choices, an additional sensitivity analysis was performed. The reference explanations were generated using 2000 perturbations and a perturbation grid of size (5×2). Subsequently, new explanations were computed using alternative perturbation settings, including different numbers of perturbations (1000 and 3000) and different perturbation grid sizes ((3×1) and (7×3)). Similarity between the resulting class-level saliency maps and the reference maps was quantified using the same metrics as employed in the stability analysis, namely Pearson correlation, Spearman correlation, and IoU.

Because the preceding stability analysis demonstrated that averagedLIME explanations were highly reproducible across independently trained models, repeating the sensitivity analysis for every cross-validation fold would provide largely redundant information while substantially increasing computational cost. Therefore, the sensitivity study was performed using the best-performing classifier (Model 4 of the fftCNN architecture), whose explanations served as the reference for all subsequent analyses.

The results of the sensitivity analysis are presented in Table 5. For the perturbation number analysis, averagedLIME explanations exhibited a remarkably high degree of robustness. Pearson and Spearman correlations exceeded 0.96 for both classes even when the number of perturbations was reduced by 50% relative to the reference setting (1000 vs. 2000 perturbations). Increasing the number of perturbations by 50% (3000 vs. 2000 perturbations) produced only marginal improvements, with correlation coefficients approaching unity. Similarly, IoU values remained high, increasing from 0.76–0.85 for 1000 perturbations to 0.92–0.96 for 3000 perturbations. These findings indicate that the principal attribution patterns identified by averagedLIME are already highly stable at 1000 perturbations and are largely unaffected by substantial changes in perturbation sampling density.

In contrast, modifying the perturbation grid size had a considerably stronger influence on the resulting explanations. Increasing the grid size to 7×3 reduced Pearson correlations to approximately 0.80 while maintaining relatively high Spearman correlations (0.77–0.79), suggesting that the general ranking of important regions was preserved despite changes in attribution magnitude and localization. A similar pattern was observed for the smaller 3×1 grid, where Pearson correlations decreased further (0.42–0.46), whereas Spearman correlations remained comparable to those obtained for the larger grid. These results indicate that the precise spatial distribution of attribution values depends on the selected superpixel size, although the relative importance ordering of the identified regions remains substantially more stable.

Taken together, the results suggest that averagedLIME explanations are highly robust to the number of perturbation samples but moderately sensitive to the spatial granularity of the perturbation grid. Combined with the previously demonstrated inter-fold stability, these findings support the conclusion that the major explanatory patterns identified by averagedLIME represent genuine properties of the trained model rather than artifacts of random perturbation sampling, while also highlighting the importance of selecting an appropriate superpixel size for detailed spatial interpretation.

Table 5. Sensitivity of LIME explanations to perturbation parameters. Metrics were computed between the reference explanations and explanations generated using modified perturbation settings.

	Pearson		Spearman		IoU	
Perturbation setting	AD	CN	AD	CN	AD	CN
1000 pert., 5×2	0.996	0.978	0.990	0.965	0.853	0.757
3000 pert., 5×2	0.999	0.991	0.997	0.988	0.961	0.924
2000 pert., 7×3	0.804	0.805	0.769	0.794	0.358	0.627
2000 pert., 3×1	0.463	0.421	0.790	0.781	0.474	0.508

Comparison and validation of explanation methods

The preceding section demonstrated that averagedLIME explanations were both stable across independently trained models and robust to perturbation parameter selection. An important remaining question is whether the identified explanatory patterns are specific to LIME or instead reflect more general properties of the trained classifier. To address this issue, two additional explanation methods, SHAP and Grad-CAM, were applied to the same model and dataset.

SHAP explanations were generated using DeepSHAP with a background distribution represented by 20 reference samples obtained through k-means clustering of the training data. This configuration was selected as a compromise between

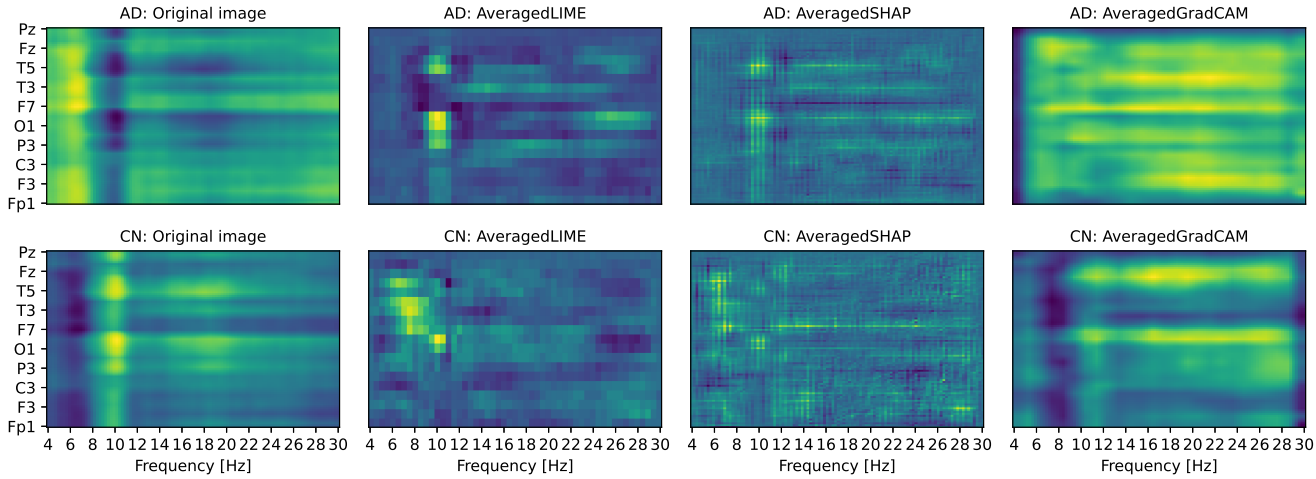

Figure 3. Global mean representations for the AD and CN classes. The columns show averaged input images, averaged LIME saliency maps, averaged SHAP saliency maps, and averaged Grad-CAM saliency maps, respectively.

computational efficiency and the ability to capture the variability of the training set while maintaining a manageable number of reference points for Shapley value estimation. Grad-CAM explanations were obtained from the final convolutional layer, where gradients of the target class score were used to estimate the importance of individual feature maps. The resulting weighted activation maps were subsequently projected back to the input space to obtain class-specific saliency maps highlighting spectral–spatial regions that contributed positively to the model prediction. As in the case of LIME, explanations were computed for all training trials and subsequently averaged within the AD and CN classes, yielding global class-level attribution maps for each explanation method. A visual comparison of the resulting explanations is presented in Fig. 3, which shows the averaged input representations together with the corresponding averagedLIME, SHAP, and Grad-CAM saliency maps for the AD and CN classes.

Visual inspection of the averaged input representations reveals clear differences between the AD and CN classes. In both groups, the dominant structures consist of two prominent vertical bands located approximately within the 4–6 Hz and 9–11 Hz frequency ranges. However, their relative prominence differs between classes. In the AD group, the lower-frequency component is more pronounced, particularly across frontal and temporal channels, whereas in the CN group greater emphasis is observed around 9–11 Hz, corresponding to the alpha frequency range. Additionally, the AD representation exhibits more pronounced horizontal structures extending across multiple channels within approximately 12–30 Hz, while the corresponding patterns are weaker and less spatially extended in the CN class.

These differences are reflected in the averagedLIME explanations, although the network does not simply focus on the strongest regions of the input images. For the CN class, the most salient attribution region is concentrated within approximately 5–9 Hz and is particularly visible over frontal and temporal channels. Interestingly, this region is shifted toward lower frequencies relative to the strongest alpha-band activity observed in the averaged input representation. In contrast, the AD explanations emphasize regions located around 8–10 Hz together with several localized vertical structures distributed across higher frequencies. The resulting attribution patterns differ substantially between classes.

The SHAP-based explanations largely confirm the regions identified by averagedLIME. Although the resulting maps appear more diffuse and less spatially localized, increased attribution values remain concentrated within similar frequency ranges for both classes. In particular, the low-frequency region highlighted by averagedLIME in the CN class remains visible in the SHAP maps, while the AD explanations preserve both the activity around 8–10 Hz and the broader structures observed at higher frequencies. This agreement between two conceptually different explanation techniques suggests that the identified patterns are unlikely to be artifacts of a particular attribution method.

Grad-CAM produced substantially different attribution patterns from those obtained using averagedLIME and SHAP. In particular, the prominent low-frequency regions identified by the perturbation-based methods were only weakly represented in the Grad-CAM maps. For the CN class, the localized activation around 5–9 Hz highlighted by averagedLIME and confirmed by SHAP was largely absent, while the dominant Grad-CAM activations were concentrated within broad horizontal structures extending across multiple frequency bands. Similarly, for the AD class, the localized vertical patterns observed around 8–10 Hz in both averagedLIME and SHAP were not clearly preserved in the Grad-CAM explanations. Despite these differences, Grad-CAM emphasized horizontal structures located within the same EEG channels highlighted by the two other

Table 6. Agreement between class-level explanation maps generated by averagedLIME, SHAP, and Grad-CAM.

Comparison	Pearson_{AD}	Spearman_{AD}	IoU_{AD}	Pearson_{CN}	Spearman_{CN}	IoU_{CN}
LIME-SHAP	0.641	0.549	0.271	0.379	0.277	0.156
LIME-GradCAM	-0.027	-0.147	0.026	-0.181	-0.112	0.071
SHAP-GradCAM	-0.052	-0.092	0.034	-0.142	-0.138	0.028

methods. These channel-specific patterns were considerably more pronounced in the Grad-CAM maps and extended across wider frequency ranges, particularly for the AD class. This observation suggests that while averagedLIME and SHAP primarily identify localized spectral regions contributing to the classification decision, Grad-CAM captures broader channel-level feature representations learned by the deeper convolutional layers of the network. Overall, averagedLIME and SHAP exhibit a high degree of agreement, consistently identifying similar spectral–channel regions despite differences in attribution smoothness and spatial localization. In contrast, Grad-CAM emphasizes broader channel-level structures and shows weaker correspondence with the localized frequency-specific patterns highlighted by the perturbation-based methods.

An important observation is that the salient attribution regions are not generally aligned with the areas of highest power visible in the averaged input representations. Instead, the explanation methods consistently assign high importance to regions corresponding to relative reductions in spectral power. This finding suggests that the classifier detected primarily deviations from dominant spectral activity rather than the strongest signal components themselves.

In addition to visual comparison, quantitative similarity analysis was performed between explanation techniques. Pairwise agreement between averagedLIME, SHAP, and Grad-CAM was quantified using Pearson correlation, Spearman correlation an IoU computed from the corresponding class-level saliency maps. The results are summarized in Table 6.

The quantitative agreement analysis largely confirmed the observations derived from visual inspection of the saliency maps. As shown in Table 6, the highest agreement was observed between averagedLIME and SHAP. For the AD class, Pearson and Spearman correlations reached 0.641 and 0.549, respectively, while for the CN class the corresponding values were lower (0.379 and 0.277). Similarly, the overlap of the most salient regions was substantially greater for the LIME–SHAP pair than for any comparison involving Grad-CAM, with IoU values of 0.271 and 0.156 for the AD and CN classes, respectively.

In contrast, comparisons involving Grad-CAM yielded near-zero or slightly negative correlation coefficients, indicating that the spatial organization of the attribution patterns identified by Grad-CAM differed substantially from those obtained using averagedLIME and SHAP. This observation is consistent with the visual comparison presented in Fig. 3, where Grad-CAM emphasized broad channel-level structures while the perturbation-based methods highlighted more localized spectral–channel regions.

Although the previous analysis quantified the similarity between different explanation methods, it did not reveal whether the resulting attribution patterns remained consistent across independently trained models. To investigate this aspect, an additional stability analysis was performed using class-level saliency maps obtained from the seven cross-validation folds. Pairwise similarity between folds was quantified separately for averagedLIME, SHAP, and Grad-CAM using the same evaluation metrics.

The inter-fold stability results are summarized in Table 7. Among the evaluated explanation techniques, averagedLIME exhibited the highest stability across independently trained models. For the AD class, mean Pearson and Spearman correlations reached 0.704 and 0.605, respectively, while the overlap of the most salient regions achieved an IoU of 0.472. Stability was lower for the CN class but remained moderate, with Pearson and Spearman correlations of 0.520 and 0.482 and an IoU of 0.298.

SHAP explanations demonstrated a similar but slightly weaker trend. For the AD class, Pearson and Spearman correlations reached 0.576 and 0.509, respectively, while the corresponding IoU was 0.329. Comparable values were obtained for the CN class, where Pearson and Spearman correlations equaled 0.473 and 0.430 and the IoU reached 0.243. These results indicate that the major explanatory patterns identified by SHAP remain reasonably consistent across folds, although with lower reproducibility than those obtained using averagedLIME.

Grad-CAM produced the least stable explanations. While moderate stability was still observed for the AD class (Pearson = 0.549, Spearman = 0.553), the overlap of the most salient regions was substantially lower (IoU = 0.225). The effect was even more pronounced for the CN class, where all stability measures decreased markedly, reaching Pearson and Spearman correlations of approximately 0.25 and an IoU of only 0.109. This finding suggests that Grad-CAM explanations are considerably more dependent on the particular train–test split than the two other methods.

A consistent pattern across all three explanation techniques was the higher stability observed for the AD class compared with the CN class. This trend indicates that disease-related attribution patterns are more reproducible across independently trained models, whereas explanations associated with healthy controls exhibit greater variability. Overall, the results demonstrate that averagedLIME provides the most reproducible class-level explanations, followed by SHAP, while Grad-CAM yields

Table 7. Inter-fold stability of class-level explanation maps generated by averagedLIME, SHAP, and Grad-CAM, quantified using Pearson correlation, Spearman correlation, and Intersection over Union (IoU). Values correspond to mean pairwise similarity between maps obtained from independent cross-validation folds.

Method	Pearson_{AD}	Spearman_{AD}	IoU_{AD}	Pearson_{CN}	Spearman_{CN}	IoU_{CN}
LIME	0.704	0.605	0.472	0.520	0.482	0.298
SHAP	0.576	0.509	0.329	0.473	0.430	0.243
GradCAM	0.549	0.553	0.225	0.253	0.254	0.109

Table 8. Train–test agreement of class-level explanation maps generated by averagedLIME, SHAP, and Grad-CAM. Values quantify the similarity between explanation maps obtained from the training and unseen test subsets of the representative cross-validation fold.

Method	Pearson_{AD}	Spearman_{AD}	IoU_{AD}	Pearson_{CN}	Spearman_{CN}	IoU_{CN}
LIME	0.899	0.849	0.591	0.514	0.525	0.232
SHAP	0.877	0.837	0.528	0.409	0.287	0.210
Grad-CAM	0.755	0.686	0.181	0.046	0.096	0.075

substantially less stable attribution maps across cross-validation folds.

Having established that the major attribution patterns are reasonably stable across independently trained models, the next question concerns their generalization beyond the training data. To evaluate this aspect, a train–test agreement analysis was performed using the best-performing fftCNN model (Model 4).

Class-level explanation maps were generated independently for the training and test subsets and subsequently compared using Pearson correlation, Spearman correlation, and IoU. This analysis assessed whether the spectral–channel structures identified as important during training were preserved when explanations were computed from previously unseen EEG segments. The corresponding train–test agreement results for all three XAI techniques are presented in Table 8.

Overall, the obtained similarities were higher than those observed in the inter-fold stability analysis, indicating that the principal explanatory patterns identified within the selected model generalize well from the training data to previously unseen EEG segments. Among the evaluated methods, averagedLIME exhibited the strongest train–test consistency. For the AD class, Pearson and Spearman correlations reached 0.899 and 0.849, respectively, while the overlap of the most salient regions achieved an IoU of 0.591. Lower values were obtained for the CN class, with Pearson and Spearman correlations of 0.514 and 0.525 and an IoU of 0.232. These results indicate that the major spectral–channel structures identified by averagedLIME are largely preserved when explanations are computed from unseen samples.

SHAP produced a similar pattern, yielding high train–test correlations for the AD class (Pearson = 0.877, Spearman = 0.837) and moderate agreement for the CN class (Pearson = 0.409, Spearman = 0.287). The overlap of the most salient regions was lower than for averagedLIME but remained non-negligible for both classes. Together with the previously observed agreement between averagedLIME and SHAP, these findings suggest that these two approaches capture explanatory patterns that generalize beyond the training set.

Grad-CAM demonstrated substantially weaker train–test consistency. Although moderate Pearson and Spearman correlations were obtained for the AD class (0.755 and 0.686, respectively), the overlap of the most salient regions remained low (IoU = 0.181). For the CN class, all similarity measures decreased markedly, reaching Pearson and Spearman correlations of only 0.046 and 0.096, respectively, with an IoU of 0.075. These results indicate that Grad-CAM explanations are considerably less reproducible across datasets than the perturbation-based methods.

Consistent with the previous agreement and stability analyses, higher train–test agreement was observed for the AD class than for the CN class across all three explanation techniques. This observation suggests that the disease-related attribution patterns learned by the classifier are more distinct and reproducible than those associated with healthy controls. Overall, the results further support averagedLIME as the most reliable explanation method among the evaluated approaches, with SHAP providing partially convergent evidence and Grad-CAM exhibiting substantially weaker generalization properties.

Agreement and generalization analyses indicated that the identified attribution patterns were reproducible across methods, folds, and datasets. However, reproducibility alone did not guarantee that the highlighted regions were genuinely involved in the classifier's decision-making process. To address this question, a perturbation-based faithfulness analysis was performed.

Faithfulness was quantified by masking the most salient regions identified by each explanation method and measuring the resulting decrease in prediction confidence. Larger confidence reductions indicate that the removed regions contribute more strongly to the model's decisions. The corresponding results are summarized in Table 9.

AveragedLIME and SHAP produced the largest confidence reductions for both classes. For the AD class, masking the

Table 9. Faithfulness evaluation based on confidence drop after masking the most salient regions identified by each explanation method. Results are reported separately for the Alzheimers disease (AD) and normal control (CN) classes.

Method	AD	CN
LIME	0.401±0.195	0.142±0.095
SHAP	0.384±0.229	0.159±0.096
GradCAM	0.035±0.075	0.002±0.046

regions identified by averagedLIME reduced prediction confidence by 0.401 ± 0.195, while SHAP produced a comparable reduction of 0.384 ± 0.229. Similar behavior was observed for the CN class, with confidence decreases of 0.142 ± 0.095 and 0.159 ± 0.096 for averagedLIME and SHAP, respectively. These results indicate that the regions highlighted by these two approaches are strongly associated with the classifier's predictions. In contrast, Grad-CAM exhibited substantially lower faithfulness scores. The average confidence reduction reached only 0.035 ± 0.075 for the AD class and 0.002 ± 0.046 for the CN class, indicating that masking the regions emphasized by Grad-CAM had only a limited effect on the model's output. This finding suggests that the broad attribution patterns produced by Grad-CAM are less directly related to the specific input regions driving the classification decision.

The faithfulness results were consistent with the previous agreement, stability, and train–test analyses. AveragedLIME and SHAP not only produced more reproducible explanations and higher cross-method agreement, but also identified regions whose removal substantially affected model predictions. Grad-CAM, in contrast, exhibited lower agreement with alternative explanation methods, reduced stability across folds, weaker train–test consistency, and markedly lower faithfulness scores.

The previous analyses collectively supported the reliability of the obtained explanations. The remaining question was whether the identified attribution patterns corresponded directly to the physiological differences observed between the AD and CN groups. To investigate this relationship, an attribution spectral difference analysis was performed.

For each explanation method, class-level attribution differences (AD–CN) were compared with the corresponding spectral power differences between the two groups. Pearson and Spearman correlations were computed to quantify the relationship between attribution strength and physiological signal differences. The results are summarized in Table 10.

Table 10. Correlation between attribution differences (AD–CN) and spectral power differences (AD–CN).

Method	Pearson	Spearman
LIME	-0.396	-0.209
SHAP	-0.340	-0.381
GradCAM	0.002	0.020

Moderate negative correlations were observed for both averagedLIME and SHAP. Pearson correlations reached -0.396 and -0.340 for averagedLIME and SHAP, respectively, while the corresponding Spearman correlations equaled -0.209 and -0.381. In contrast, Grad-CAM exhibited virtually no association with the spectral power differences, with both correlation coefficients remaining close to zero.

The observed negative correlations indicate that regions receiving high attribution values frequently did not correspond to locations exhibiting the strongest spectral power differences between AD and CN subjects. In many cases, the opposite tendency was observed, with highly attributed regions occurring in areas where group-level power differences were relatively weak or exhibited an opposite direction. Consequently, attribution maps cannot be interpreted as direct visualizations of disease-related spectral alterations.

This finding has important implications for interpretation of the obtained explanations. Attribution maps identify spectral–spatial regions that contribute to the classifier's discrimination between AD and CN subjects, whereas the underlying spectral representations indicate the direction and magnitude of physiological signal differences. Therefore, physiological interpretation requires joint analysis of both information sources. Considered in isolation, attribution maps reveal which regions are important for classification, but they do not indicate whether the associated spectral activity is increased or decreased in either group.

Together, these results suggest that the classifier does not primarily rely on the largest univariate spectral power differences between groups. Instead, it appears to exploit more complex spectral–channel relationships that become apparent only through the learned representation. This observation further supports the interpretation framework adopted in the present study, in which saliency maps are analyzed jointly with the corresponding averaged spectral representations rather than being interpreted as direct physiological activation maps.

Patterns' Analysis

Although individual local saliency maps shown in Fig. 2 vary considerably across trials, averaging the local explanations reveals stable class-level attribution patterns (Fig. 3). The previous analyses demonstrated that these patterns are reproducible across folds, generalize to unseen data, and are strongly associated with the classifier's decisions. The remaining question concerns their physiological interpretation.

To facilitate interpretation of the identified spectral–channel patterns, the attribution maps were summarized within three standard EEG frequency bands: theta (4–8 Hz), alpha (8–13 Hz), and beta (13–30 Hz). Mean band-specific attribution values were subsequently visualized as scalp maps for averagedLIME (Fig. 5), DeepSHAP (Fig. 6), and Grad-CAM (Fig. 7). For reference, Fig. 4 presents analogous scalp maps derived from the mean spectral power of the input representations.

Now, the spectral–spatial patterns driving the CNN decisions become more clearly discernible. In the global LIME saliency map for the AD class, the most prominent model sensitivity was localized within the alpha band, predominantly over occipital regions, with additional foci of increased relevance in the parietal and frontal areas. In contrast, reduced relevance was observed bilaterally across the frontotemporal regions, indicating that these locations contributed less to the network's AD-related discriminative features. In the beta band, enhanced model relevance was most evident over the left temporal area (around electrode T5), with significantly weaker effects extending to the right centro-temporal, occipital, and left frontal regions. Finally, in the theta band, both temporal lobes exhibited lowered contribution to the model's decision process. As expected for a binary classification problem, the relevance distribution associated with the CN class reflected an inverse pattern relative to that observed for AD, indicating that the network relied on complementary spectral–spatial features when discriminating between the two groups.

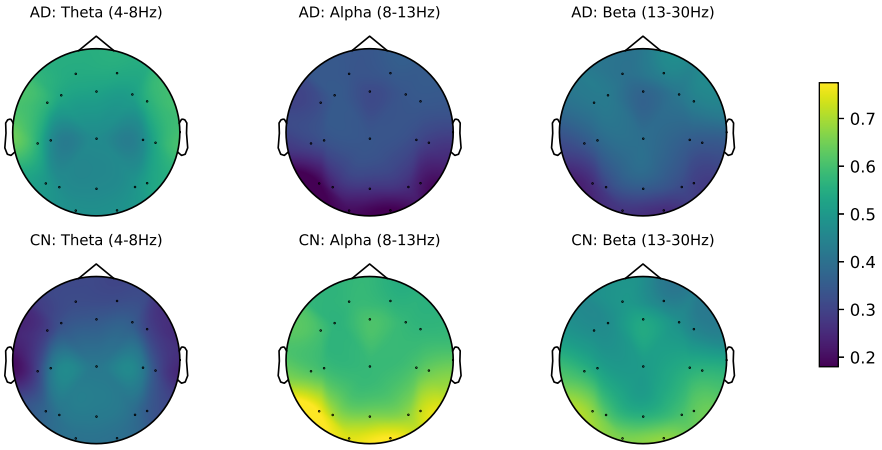

Figure 4. Scalp maps illustrating mean spectral power of the input images for the AD and CN groups across the analyzed frequency bands.

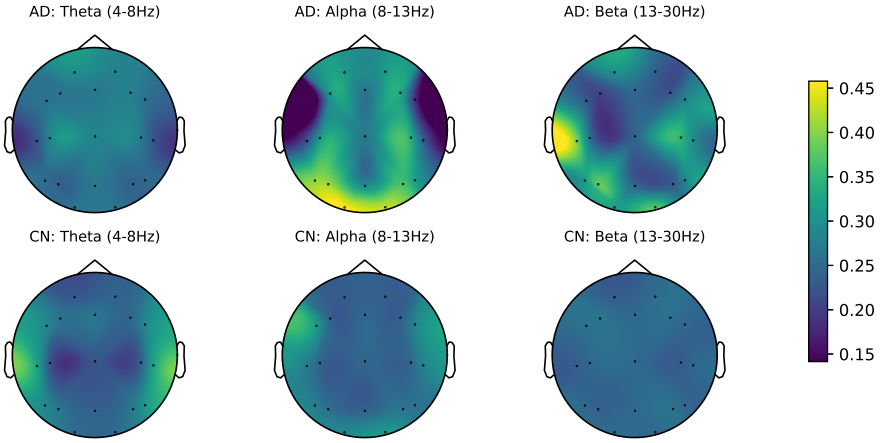

Figure 5. Scalp maps illustrating mean band-specific attribution values obtained from averagedLIME explanations for the AD and CN groups across the analyzed frequency bands.

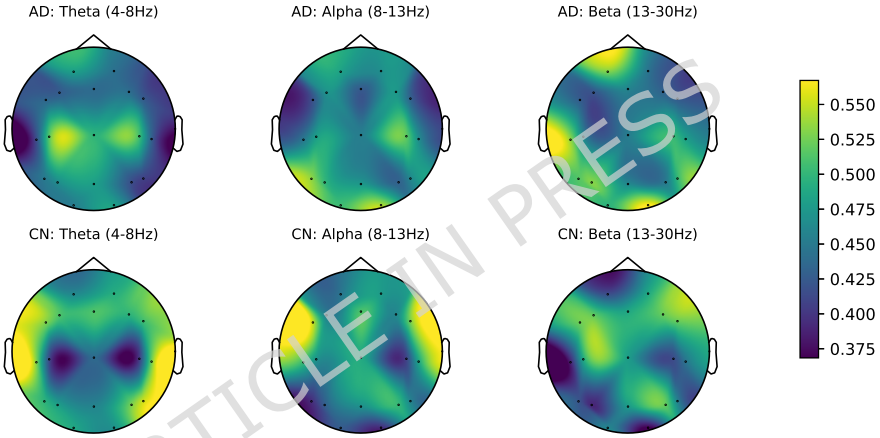

Figure 6. Scalp maps illustrating mean band-specific attribution values obtained from averaged SHAP explanations for the AD and CN groups across the analyzed frequency bands.

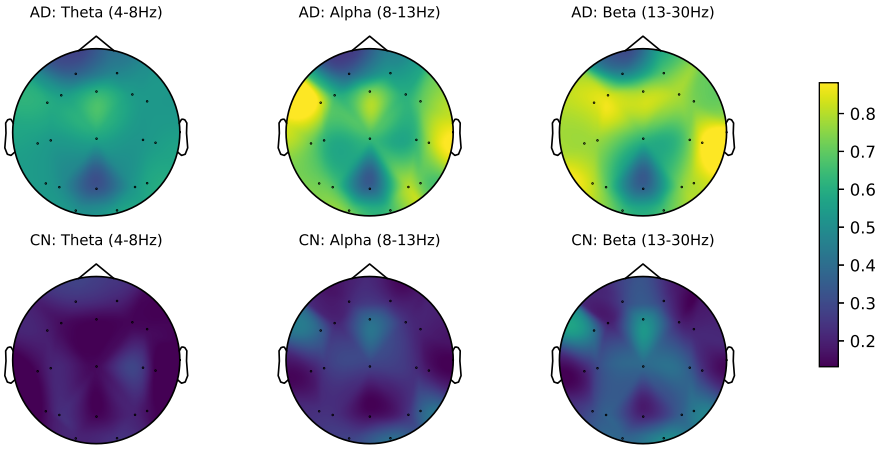

Figure 7. Scalp maps illustrating mean band-specific attribution values obtained from averaged Gram-CAM explanations for the AD and CN groups across the analyzed frequency bands.

The SHAP-derived maps exhibited spatial patterns largely consistent with those revealed by averagedLIME; however, these effects were substantially less apparent under standard visualization conditions (Fig. 3). In Figs. 4–6, each map is shown using its native color scale rather than a standardized one, because applying identical scaling would render the SHAP patterns nearly indistinguishable — the dynamic range of SHAP attributions is roughly half that of LIME. This approach allows clearer visualization of the frequency–spatial motifs identified by both methods, illustrating that, despite their lower contrast, SHAP maps corroborate the patterns highlighted by averagedLIME. The attenuation of SHAP activations becomes particularly evident in Fig. 3, where all saliency maps are displayed using a unified color scale, causing the SHAP-derived features to appear more diffuse and less distinct. Overall, this comparison underscores that SHAP confirms the same spectral–spatial tendencies as LIME, but with more subtle and distributed relevance.

The Grad-CAM maps revealed a markedly different attribution structure. Unlike averagedLIME and SHAP, which highlighted localized spectral–spatial regions, Grad-CAM produced broad and spatially diffuse activation patterns extending across large portions of the scalp. The most prominent activations were observed for the AD class, particularly within the alpha and beta bands. In contrast, the CN maps exhibited uniformly low attribution values across all three frequency bands. This behavior is consistent with the previous quantitative analyses. Grad-CAM demonstrated substantially lower agreement with both averagedLIME and SHAP, reduced stability across cross-validation folds, weaker train–test consistency, and markedly lower faithfulness scores. Consequently, although Grad-CAM broadly distinguishes between the AD and CN classes, the resulting patterns provide considerably less spatial specificity than those provided by the two other approaches. For this reason, the detailed interpretation presented in the remainder of this section focuses primarily on the spectral–spatial patterns consistently identified by averagedLIME and independently corroborated by SHAP.

At first glance, the spectral–spatial patterns identified by averagedLIME and corroborated by SHAP appear inconsistent with classical electrophysiological findings reported for Alzheimer's disease [5]. Resting-state EEG studies typically describe reduced alpha activity, particularly over posterior and frontal regions, together with increased low-frequency activity in the theta and delta bands [5, 6]. In contrast, the attribution maps presented in Figs. 5 and 6 exhibit pronounced alpha-related relevance for the AD class and enhanced theta-related relevance for the CN class.

This apparent discrepancy arises from the fundamentally different meaning of attribution maps and physiological signal representations. LIME and SHAP do not quantify neural activity or spectral power directly; instead, they identify the spectral–spatial features that most strongly influence the classifier's decision. Consequently, high attribution values should not be interpreted as increased EEG power in a given region or frequency band. Rather, they indicate that variations within these regions contribute substantially to discrimination between AD and CN subjects. In this sense, attribution maps represent decision evidence within the learned feature space rather than direct physiological activity.

Consequently, attribution maps alone are insufficient to determine whether a given spectral–spatial feature corresponds to increased or decreased activity in the AD group. To infer the direction of the underlying effects, the explanation maps must be interpreted jointly with the averaged input representations, which provide information about the corresponding group-level spectral power differences. When the attribution maps are examined together with the mean EEG representations (Fig. 4), a coherent set of disease-related patterns emerges. Specifically, the results suggest: (i) reduced alpha activity over occipital regions; (ii) a less pronounced alpha attenuation extending to parietal and frontal areas; (iii) decreased beta activity, most prominently over the left temporal region and, to a lesser extent, over right temporal, frontal, and occipital regions; and (iv) increased theta activity within lateral temporal areas. These patterns were consistently identified by averagedLIME and independently supported by SHAP.

Importantly, the resulting hypotheses are broadly consistent with previous EEG studies of Alzheimer's disease. Alpha attenuation has been consistently reported in resting-state EEG studies of Alzheimer's disease and is widely regarded as one of the characteristic electrophysiological manifestations of the disorder [10, 43, 44]. Previous studies have associated this phenomenon with cholinergic deficits and impaired neuronal synchronization accompanying disease progression [45]. Likewise, a shift of the EEG power spectrum toward lower frequencies, accompanied by reduced coherence of faster rhythms in Alzheimer's disease has been documented in previous studies [43, 46]. The reduction in beta-band activity observed in the present study, particularly over temporal regions, is also in agreement with previous reports describing decreases in alpha- and beta-band power in AD [47]. Furthermore, the combination of increased theta activity and reduced beta activity identified by the proposed framework resembles the characteristic spectral slowing frequently associated with disease progression [17].

At the same time, these findings should be interpreted with appropriate caution. Attribution maps indicate sensitivity of the classifier to discriminative spectral–spatial structures rather than direct neurophysiological mechanisms. Therefore, the identified patterns should not be regarded as validated biomarkers or definitive evidence of underlying disease processes. Instead, they should be viewed as candidate discriminative structures learned by the CNN that warrant further neuroscientific investigation and validation on independent datasets and populations.

Discussion and Limitations

The present study demonstrates that explainability methods can be used not only to visualize individual CNN decisions but also to identify stable class-level spectral–spatial patterns associated with Alzheimer's disease. By aggregating local explanations across large numbers of EEG segments, the proposed framework revealed recurrent attribution structures that were subsequently validated through multiple complementary analyses, including inter-fold stability, train–test agreement, cross-method comparison, and perturbation-based faithfulness evaluation. The convergence of these independent analyses suggests that the identified patterns reflect meaningful properties learned by the classifier rather than artifacts of a particular explanation method or dataset partition.

At the same time, interpretation of attribution maps requires careful consideration of their relationship to the underlying physiological signals. Explainability methods identify regions that contribute to model decisions, whereas physiological activity reflects neural processes occurring in the brain. Consequently, attribution values and spectral power represent fundamentally different quantities, and direct correspondence between them should not necessarily be expected. This distinction was supported by the observed negative correlations between attribution differences and spectral power differences, indicating that the classifier did not simply rely on regions exhibiting the largest group-level power changes. Instead, the model appeared to exploit more complex spectral–spatial relationships that became evident only through the learned feature representations of the network.

Despite the physiological plausibility of the identified patterns, they should not be interpreted as direct evidence of neurophysiological mechanisms or validated biomarkers. Rather, they represent candidate discriminative structures learned by the CNN within the analyzed dataset. The observed consistency with established findings, including alpha attenuation, increased prominence of lower-frequency activity, and reduced beta-band contributions, supports the neurobiological plausibility of these hypotheses. Nevertheless, model attention does not necessarily correspond to physiological truth, and attribution maps should therefore be viewed as tools for generating and prioritizing hypotheses rather than confirming underlying mechanisms.

An additional challenge arises from the intrinsic properties of resting-state EEG recordings. Unlike task-based paradigms, resting-state recordings exhibit substantial temporal variability, non-stationarity, and lack of event synchronization across subjects. Consequently, identical physiological processes may manifest differently between recordings due to spontaneous fluctuations in brain activity and individual variability. While these factors may limit the interpretability of individual saliency maps, the proposed framework partially addresses this challenge through aggregation within a common spectral–spatial representation. Nevertheless, variability between individuals, preprocessing procedures, and recording conditions may influence the extracted patterns and contribute to the performance fluctuations observed across cross-validation folds. Therefore, the identified structures should be regarded as exploratory findings rather than fixed physiological signatures.

Another important limitation of explanation-based analyses is that the identified patterns may not represent the full spectrum of disease-related differences present in the data. Neural networks are optimized to achieve accurate classification using only the subset of features necessary for decision making. Consequently, regions receiving low attribution values should not automatically be regarded as physiologically irrelevant. The patterns identified in the present study should therefore be interpreted as sufficient features for discrimination rather than an exhaustive characterization of Alzheimer's-related EEG alterations. This observation suggests several directions for future research. Repeated model training with different random initializations, as well as targeted ablation of highly attributed regions, may encourage the discovery of alternative discriminative structures and provide a more comprehensive view of the feature hierarchy learned by CNN models.

Beyond these conceptual limitations, several methodological limitations should also be acknowledged. First, all analyses were performed using a single publicly available dataset, and both classification performance and extracted attribution patterns may therefore be influenced by characteristics specific to the analyzed population, including demographic composition, disease-stage distribution, and participant selection procedures. Second, EEG recordings may vary substantially across acquisition systems, electrode configurations, recording protocols, and preprocessing pipelines. These factors may affect both model performance and the corresponding explanation maps. Third, SHAP explanations depend on the selected background distribution used for estimation of Shapley values. Although k-means centroids were employed to provide a compact approximation of the training-data distribution, alternative background definitions may produce different attribution patterns. Finally, the present work focused exclusively on CNN-based EEG analysis. Whether similar explanation aggregation mechanisms remain applicable to transformer-based architectures or traditional feature-based classifiers remains an open question.

From a practical perspective, the proposed framework may improve interpretability of CNN-based EEG analysis by providing visual representations of spectral–spatial structures contributing to model decisions. Such information could potentially support neurologists by offering complementary insight into discriminative signal characteristics rather than functioning as an autonomous diagnostic system. Future studies should therefore focus on external validation across heterogeneous datasets, evaluation of different disease stages and demographic groups, assessment of explanation robustness across alternative EEG modeling approaches, and prospective neuroscientific validation of the identified patterns. Ultimately, transformation of candidate discriminative structures into clinically meaningful biomarkers will require substantial additional validation. Until such

validation is performed, the proposed framework should be regarded as a research-oriented decision-support approach rather than a clinically validated diagnostic system.

Conclusion

The present study demonstrates that CNNs trained on resting-state EEG can serve not only as predictive models but also as interpretable frameworks for investigating the spectral–spatial features underlying their classifications. By transforming instance-level explanations into stable class-level saliency maps using averagedLIME, we extracted spectral–spatial patterns that were consistent with previously reported electrophysiological alterations associated with Alzheimer's disease while also revealing more specific candidate discriminative structures that warrant further investigation.

The obtained results suggest that the discriminative information learned by the network extends beyond simple group-level spectral power differences. While the aggregated explanations reproduced characteristic manifestations of EEG slowing, including posterior alpha attenuation, they also highlighted more localized spectral–spatial structures, such as asymmetric beta reductions and symmetric theta enhancements over temporal regions. Moreover, the observed inverse relationship between attribution patterns and spectral power differences demonstrated that the identified saliency maps cannot be interpreted as direct representations of physiological activity. Instead, meaningful interpretation requires joint analysis of the explanation maps and the corresponding averaged EEG representations.

Collectively, these findings highlight the methodological value of structured aggregation of local explanations. Rather than serving solely as a transparency tool, explainability can be used to identify stable cross-subject candidate discriminative patterns encoded in CNN models. The performed stability, generalization, and faithfulness analyses further suggest that such patterns are not merely artifacts of individual samples or model realizations, but reflect reproducible structures learned during classification.

Future work should investigate the robustness of the proposed framework across independent cohorts, alternative pre-processing pipelines, and diverse model architectures, as well as evaluate the clinical and neuroscientific relevance of the identified candidate patterns in larger and more heterogeneous populations.

References

1. Devi, A. N. & Latha, M. Dcnn-sbil: Eeg signal based mild cognitive impairment classification using compact convolutional network. *Expert. Syst. with Appl.* **273**, 126553 (2025).

2. Abinayaa, S. & Sridhar, S. An explainable hybrid bio-inspired feature selection framework using rsvp-evoked p300 eeg signals for identity authentication. *Expert. Syst. with Appl.* 129674 (2025).

3. Murugan, T. K. & Kameswaran, A. Employing convolutional neural networks and explainable artificial intelligence for the detection of seizures from electroencephalogram signal. *Results Eng.* **24**, 103378 (2024).

4. Rejer, I. & Gago, I. AveragedLime for general explanations in EEG domain. *NeuroImage* **323**, 121588 (2025).

5. Coben, L. A., Danziger, W. & Storandt, M. A longitudinal EEG study of mild senile dementia of Alzheimer type: changes at 1 year and at 2.5 years. *Electroencephalogr. Clin. Neurophysiol.* **61**, 101–112 (1985).

6. Canuet, L. *et al.* Resting-state network disruption and APOE genotype in Alzheimer's disease: a lagged functional connectivity study. *PLoS ONE* **7**, e46289, DOI: 10.1371/journal.pone.0046289 (2012).

7. Baik, K. *et al.* Implication of EEG theta/alpha and theta/beta ratio in Alzheimer's and lewy body disease. *Sci. Reports* **12**, 18706 (2022).

8. Fahimi, G., Tabatabaei, S. M., Fahimi, E. & Rajebi, H. Index of theta/alpha ratio of the quantitative electroencephalogram in Alzheimer's disease: a case-control study. *Acta Medica Iran.* 502–506 (2017).

9. David, M. C. *et al.* Pupil-linked arousal, cortical activity, and cognition in Alzheimer's disease. *Brain communications* **7** (2025).

10. Miltiadous, A. *et al.* A dataset of scalp EEG recordings of Alzheimer's disease, frontotemporal dementia and healthy subjects from routine EEG. *Data* **8**, 95 (2023).

11. Morabito, F. C. *et al.* Deep convolutional neural networks for classification of mild cognitive impaired and Alzheimer's disease patients from scalp EEG recordings. In *2016 IEEE 2nd International Forum on Research and Technologies for Society and Industry Leveraging a better tomorrow (RTSI)*, 1–6 (2016).

12. Ieracitano, C., Mammone, N., Bramanti, A., Hussain, A. & Morabito, F. C. A convolutional neural network approach for classification of dementia stages based on 2D-spectral representation of EEG recordings. *Neurocomputing* **323**, 96–107 (2019).

13. Huggins, C. J. *et al.* Deep learning of resting-state electroencephalogram signals for three-class classification of Alzheimer's disease, mild cognitive impairment and healthy ageing. *J. Neural Eng.* **18**, 046087 (2021).

14. Xia, W., Zhang, R., Zhang, X. & Usman, M. A novel method for diagnosing Alzheimer's disease using deep pyramid CNN based on EEG signals. *Heliyon* **9** (2023).

15. Kachare, P. *et al.* LCADNet: a novel light CNN architecture for EEG-based Alzheimer disease detection. *Phys. Eng. Sci. Medicine* **47**, 1037–1050 (2024).

16. Nour, M., Senturk, U. & Polat, K. A novel hybrid model in the diagnosis and classification of Alzheimer's disease using EEG signals: Deep ensemble learning (DEL) approach. *Biomed. Signal Process. Control.* **89**, 105751 (2024).

17. Stefanou, K. *et al.* A novel CNN-based framework for Alzheimer's disease detection using EEG spectrogram representations. *J. Pers. Medicine* **15**, 27 (2025).

18. Fouad, I. A. & Labib, F. E.-Z. M. Identification of Alzheimer's disease from central lobe EEG signals utilizing machine learning and residual neural network. *Biomed. Signal Process. Control.* **86**, 105266 (2023).

19. Amini, M., Pedram, M. M., Moradi, A. & Ouchani, M. Diagnosis of Alzheimer's disease by time-dependent power spectrum descriptors and convolutional neural network using EEG signal. *Comput. Math. Methods Medicine* **2021**, 5511922 (2021).

20. Calub, G. I. A. *et al.* EEG-based classification of stages of Alzheimer's disease (AD) and mild cognitive impairment (MCI). In *2023 5th international conference on bio-engineering for smart technologies (BioSMART)*, 1–6 (IEEE, 2023).

21. Siuly, S., Alçin, Ö. F., Wang, H., Li, Y. & Wen, P. Exploring rhythms and channels-based EEG biomarkers for early detection of Alzheimer's disease. *IEEE Transactions on Emerg. Top. Comput. Intell.* **8**, 1609–1623 (2024).

22. Amezquita-Sanchez, J. P., Mammone, N., Morabito, F. C., Marino, S. & Adeli, H. A novel methodology for automated differential diagnosis of mild cognitive impairment and the Alzheimer's disease using EEG signals. *J. neuroscience methods* **322**, 88–95 (2019).

23. Kulkarni, N. & Bairagi, V. Extracting salient features for EEG-based diagnosis of Alzheimer's disease using support vector machine classifier. *IETE J. Res.* **63**, 11–22 (2017).

24. Santos Toural, J. E., Montoya Pedrón, A. & Marañón Reyes, E. J. Classification among healthy, mild cognitive impairment and Alzheimer's disease subjects based on wavelet entropy and relative beta and theta power. *Pattern analysis applications* **24**, 413–422 (2021).

25. Alvi, A. M., Siuly, S. & Wang, H. A long short-term memory based framework for early detection of mild cognitive impairment from EEG signals. *IEEE Transactions on Emerg. Top. Comput. Intell.* **7**, 375–388 (2022).

26. Sidulova, M., Nehme, N. & Park, C. H. Towards explainable image analysis for Alzheimer's disease and mild cognitive impairment diagnosis. In *2021 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, 1–6 (IEEE, 2021).

27. Lopes, M., Cassani, R. & Falk, T. H. Using CNN Saliency Maps and EEG Modulation Spectra for improved and more interpretable machine learning-based Alzheimer's disease diagnosis. *Comput. Intell. Neurosci.* **2023**, 3198066 (2023).

28. Khan, W. *et al.* An explainable and efficient deep learning framework for EEG-Based diagnosis of Alzheimer's and frontotemporal dementia. *Front. Medicine* **12**, 1590201 (2025).

29. Montavon, G., Samek, W. & Müller, K.-R. Methods for interpreting and understanding deep neural networks. *Digit. Signal Process.* **73**, 1–15 (2018).

30. Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K. & M"uller, K.-R. *Explaining Deep Neural Networks and Beyond* (Springer, 2021).

31. Ras, G., Xie, N., van Gerven, M. & Doran, D. Explainable deep learning for eeg analysis: a review. *IEEE Rev. Biomed. Eng.* **15**, 29–44 (2022).

32. Molnar, C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* (Christoph Molnar, 2022), 2 edn.

33. Lundberg, S. M., Erion, G., Chen, H. *et al.* From local explanations to global understanding with explainable ai for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).

34. Ribeiro, M. T., Singh, S. & Guestrin, C. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32 (2018).

35. Jasper, H. The ten-twenty electrode system of the international federation. *Electroencephalogr Clin Neurophysiol* **10**, 367–380 (1958).

36. Lawhern, V. J. *et al.* EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *J. neural engineering* **15**, 056013 (2018).

37. Liu, K., Yang, M., Yu, Z., Wang, G. & Wu, W. FBMSNet: A filter-bank multi-scale convolutional neural network for EEG-based motor imagery decoding. *IEEE Transactions on Biomed. Eng.* **70**, 436–445 (2022).

38. Ribeiro, M. T., Singh, S. & Guestrin, C. Why should I trust you? explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144 (2016).

39. Garreau, D. & von Luxburg, U. Looking deeper into tabular LIME. *arXiv preprint arXiv:2008.11092* (2020).

40. ÖNLER, E. Explaining image classification model (CNN-based) predictions with LIME. *World J. Adv. Eng. Technol. Sci.* **7**, 275–280 (2022).

41. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Adv. neural information processing systems* **30** (2017).

42. Selvaraju, R. R. *et al.* Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 618–626 (2017).

43. Jeong, J. EEG dynamics in patients with Alzheimer's disease. *Clin. neurophysiology* **115**, 1490–1505 (2004).

44. Cassani, R., Estarellas, M., San-Martin, R., Fraga, F. J. & Falk, T. H. Systematic review on resting-state EEG for Alzheimer's disease diagnosis and progression assessment. *Dis. markers* **2018**, 5174815 (2018).

45. Papaliagkas, V. The role of quantitative EEG in the diagnosis of Alzheimer's disease. *Diagnostics* **15**, 1965 (2025).

46. Smailović, U. *et al.* Neurophysiological markers of Alzheimer's disease: a review. *Neurol. Ther.* **8**, 51–76 (2019).

47. Koelewijn, L. *et al.* Alzheimer's disease disrupts alpha and beta-band resting-state oscillatory network connectivity. *Clin. Neurophysiol.* **128**, 2347–2357 (2017).

Acknowledgements

This study was financially supported by the Minister of Science under the 'Regional Initiative of Excellence' (RID) programme.

Author contributions statement

I.R. conceived the study, developed the methodology, performed the experiments, analysed the data, and wrote the original draft. I.G. contributed to investigation, visualization, and writing – original draft and review. V.M. contributed to formal analysis, supervision, and manuscript review. All authors reviewed the manuscript.

Additional information

Accession codes: The EEG dataset used in this study, *A dataset of EEG recordings from: Alzheimer's disease, Frontotemporal dementia and Healthy subjects* [10], is publicly available via the OpenNeuro repository.
Code availability: The code used for preprocessing, model training, XAI analysis, and averagedLIME aggregation will be made publicly available in an open-source GitHub repository upon acceptance of the manuscript.
Data availability: Data will be made available on reasonable request.
Competing interests: The authors declare no competing interests.