Research Square

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

FAT-Net: A Frequency-Aware Temporal Graph Network Recovering Clinically Valid Dementia Signatures Where Aggregation May Bias Gradient-Based XAI

Mahshad Faramarzi
Kharazmi University
Mohammad Fathi
Kharazmi University
Mohammadreza Gholami
Kharazmi University
Ahmad Haj Reza Jafarabadi
Kharazmi University
Mir Mohsen Pedram
pedram@khu.ac.ir

Kharazmi University

Research Article

Keywords: Spatiotemporal graph neural network, Dynamic brain connectivity, Granger causality, Alzheimer's disease, Frontotemporal dementia, Explainable AI

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11208459/v1

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

Additional Declarations: No competing interests reported.

FAT-Net: A Frequency-Aware Temporal Graph Network Recovering Clinically Valid Dementia Signatures Where Aggregation May Bias Gradient-Based XAI

Mahshad Faramarzi[1†], Mohammad Fathi[2†], Mohammadreza Gholami[1], Ahmad Haj Reza Jafarabadi[1], Mir Mohsen Pedram[1,2*]

[1]Machine Learning & Cognitive Modeling Lab, Faculty of Engineering, Kharazmi University, Tehran, Iran.
[2]Department of Electrical and Computer Engineering, Faculty of Engineering, Kharazmi University, Tehran, Iran.

*Corresponding author(s). E-mail(s): pedram@khu.ac.ir;
Contributing authors: m.faramarzi@modares.ac.ir;
mohammadsadegh.fathi@khu.ac.ir;
Mohammadreza.gholami@modares.ac.ir; ajafarabadi@neurtex.com;
†These authors contributed equally to this work and share first authorship; they are listed alphabetically by last name.

Abstract

Distinguishing Alzheimer's disease (AD) and frontotemporal dementia (FTD) from healthy controls (HC) and from each other, remains challenging due to overlapping symptoms. Electroencephalography (EEG) provides a cost-effective view of neurodegeneration, yet conventional methods often overlook temporal and frequency-specific changes in brain connectivity. We propose a temporal multi-band graph-based framework for classifying AD and FTD from resting-state EEG while identifying informative frequency bands and disease-relevant brain regions. EEG data from 88 subjects were analyzed, with temporal graphs constructed using Granger causality, retaining connections above the 95th percentile. Nodes were represented by 18 statistical, spectral, and information-theoretic features, and the proposed graph neural network architecture employed GraphSAGE for spatial feature extraction and Bidirectional LSTM with temporal attention

to capture connectivity dynamics. To integrate frequency-specific information, a multi-band attention layer fused alpha and beta bands. Model performance was evaluated using subject-level stratified 5-fold cross-validation repeated 10 times. The model achieved $\boldsymbol{F_1}$-scores of $\mathbf{0.8891 \pm 0.0224}$ for HC vs. FTD, $\mathbf{0.8646 \pm 0.0172}$ for HC vs. AD, and $\mathbf{0.8057 \pm 0.0193}$ for FTD vs. AD. Lobe-wise analysis identified the temporal lobe as the most discriminative region, nearly matching whole-brain performance. It also aligned well with established clinical evidence, whereas XAI methods showed weaker clinical alignment. These findings show that integrating spatiotemporal and spectral dynamics can support accurate dementia classification and provide neurobiological insight. Lobe-wise analysis further suggested that FTD is more regionally focused, whereas AD shows a more widespread pattern, while current graph-based XAI methods may require further refinement for reliable clinical interpretation.

Keywords: Spatiotemporal graph neural network, Dynamic brain connectivity, Granger causality, Alzheimer's disease, Frontotemporal dementia, Explainable AI

1 Introduction

Dementia describes a group of progressive neurodegenerative disorders that disrupt memory, reasoning, and behavior to a degree that interferes with daily functioning. Alzheimer's disease (AD) is the most frequently diagnosed form [1, 2], while frontotemporal dementia (FTD) is one of the main causes of dementia in middle age [3, 4]. As global life expectancy increases, the number of individuals affected by these disorders continues to rise, creating a growing demand for diagnostic tools that can support early detection and help distinguish between dementia subtypes [5, 6]. These conditions do not follow identical pathological pathways, and their clinical presentations often overlap, making differential diagnosis challenging [7–11].

Electroencephalography (EEG) remains one of the cheapest and most accessible techniques for studying brain dysfunction [12, 13]. Previous studies have shown activity and connectivity changes in both AD and FTD, such as shifts in dominant rhythms, alterations in synchrony, and disruptions in network organization across different brain regions [14]. These observations suggest that EEG carries informative markers the progression of neurodegenerative processes.

Machine learning and deep learning methods have gained increasing attention as tools for extracting clinically relevant information from neurophysiological signals [15, 16]. Applications of convolutional, recurrent, and transformer-based models have reported promising results across a range of neuroimaging modalities, including MRI, fMRI, PET, DTI, and EEG [16, 17]. Despite the progression of deep learning methods, many studies still model EEG channels as independent signals [18, 19] or convert them into two-dimensional representations, which limits their ability to reflect the brain's connectivity and spatial topology [18, 20]. Recent graph-based approaches have mainly focused on capturing interactions between brain regions using static connectivity representations [21–23]. These models offer a more structured view of neural activity, but important gaps still remain. Temporal changes in connectivity are often

only partially represented [23, 24], and frequency bands are usually analyzed separately, despite evidence that cross-band interactions carry meaningful information in dementia [25, 26].

Interpretability also remains underexplored, leading to limited insight into which cortical regions or connectivity patterns are most relevant [27–29]. To address these gaps, we develop a temporal, multi-band graph-based framework for classifying AD, FTD, and healthy control (HC) participants using resting-state, eyes-closed EEG. The analysis is performed on a publicly available dataset of 88 individuals: 36 diagnosed with AD, 23 with FTD, and 29 HC. EEG recordings were collected using a 19-channel setup following the 10–20 international system, with A1 and A2 placed on the mastoids as reference electrodes. All recordings were obtained according to the clinical protocol, with participants seated and their eyes closed. The dataset was preprocessed by the data collectors through a standardized preprocessing pipeline that included band-pass filtering, re-referencing, artifact correction using ASR [30, 31], and ICA-based removal of ocular and jaw-related components [32, 33].

A key element of our approach is the construction of temporal graphs that reflect the evolving patterns of brain connectivity across the recording. Continuous EEG signals are divided into fixed-length windows using a sliding-window. We evaluate several configurations, including window lengths of 4 and 6 seconds with either no overlap or 50% overlap. Each window is transformed into an individual graph, producing a time-ordered sequence that summarizes how connectivity patterns unfold. Node attributes—representing individual EEG channels—are defined by a feature vector selected through an XGBoost-based feature ranking procedure. The final set includes statistical features (minimum, maximum, mean, standard deviation, kurtosis, skewness, slope), information-theoretic measures (Shannon entropy, spectral entropy, SVD entropy), spectral properties (mean and relative power, Hjorth parameters, RMS, zero-crossing rate), and one network-derived attribute (hub score computed using the HITS algorithm).

In addition to this representation, we design a temporal graph neural network (GNN) pipeline that processes the sequence of graphs for each frequency band. GraphSAGE layers [34] extract spatial embeddings from successive graph snapshots, and bidirectional LSTM units capture temporal dependencies across the recording. A temporal attention mechanism identifies the segments that carry the most discriminative information, while a band-attention module integrates contributions from different frequency bands. The fused embedding is passed to a classifier that distinguishes AD and FTD patients from healthy participants.

Interpretability is incorporated through complementary analyses at both the lobe and node levels. Brain lobe-wise evaluation quantifies how restricting the model to frontal, central, parietal, temporal, or occipital channels affects performance. In addition, Integrated Gradients [35], Deep Lift [36], Smooth Grad (Saliency-based) [37] are applied to the graph embeddings to estimate each nodes's contribution over time. These analyses reveal which brain regions and frequency-specific patterns the model relies on for predictions, providing a link between algorithmic output and known neurophysiological alterations in dementia. The rest of the manuscript describes the dataset,

preprocessing procedures, and graph construction pipeline; presents the model architecture and training setup; and reports the classification and interpretability results. The study highlights how temporal graph modeling can leverage the structure of EEG recordings to support automated dementia analysis and provide insights into disease-related network alterations.

2 Materials and Methods

2.1 The EEG Dataset

The EEG dataset analyzed in this study includes 88 subjects, consisting of 36 patients diagnosed with Alzheimer's disease (AD group), 23 patients diagnosed with Frontotemporal Dementia (FTD group), and 29 healthy control subjects (HC group). The data consists of resting-state, eyes-closed scalp EEG recordings. The mean age was 66.4 ± 7.9 years for the AD group, 63.6 ± 8.2 for the FTD group, and 67.9 ± 5.4 for the HC group. Cognitive state was assessed using the Mini-Mental State Examination (MMSE), with average scores of 17.75 ± 4.5 (AD), 22.17 ± 8.22 (FTD), and 30 (HC).

Recordings were acquired using a Nihon Kohden EEG 2100 device with 19 scalp electrodes (Fp1, Fp2, F7, F3, Fz, F4, F8, T3, C3, Cz, C4, T4, T5, P3, Pz, P4, T6, O1, O2) placed according to the 10–20 international system, using a referential montage with Cz as the common reference. The sampling rate was 500 Hz. The dataset, fully described in [38, 39], was used in its preprocessed form. The preprocessing pipeline, performed by the data collectors, included a 0.5–45 Hz Butterworth band-pass filter, re-referencing to A1–A2, artifact removal using Artifact Subspace Reconstruction (ASR), and Independent Component Analysis (ICA) with ICLabel for rejecting eye and jaw artifact components.

Table 1 Demographic and cognitive characteristics of the EEG dataset.

Group	Subjects	MMSE (Mean ± SD)	Age (Mean ± SD)
AD	36	17.75 ± 4.5	66.4 ± 7.9
FTD	23	22.17 ± 8.22	63.6 ± 8.2
HC	29	30	67.9 ± 5.4

2.2 Graph analysis procedure

Our analysis pipeline consists of four primary phases: participant data acquisition and windowing, temporal graph construction, feature extraction, and graph classification with explanation, as depicted in Fig. 1. First, EEG signals recorded from each participant (Fig. 1a) are preprocessed and utilized to construct a series of temporal graphs for the Alpha and Beta frequency bands (Fig. 1b). Node-level features are then extracted and organized into structured embeddings (Fig. 1c). These graphs are subsequently input into a GraphSAGE-based network for classification (Fig. 1d). Finally,

a graph explanation phase is employed to identify and highlight the most influential nodes to the model's decisions, using two distinct explainability methods: a lobe-wise and gradient-based analysis via the Captum library. The following sections provide a detailed description of each phase.

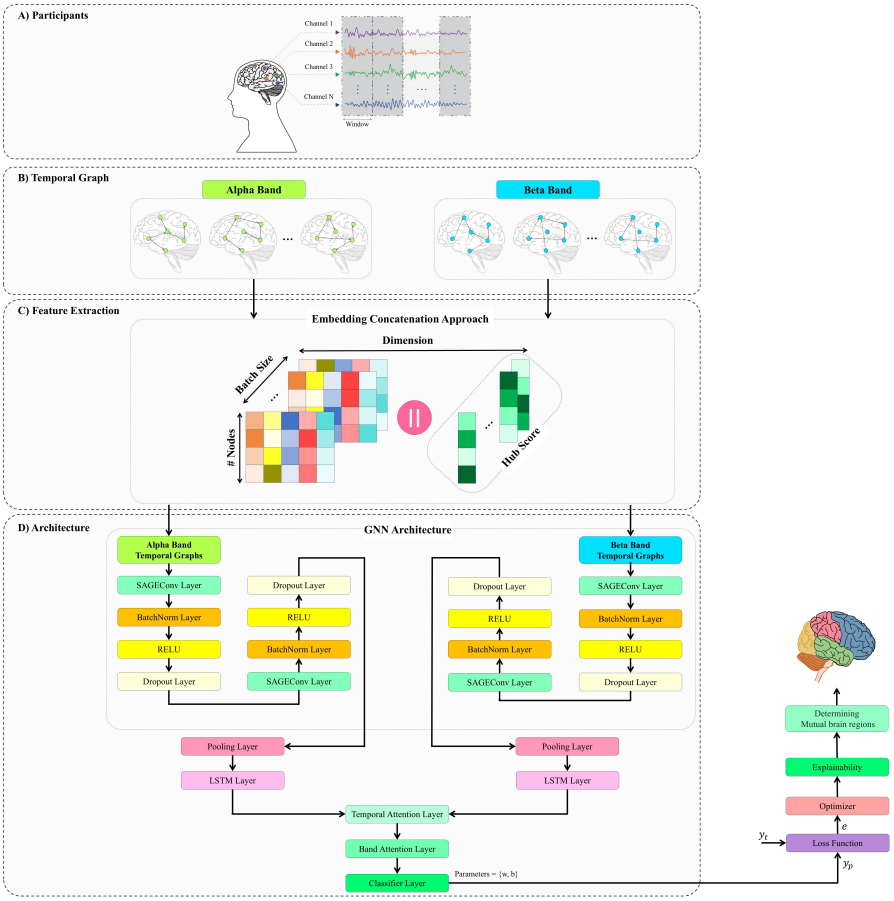


Fig. 1 Overview of the proposed pipeline. (a) EEG channels are recorded from participants and segmented into fixed-length windows. (b) Temporal graphs are constructed separately for the Alpha and Beta frequency bands. (c) Node-level features are extracted and concatenated, including the hub score derived from graph topology. (d) The extracted temporal graphs are passed through the band-specific GraphSAGE–BiLSTM–attention architecture, fused, and classified; the resulting predictions are subsequently analyzed through the explainability module to determine the brain regions most relevant to the classification decision.

2.3 Graph Extraction

Graph theory has become one of the essential analytical tools in neuroscience research, particularly for investigating brain connectivity patterns. Two primary connectivity types are typically analyzed: functional connectivity (FC), which measures statistical dependencies between brain regions, and effective connectivity (EC), characterizing the directional and causal influences among neural populations. In this study, we focused on EC analysis to understand information flow and causal dynamics within brain networks. A graph consists of nodes and edges, formally denoted as $G = (V, E)$, and can be characterized as directed or undirected depending on whether its edges possess directionality [40].

We developed temporal EEG-based graphs using Granger causality, a causal discovery approach that identifies directed edges between brain regions. The Granger causality analysis was performed on all channel pairs using a maximum time lag of 1, allowing rapid directional dependencies to be captured at the 500 Hz sampling rate. To maintain sparse network topology and preserve only robust connections, we retained edges above the 95th percentile threshold, computed independently for each subject and window to yield comparable edge density and normalize for inter-subject differences in raw connectivity magnitude, excluding self-connection. Graph structures can encompass three distinct association types: 1) causal associations (direct or indirect), 2) confounding associations (i.e., two variables have a common cause), and 3) noisy associations.

Methods based on simple correlation capture all of these relationship types without distinguishing among them, whereas causal discovery approaches such as Granger causality are intended to identify directed influences rather than purely statistical dependencies, although the reliability of these estimates depends on the estimator and model specification [41]. This approach allowed us to characterize the temporal dynamics and directional information flow within neural networks.

Our temporal graph framework involved partitioning continuous EEG data into fixed-length segments using a sliding window approach. We experimented with multiple configurations: window lengths of 4 seconds and 6 seconds, combined with overlap settings of 0% (non-overlapping windows) and 50% overlap. Each segment was transformed into an individual graph, and through sequential processing of the entire recording, we obtained a time-ordered series of graphs capturing the evolving patterns of brain connectivity. Node characteristics (representing individual EEG channels) were defined by a feature vector selected using an XGBoost-based feature importance analysis. Feature selection was performed once using a single 5-fold cross-validation procedure. Within each fold, feature importance was computed, and features that were consistently ranked as important across the five folds were selected based on their occurrence frequency. The resulting feature set was then fixed and used in all subsequent experiments, including the 10 repetitions of the 5-fold cross-validation used for model evaluation. The final node feature set encompassed: statistical features (minimum, maximum, mean, standard deviation, kurtosis, skewness, slope), information-theoretic measures (Shannon entropy, spectral entropy, SVD entropy), spectral characteristics (mean power spectral density, relative power, Hjorth activity,

mobility, and complexity parameters, root mean square, zero-crossing rate), and network topology metrics (hub score derived from the HITS algorithm). While several connectivity metrics (PLV, phase lag, spectral coherence, Pearson correlation, Granger causality, and partial correlation) were computed as potential edge features, they were not incorporated into the final model because the adopted GraphSAGE architecture aggregates information based on node features and graph topology rather than explicit edge features, thereby maintaining a simpler model and reducing computational cost.

2.4 Graph Classifier Network

To capture both spatial and temporal dynamics of brain connectivity, we developed a multi-band temporal GNN architecture that processes EEG data across different frequency bands while preserving temporal dependencies. Each subject's EEG recording was transformed into a structured sequence of temporal graphs. Specifically, for each subject, we generated a list of graphs representing consecutive time windows, creating a temporal sequence that captures the evolution of brain connectivity patterns over time.

This sequential structure was further organized by frequency band, resulting in each subject being represented as multiple parallel sequences, one for each frequency band. Each graph within these sequences contained node features (derived from the 18 selected features per EEG channel) and graph topology defined by Granger causality values, forming the foundation for our temporal graph learning framework. We experimented with various frequency band combinations, including individual bands and multiple band configurations. After systematic evaluation, we selected Alpha and Beta bands for final model training, as this combination achieved the highest accuracy and F_1-scores compared to other configurations.

2.4.1 Band-Specific Spatial-Temporal Processing

The core architecture consists of band-specific processing modules, each designed to handle the temporal sequence of graphs for a single frequency band. Before designing the final architecture, we first compared several GNN backbones, including GCN, GIN, Weighted-GATConv, GATConv, SAGEConv, and GraphSAGE, using a standard 20/80 train–test split. GraphSAGE achieved the highest and most stable performance among all architectures, therefore, it was selected as the core GNN layer of the proposed model. Specifically, we employed a two-layer GraphSAGE (Graph Sample and Aggregate) network to extract spatial features from individual graphs. The first GraphSAGE layer transforms the input node features (18-dimensional) into a 64-dimensional hidden representation, followed by batch normalization, ReLU activation, normalization layer, and dropout regularization. The second GraphSAGE layer further refines these spatial representations while maintaining the 64-dimensional feature space, again followed by batch normalization and ReLU activation. This dual-layer structure enables the network to capture both local neighborhood information and higher-order graph topological patterns within each time window. After spatial feature extraction, node embeddings of each graph were flattened and concatenated to form a single vector representation per time window.

These flattened representations were then fed into a bidirectional LSTM network (2 layers, 64 hidden units per direction) to capture temporal dependencies across the sequence of graphs. The bidirectional architecture allows the network to capture both forward and backward temporal patterns, resulting in 128-dimensional temporal features (64 from each direction). To assign more weight to important windows, we applied an attention mechanism over the LSTM outputs, computing attention weights via a linear layer followed by softmax normalization. The attention-weighted temporal representation serves as the band-specific embedding, encapsulating both spatial graph structure and temporal dynamics for that particular frequency band.

2.4.2 Multi-Band Attention Fusion

To integrate information across the Alpha and Beta frequency bands, we implemented an attention-based fusion mechanism. Each band-specific embedding (128-dimensional) was first projected through a learned linear transformation, and then band-level attention weights were computed using another linear layer followed by softmax activation. These attention weights determine the relative importance of each frequency band for classification, allowing the model to focus on the most discriminative spectral components. The final fused representation was obtained through a weighted combination of all band-specific embeddings according to their attention weights, followed by layer normalization. To encourage balanced utilization of both frequency bands and prevent the model from relying too heavily on a single band, we incorporated an L1 regularization term (with factor 0.05) that penalizes deviations from uniform band importance during training. The fused multi-band representation was then passed through a two-layer classification head consisting of fully connected layers with ReLU activation and dropout (rate = 0.5). The final layer projected the learned features into a 2-dimensional output space for binary classification (HC/AD, HC/FTD, and AD/FTD).

2.4.3 Training Strategy and Cross-Validation

Before training, all directed Granger causality graphs were converted to undirected graphs, as GraphSAGE is designed to operate on undirected graph structures. The resulting undirected graphs, together with the node features, were then used as inputs to the GraphSAGE classifier. We employed stratified 5-fold cross-validation, repeated 10 times to obtain robust and reliable performance estimates. The data splitting was performed at the subject level rather than the graph level, ensuring that all graphs from a single subject remained within the same fold. This subject-level splitting prevents data leakage, as temporal graphs from the same subject are inherently correlated and treating them as independent samples would artificially inflate performance metrics. The stratification strategy preserved the original class distribution across all folds, maintaining proportional representation of each class in both the training and validation sets. Within each fold, we addressed class imbalance by computing inversely proportional class weights, with an additional $1.5\times$ boost applied to the minority class.

These weights were incorporated into the cross-entropy loss function to prevent the model from learning to predict only the majority class and to ensure balanced learning across both classes. This weighted loss approach was essential for achieving meaningful

classification performance rather than trivial accuracy through biased predictions. To improve model stability and convergence, we implemented several optimization strategies. First, we tested three random initializations at the beginning of each fold and selected the one with the lowest validation loss. All linear layers were initialized using Xavier uniform initialization. Second, we used the AdamW optimizer with a learning rate of 0.0005 and weight decay of 1×10^{-3} for effective regularization. Third, we applied a cosine annealing learning rate scheduler with warm restarts ($T_0 = 10$) to facilitate better convergence. Additionally, gradient clipping (max norm = 0.5) was employed to prevent exploding gradients during backpropagation.

For each fold, we selected the best model based on validation F_1-score, with a minimum improvement threshold of 0.001 to avoid selecting models based on negligible fluctuations. Training was conducted for 100 epochs per fold, with the best-performing epoch saved based on F_1-score. We tracked comprehensive metrics including F_1-score, accuracy, precision, recall, loss values, and prediction confidence throughout training. By running the entire 5-fold cross-validation procedure 10 times, we obtained a total of 50 independent evaluations (5 folds $\times$ 10 repetitions). The reported performance metrics represent the average across all 10×5 folds, providing a robust estimate of model performance and reducing the impact of random variations in data splitting and model initialization. This evaluation framework ensures that our reported results reflect the model's generalization capability.

2.5 Model Explainability

Throughout training, we monitored the learned attention weights to analyze the relative importance of the Alpha and Beta frequency bands. By averaging attention weights across training epochs, cross-validation folds, and repetitions, we obtained interpretable measures of each band's contribution to the classification task, providing neuroscientific insights into which spectral components are most discriminative for the classification problem at hand. To gain insights into the neurobiological mechanisms underlying classification and identify which brain regions are most relevant for distinguishing between HC and patient groups (FTD and AD), we employed two complementary interpretability approaches: lobe-wise analysis and gradient-based using Captum, as illustrated in Fig. 2.

2.5.1 Lobe-Wise Analysis

We first conducted a systematic lobe-wise analysis to determine the relative importance of different brain regions for classification. EEG channels were grouped according to their anatomical locations into five major lobes: frontal, central, parietal, temporal, and occipital. For each lobe, we trained separate models using only the channels corresponding to that specific brain region, while maintaining the same network architecture and training protocol (stratified 5-fold cross-validation repeated 10 times). The performance of each model was evaluated using F_1-score, accuracy, precision, and recall, and compared with the whole-brain model that utilized all channels simultaneously. This approach allowed us to identify which brain regions carry the most discriminative

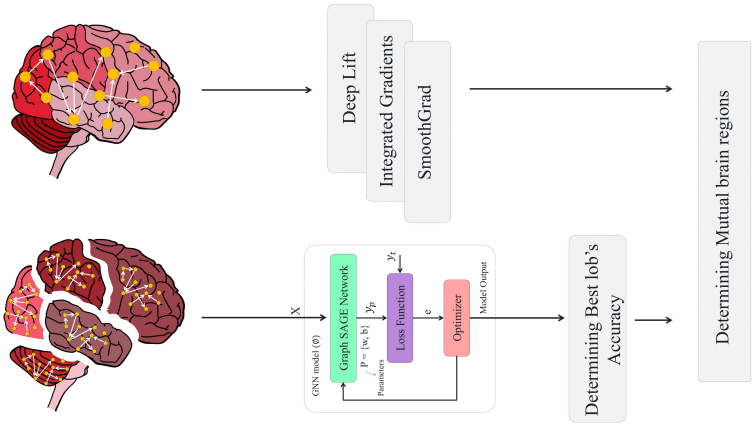


Fig. 2 Overview of the two complementary interpretability approaches used in this study. Top: gradient-based analysis, where node-level connectivity graphs are passed through three explanation methods (DeepLIFT, Integrated Gradients, and SmoothGrad) to estimate feature-level importance. Bottom: lobe-wise analysis, where lobe-specific sub-graphs are passed through the GraphSAGE network, and model accuracy is compared across lobes to determine the best-performing region. Both approaches converge on a final step that determines the mutual brain regions most relevant to classification.

information for differentiating HC from FTD and HC from AD. By comparing lobe-specific performance metrics to whole-brain performance, we could assess the relative contribution of each anatomical region to the classification task. Lobes with performance close to or matching that of the whole-brain model were considered particularly important, while those showing substantially lower performance suggested less critical roles in the classification.

2.5.2 Gradient-Based Analysis

As a complementary approach, we employed Captum's Integrated Gradients method, a gradient-based explainability framework, to identify important nodes, at the individual-subject level. The mathematical formulations of the XAI methods considered in this study are summarized in Table 2. Unlike the lobe-wise analysis, which evaluates regional importance across the entire dataset, Integrated Gradients provides instance-specific explanations by computing attribution scores for each input feature in a given subject's data. The method works by measuring how much each input feature contributes to the model's prediction. It computes the contribution of each feature by integrating gradients along a path from a baseline, or reference input, to the actual input. The baseline typically represents the absence of signal, such as zero-valued features or average feature values across the dataset. By calculating how the model's prediction changes as we gradually transition from this baseline to the actual input, Integrated Gradients assigns an importance score to each feature that reflects its contribution to the final prediction.

Table 2 Summary of explanation algorithms used to identify important nodes. Each method estimates node or feature importance via a distinct mathematical principle.

Method	Core concept	Key formulation(s)
DeepLIFT	Measures contributions relative to a reference input x^0	$C_i = (x_i - x_i^0) \cdot \left(\frac{\partial y}{\partial x_i}\right)_{\text{ref}}$ where x_i^0 is the reference feature value.
Integrated Gradients (IG)	Aggregates gradients along a path from baseline to input	$\text{IG}_i = (x_i - x_i') \cdot \int_0^1 \frac{\partial f(x' + \alpha(x - x'))}{\partial x_i} \, d\alpha$ where x' is the baseline input and $\alpha \in [0, 1]$.
SmoothGrad (Saliency-based)	Computes feature attributions by averaging saliency maps obtained from multiple noisy perturbations of the input, thereby reducing gradient noise and improving attribution stability.	$\text{SmoothGrad}_i(x) = \frac{1}{N} \sum_{k=1}^{N} \left\lvert \frac{\partial f(x + \epsilon_k)}{\partial x_i} \right\rvert, \quad \epsilon_k \sim \mathcal{N}(0, \sigma^2)$ where N is the number of noisy samples, ϵ_k is Gaussian noise added to the input, and σ is the noise standard deviation.

For our temporal graph-based architecture, we applied DeepLIFT, Integrated Gradients, and SmoothGrad to each subject's sequence of graphs in both the Alpha and Beta frequency bands. The overall procedure for computing subject-level, node-level, and lobe-level importance scores is summarized in Algorithm 1. The attribution scores were computed for each node's features at every time window, providing a comprehensive view of which nodes contributed most to the classification decision. To obtain an overall measure of node importance for each subject, we aggregated the attribution scores across temporal windows and frequency bands by taking their absolute values and averaging them. This aggregation process yielded a single importance score for each node, indicating how strongly that node influenced the model's prediction for that particular subject. To obtain group-level insights, we averaged the node importance scores across all subjects within each class, including HC, FTD, and AD. This allowed us to identify nodes that consistently showed high importance across multiple subjects. Furthermore, we aggregated node-level importance scores by anatomical lobe to enable direct comparison with our lobe-wise analysis results, providing convergent evidence for regional brain dysfunction patterns.

3 Results

In this section, we present the results, including the evaluation of the proposed architecture, model explainability, and ablation study.

3.1 Model's Performance

Table 3 demonstrates the model performance for the HC vs FTD, HC vs AD, and FTD vs AD classification tasks across different window-length and overlap configurations. For the HC vs FTD task, the best overall performance was obtained using WL=4 and

Algorithm 1 Band-wise Temporal Graph Processing and Attention-based Fusion

function ProcessBandSequence(*band_graphs*)
  Initialize an empty list *embeddings*.
  for each graph G_t in *band_graphs* **do**
    $x \leftarrow \text{GraphSAGE}_1(G_t)$
    $x \leftarrow \text{ReLU}(\text{BatchNorm}_1(x))$
    $x \leftarrow \text{LayerNorm}_1(x)$
    if training **then**
      $x \leftarrow \text{Dropout}(x)$
    end if
    $x \leftarrow \text{GraphSAGE}_2(x)$
    $x \leftarrow \text{ReLU}(\text{BatchNorm}_2(x))$
    if training **then**
      $x \leftarrow \text{Dropout}(x)$
    end if
    Append Flatten(x) to *embeddings*.
  end for
  $T \leftarrow \text{Stack}(\textit{embeddings})$
  $H \leftarrow \text{BiLSTM}_{\text{2-layer}}(T)$
  $H \leftarrow \text{LayerNorm}_2(H)$
  $e_t \leftarrow \text{Linear}_{\text{temp-attn}}(H_t)$
  $\alpha_t \leftarrow \text{Softmax}_t(e_t)$
  $z_{band} \leftarrow \sum_t \alpha_t H_t$
  return z_{band}
end function

function FuseBands(*subject_bands* $= \{\textit{band_graphs}_1, \ldots, \textit{band_graphs}_B\}$)
  Initialize an empty list *band_embeddings*.
  for $b \leftarrow 1$ to B **do**
    $z_b \leftarrow$ ProcessBandSequence($\textit{band_graphs}_b$)
    $z_b \leftarrow \text{Linear}_{\text{proj}}(z_b)$
    Append z_b to *band_embeddings*.
  end for
  $g_b \leftarrow \text{Linear}_{\text{band-attn}}(z_b)$ for each b
  $\beta_b \leftarrow \text{Softmax}_b(g_b)$
  $z_{fused} \leftarrow \sum_{b=1}^{B} \beta_b \, z_b$
  $z_{fused} \leftarrow \text{LayerNorm}(z_{fused})$
  return z_{fused}, β_b
end function

for each subject s in the dataset **do**
  $z_{fused}, \beta_b \leftarrow$ FuseBands(s)
  $h \leftarrow \text{ReLU}(\text{Linear}_{fc1}(z_{fused}))$
  if training **then**
    $h \leftarrow \text{Dropout}(h)$
  end if
  $\hat{y} \leftarrow \text{Linear}_{fc2}(h)$
  $\mathcal{L}_{cls} \leftarrow \text{CrossEntropy}(\hat{y}, y)$
  $\mathcal{L}_{reg} \leftarrow \lambda \sum_{b=1}^{B} \left|\beta_b - \frac{1}{B}\right|$
  $\mathcal{L} \leftarrow \mathcal{L}_{cls} + \mathcal{L}_{reg}$
  Backpropagate $\mathcal{L}$; clip gradients to max norm 0.5; update model parameters via AdamW.
end for

OL=50, with an F_1-score of 0.8891 ± 0.0224 and an accuracy of 0.8896 ± 0.0225. This suggests that shorter windows with 50% overlap provided the most effective temporal representation for distinguishing HC from FTD. For the HC vs AD task, the highest

F_1-score was achieved using WL=6 and OL=50, reaching 0.8646 ± 0.0172, although the highest accuracy was observed for WL=6 and OL=0. This indicates that longer windows were generally more effective for HC vs AD classification, possibly because they captured more stable disease-related temporal patterns. In the FTD vs AD task, classification performance was lower than in the HC-based comparisons, reflecting the greater difficulty of distinguishing between the two dementia groups. The best performance for this task was obtained using WL=6 and OL=0, with an F_1-score of 0.8057 ± 0.0193 and an accuracy of 0.8144 ± 0.0175.

In addition, Table 3 shows that across all window and overlap configurations, Alpha- and Beta-band contributions to FTD vs AD classification remained approximately balanced, with each band accounting for roughly 47–53% of the combined weight. The near-equal split and overlapping standard deviations ($\pm$0.05–0.06) across all configurations suggest that neither frequency band provided a substantially stronger discriminative signal than the other. This indicates that Alpha and Beta oscillations contributed comparably to distinguishing between the two classes, regardless of the windowing parameters.

Table 4 shows the model performance gaps across different lobes. The temporal lobe provides the most critical input, as its independent predictive power nearly matches the whole-brain model. The parietal lobe also offers substantial predictive value, although it remains less effective than the temporal lobe. The frontal and central lobes provide useful but notably weaker independent signals, resulting in larger performance deficits. While the occipital lobe is the least reliable when isolated, it benefits the most from the compensatory mechanisms of the integrated model, which may filter noisy information and help maintain system-wide precision.

3.2 Explainability Results

The XAI analysis, including SmoothGrad, DeepLIFT, and Integrated Gradients, was applied to the whole-brain graphs and revealed consistent patterns of node importance across different attribution methods. However, these patterns differed from those observed in the lobe-wise study.

As shown in Table 5, the temporal lobe consistently achieved the strongest single-lobe performance, closely approaching or even exceeding whole-brain performance in some HC vs. AD configurations, while the occipital lobe performed poorly across all settings ($F_1 \approx 0.50$–0.57). Despite the temporal lobe's strength, the whole-brain model consistently matched or outperformed it with comparable or lower variance, indicating that the remaining lobes contribute complementary, stabilizing information.

As shown in Fig. 3, central electrodes (e.g., Cz) consistently received the highest attribution scores across all three XAI methods and both classification tasks. Frontal electrodes scored lower, despite their established role in FTD [42] and the temporal and parietal abnormalities reported in AD [43]. This pattern likely reflects the topological position of central nodes. Being connected to multiple neighboring regions, they play a disproportionate role in message-passing aggregation and come to dominate the global graph embedding, regardless of their actual discriminative value. Indeed, the lobe-wise evaluation shows this value to be limited. Clinically established regions such as the

Table 3 Performance comparison of temporal window configurations. Abbreviations: WL, window length; OL, overlap; Acc., accuracy; Rec., recall; Prec., precision; α imp., Alpha-band importance; β imp., Beta-band importance.

Task	Configuration	F_1	Acc.	Rec.	Prec.	α imp.	β imp.
HC vs FTD	WL=4, OL=0	0.8830 ± 0.0204	0.8836 ± 0.0206	0.8713 ± 0.0210	0.8950 ± 0.0237	0.5054 ± 0.0392	0.4946 ± 0.0392
	WL=4, OL=50	$\mathbf{0.8891 \pm 0.0224}$	$\mathbf{0.8896 \pm 0.0225}$	$\mathbf{0.8766 \pm 0.0316}$	$\mathbf{0.9021 \pm 0.0234}$	$\mathbf{0.4897 \pm 0.0400}$	$\mathbf{0.5103 \pm 0.0400}$
	WL=6, OL=0	0.8811 ± 0.0239	0.8822 ± 0.0235	0.8665 ± 0.0333	0.8965 ± 0.0169	0.5138 ± 0.0264	0.4862 ± 0.0264
	WL=6, OL=50	0.8870 ± 0.0202	0.8887 ± 0.0190	0.8685 ± 0.0279	0.9065 ± 0.0139	0.4893 ± 0.0459	0.5107 ± 0.0459
HC vs AD	WL=4, OL=0	0.8458 ± 0.0315	0.8492 ± 0.0273	0.8442 ± 0.0454	0.8692 ± 0.0202	0.4935 ± 0.0157	0.5065 ± 0.0157
	WL=4, OL=50	0.8539 ± 0.0180	0.8569 ± 0.0169	0.8339 ± 0.0245	0.8751 ± 0.0157	0.5091 ± 0.0388	0.4909 ± 0.0388
	WL=6, OL=0	0.8539 ± 0.0189	0.8569 ± 0.0178	0.8339 ± 0.0245	0.8751 ± 0.0166	0.5091 ± 0.0409	0.4909 ± 0.0409
	WL=6, OL=50	$\mathbf{0.8646 \pm 0.0172}$	$\mathbf{0.8662 \pm 0.0178}$	$\mathbf{0.8517 \pm 0.0213}$	$\mathbf{0.8781 \pm 0.0182}$	$\mathbf{0.4888 \pm 0.0155}$	$\mathbf{0.5112 \pm 0.0151}$
FTD vs AD	WL=4, OL=0	0.7686 ± 0.0361	0.7802 ± 0.0353	0.7364 ± 0.0718	0.8037 ± 0.0327	0.4863 ± 0.0549	0.5137 ± 0.0549
	WL=4, OL=50	0.7796 ± 0.0163	0.7870 ± 0.0135	0.7407 ± 0.0331	0.8227 ± 0.0187	0.5141 ± 0.0513	0.4859 ± 0.0513
	WL=6, OL=0	0.8057 ± 0.0193	0.8144 ± 0.0175	0.7677 ± 0.0386	0.8477 ± 0.0197	0.4673 ± 0.0560	0.5327 ± 0.0560
	WL=6, OL=50	0.7956 ± 0.0289	0.8024 ± 0.0275	0.7636 ± 0.0586	0.8302 ± 0.0289	0.4764 ± 0.0629	0.5236 ± 0.0629

Table 4 Performance gap analysis for individual lobes relative to the whole-brain model. Gap is defined as individual-lobe F_1-score minus whole-brain F_1-score; negative values indicate underperformance relative to the whole-brain model.

Lobe	HC vs FTD gap	HC vs AD gap
Temporal	-0.02 to $+0.01$	-0.01 to $+0.04$
Parietal	-0.11 to -0.07	-0.14 to -0.07
Frontal	-0.12 to -0.11	-0.20 to -0.14
Central	-0.18 to -0.11	-0.10 to -0.07
Occipital	-0.39 to -0.32	-0.34 to -0.28

temporal lobe, by contrast, receive comparatively little attribution despite carrying stronger diagnostic signal.

3.3 Ablation Study

To assess the contribution of frequency bands to classification performance, we conducted a frequency-band ablation study, evaluating different frequency-band combinations on the whole-brain graph.

Table 6 reports results for single- and multi-band combinations using the whole-brain graph with Window Length of 4 seconds and 50 percent of Overlap.

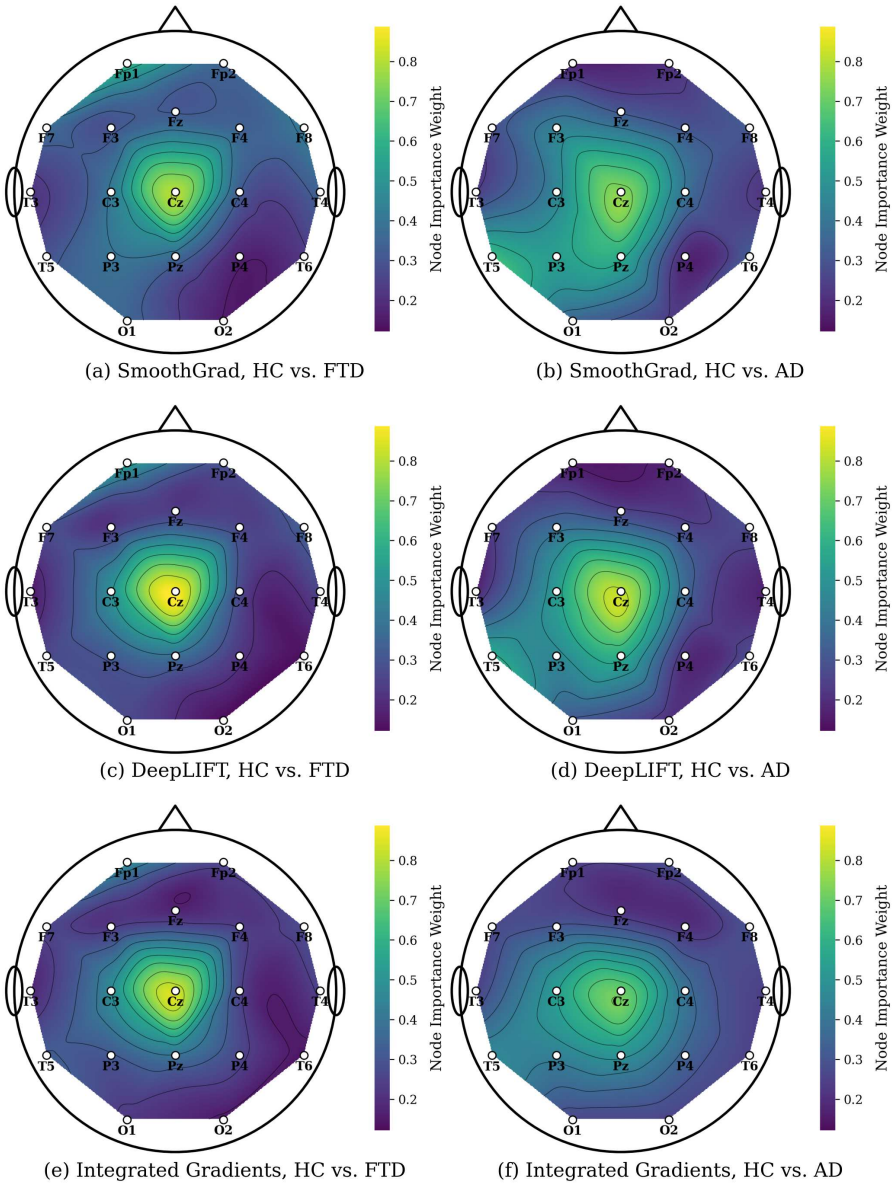
As shown in Table 6, Beta alone achieved the best performance for HC vs. FTD ($F_1 = 0.8987 \pm 0.0180$), while Alpha+Beta+Delta performed best for HC vs. AD ($F_1 = 0.8557 \pm 0.0341$). Delta alone was the weakest configuration in both tasks, and Delta-containing combinations showed higher variance, helping HC vs. AD only marginally while harming HC vs. FTD. We selected Alpha+Beta as the final configuration because it remained close to the best result in both tasks (HC vs. FTD: $F_1 = 0.8891 \pm 0.0236$; HC vs. AD: $F_1 = 0.8539 \pm 0.0189$), offering a consistent, generalizable middle ground rather than requiring task-specific models.

4 Discussion

Our approach introduces several methodological contributions to dementia classification using EEG signals, achieving mean accuracies of 0.8896 ± 0.0225 (F_1-score: 0.8891 ± 0.0224) for HC vs FTD and 0.8662 ± 0.0178 (F_1-score: 0.8646 ± 0.0172) for HC vs AD across 10 repetitions of stratified 5-fold cross-validation. For the more challenging FTD vs AD discrimination task, our approach achieved a mean accuracy of 0.8144 ± 0.175 (F_1-score: 0.8057 ± 0.193). From an architectural perspective, we integrate temporal GraphSAGE with bidirectional LSTM-Attention to model the dynamic evolution of brain connectivity over time, representing a departure from static approaches in prior work.

Table 5 Lobe-wise classification performance for HC/FTD and HC/AD tasks across temporal window configurations. Values are reported as mean ± standard deviation. Abbreviations: WL, window length; OL, overlap.

Task	Configuration	Whole-brain	Temporal	Central	Parietal	Frontal	Occipital
HC vs FTD	WL=4, OL=0	0.8830 ± 0.0204	0.8749 ± 0.0367	0.7265 ± 0.0396	0.7914 ± 0.0378	0.7687 ± 0.0317	0.5187 ± 0.0732
	WL=4, OL=50	0.8891 ± 0.0224	0.8807 ± 0.0176	0.7104 ± 0.0541	0.7815 ± 0.0336	0.7646 ± 0.0499	0.4968 ± 0.0916
	WL=6, OL=0	0.8811 ± 0.0239	0.8874 ± 0.0291	0.7662 ± 0.0446	0.8096 ± 0.0272	0.7615 ± 0.0376	0.5635 ± 0.0597
	WL=6, OL=50	0.8870 ± 0.0202	0.8660 ± 0.0320	0.7577 ± 0.0344	0.8094 ± 0.0301	0.7648 ± 0.0450	0.5703 ± 0.0707
HC vs AD	WL=4, OL=0	0.8458 ± 0.0315	0.8831 ± 0.0206	0.7439 ± 0.0694	0.7150 ± 0.0333	0.7016 ± 0.0417	0.5085 ± 0.0473
	WL=4, OL=50	0.8539 ± 0.0180	0.8796 ± 0.0213	0.7576 ± 0.0348	0.7815 ± 0.0336	0.6580 ± 0.0450	0.5735 ± 0.0702
	WL=6, OL=0	0.8635 ± 0.0265	0.8626 ± 0.0262	0.7960 ± 0.0368	0.7364 ± 0.0505	0.6979 ± 0.0237	0.5224 ± 0.0314
	WL=6, OL=50	0.8646 ± 0.0172	0.8537 ± 0.0381	0.7801 ± 0.0351	0.7205 ± 0.0357	0.7265 ± 0.0417	0.5653 ± 0.0589



(a) SmoothGrad, HC vs. FTD (b) SmoothGrad, HC vs. AD

(c) DeepLIFT, HC vs. FTD (d) DeepLIFT, HC vs. AD

(e) Integrated Gradients, HC vs. FTD (f) Integrated Gradients, HC vs. AD

Fig. 3 Brain topomap heatmaps showing the node importance weights assigned by each XAI method to the EEG electrode nodes. Left column: HC vs. FTD classification; right column: HC vs. AD classification. (a, b) SmoothGrad, (c, d) DeepLIFT, (e, f) Integrated Gradients.

As summarized in Table 7, none of these studies employed a temporal GNN framework, and most lacked explainability-based methods. While previous studies, such as Aljanabi et al. [44] and Bunterngchit et al. [45], achieved promising results for

Table 6 Performance comparison of different frequency-band combinations using the whole-brain graph with WL=4 and OL=50. Values are reported as mean ± standard deviation. Abbreviations: WL, window length; OL, overlap; Avg., average; F_1, F_1-score; Acc., accuracy; Prec., precision.

Task	Band combination	Avg. F_1	Avg. Acc.	Avg. Prec.
HC vs FTD	Alpha	0.8737 ± 0.0229	0.8756 ± 0.0225	0.8954 ± 0.0202
	Beta	$\mathbf{0.8987 \pm 0.0180}$	$\mathbf{0.8987 \pm 0.0181}$	$\mathbf{0.9114 \pm 0.0178}$
	Alpha+Beta	0.8891 ± 0.0224	0.8896 ± 0.0225	0.9021 ± 0.0234
	Alpha+Beta+Theta	0.8822 ± 0.0315	0.8836 ± 0.0305	0.8985 ± 0.0270
	Alpha+Beta+Delta+Theta	0.8673 ± 0.0392	0.8713 ± 0.0314	0.8825 ± 0.0463
	Alpha+Beta+Delta	0.8647 ± 0.0640	0.8716 ± 0.0516	0.8758 ± 0.0676
	Beta+Delta	0.8558 ± 0.0608	0.8689 ± 0.0470	0.8604 ± 0.0707
	Theta	0.7752 ± 0.0375	0.7793 ± 0.0365	0.8079 ± 0.0325
	Alpha+Delta	0.7687 ± 0.0820	0.7927 ± 0.0618	0.7861 ± 0.0959
	Delta+Theta	0.6906 ± 0.0486	0.7285 ± 0.0339	0.6964 ± 0.0507
	Delta	0.4392 ± 0.0415	0.5655 ± 0.0307	0.3881 ± 0.0739
HC vs AD	Alpha	0.8539 ± 0.0154	0.8554 ± 0.0149	0.8649 ± 0.0141
	Alpha+Beta	0.8539 ± 0.0180	0.8569 ± 0.0169	0.8751 ± 0.0157
	Alpha+Beta+Delta	$\mathbf{0.8557 \pm 0.0341}$	$\mathbf{0.8600 \pm 0.0303}$	$\mathbf{0.8804 \pm 0.0203}$
	Alpha+Beta+Delta+Theta	0.8391 ± 0.0524	0.8446 ± 0.0456	0.8570 ± 0.0515
	Alpha+Beta+Theta	0.8347 ± 0.0292	0.8385 ± 0.0283	0.8592 ± 0.0290
	Alpha+Delta	0.8375 ± 0.0223	0.8400 ± 0.0208	0.8538 ± 0.0207
	Beta	0.8362 ± 0.0267	0.8400 ± 0.0273	0.8551 ± 0.0295
	Theta	0.8222 ± 0.0299	0.8262 ± 0.0291	0.8485 ± 0.0347
	Beta+Delta	0.8173 ± 0.0278	0.8215 ± 0.0263	0.8449 ± 0.0238
	Delta+Theta	0.7890 ± 0.0581	0.8000 ± 0.0487	0.8110 ± 0.0642
	Delta	0.6123 ± 0.1022	0.6723 ± 0.0733	0.6542 ± 0.1230

dementia classification, their approaches focused on extracting spatial patterns without explicitly modeling the dynamics of brain connectivity. Similarly, Jiang et al. [46] and Zheng et al. [47] focused on static or aggregated features rather than temporal graph structures. Although Singh et al. [48] incorporated graph theory via minimum spanning trees (MSTs), their analysis was limited to static graphs rather than time-varying connectivity. Our GraphSAGE-based architecture aggregates information across time-evolving connectivity patterns, capturing dynamic neurophysiological changes characteristic of dementia progression. Additionally, the results of our multi-band integration approach not only address a gap in frequency-band utilization but also demonstrate the importance of leveraging cross-band information when studying EEG signals. Our analysis reveals relatively balanced contributions (47–52% Alpha, 48–52% Beta) across brain regions, suggesting that both bands carry complementary discriminative information that may be underutilized in single-band approaches.

Regarding the evaluation methodology, our study emphasizes reproducibility and stability assessment through several design choices. While higher single-run accuracies

Table 7 Comparison between previous studies. Abbreviations: RF, Random Forest; DT, Decision Tree; MST, minimum spanning tree; Att., attention; Dem., dementia; LOSO, leave-one-subject-out; LOOCV, leave-one-out cross-validation.

Paper	Dataset size	Network type	Explainability	Accuracy	F_1-score	Cross-val.	Subject split	Balance control
Bunterngchit et al. [45]	88 (36/23/29)	CNN-LSTM-Att.	No	97.4%	97.43%	No (80:20)	Not stated	SMOTE
Goerttler et al. [49]	40 + 65	SVM	No	∼80–82%	N/A	10-fold	Yes	–
Jiang et al. [46]	88 (36/23/29)	3D-CNN	Post-hoc	94.32%	N/A	LOOCV	Yes	–
Aljanabi et al. [44]	88 (36/23/29)	ResNet	No	99.49%	97.92% (AD)	5-fold	Yes	–
Vargas Cruz et al. [50]	88 (36/23/29)	RF	No	97% (AD-HC), 96% (FTD-HC)	N/A	Subject-wise	Yes	–
Zheng et al. [47]	88 (36/23/29)	RF	No	95.86% (AD-HC)	N/A	LOSO	Yes	–
Miltiadous et al. [39]	88 (36/23/29)	Transformer-CNN	Kernel weights	83.28% (AD-HC)	84.12%	LOSO	Yes	–
Rostamikia et al. [51]	88 (36/23/29)	SVM	No	87.8% (AD-FTD), 93.5% (Dem.-HC)	N/A	10-fold	No	–
Singh et al. [48]	46 (23/23)	RF/DT + MST	No	90% (Alpha)	∼0.833	5-fold	Not specified	–
Ours	88 (36/23/29)	Temporal GraphSAGE-BiLSTM-Att. (multi-band)	Lobe-wise + Captum	88.96% (HC-FTD) / 86.62% (HC-AD) / 81.44% (FTD-AD)	88.91% (HC-FTD) / 86.46% (HC-AD) / 80.57% (FTD-AD)	5-fold × 10 rep.	Yes	Weighted loss

have been reported in previous studies (e.g., Aljanabi et al., 99.49% [44]; Bunterngchit et al., 97.4% [45]; Vargas Cruz et al., 97% [50]; Zheng et al., 95.86% [47]; Jiang et al., 94.32% [46]), direct comparison requires consideration of differences in evaluation protocols. First, regarding stability assessment, prior studies conducted single cross-validation runs without quantifying performance variability across different data splits. In contrast, our 10-repetition protocol generated 50 independent train–test splits, providing confidence intervals and variance estimates that were absent in previous work, and offering a more comprehensive characterization of model performance. Second, regarding data partitioning, although Rostamikia et al. reported accuracies of 87.8–93.5%, their pipeline lacked a subject-level split, potentially allowing temporal correlation between training and test segments. In contrast, our subject-based stratification ensured independence between folds and provided a better estimate of generalization to unseen patients. Third, regarding imbalance handling, Bunterngchit et al. employed SMOTE oversampling for the class distribution of 36:23:29 (AD:FTD:HC), thereby generating synthetic samples. In contrast, our weighted loss function addressed class imbalance while preserving the original signal characteristics. Fourth, regarding validation schemes, Bunterngchit et al. used an 80:20 train–test split rather than cross-validation, Jiang et al. employed leave-one-out cross-validation (LOOCV), and Singh et al. did not specify subject-level splitting. In contrast, our stratified 5-fold cross-validation with subject-level splits balanced computational feasibility with robust generalization estimation. These methodological differences suggest that our reported performance may represent a more conservative estimate that better reflects expected generalization to independent patient populations in clinical settings.

A notable gap in prior work concerns model interpretability, which is increasingly recognized as essential for the clinical adoption of machine learning systems. Of the nine comparison studies reviewed, five provided no explainability mechanism, including Bunterngchit et al. [45], Goerttler et al. [49], Vargas Cruz et al. [50], Zheng et al. [47], and Rostamikia et al. [51]. Among studies incorporating interpretability, Jiang et al. [46] referred to post-hoc methods, while Miltiadous's DICE-Net analyzed kernel weights [39]. Our work extends interpretability through systematic Captum-based attribution analysis combined with anatomically grounded lobe-wise evaluation, including the frontal, temporal, parietal, central, and occipital lobes.

The lobe-wise analysis showed that the temporal lobe achieved the highest accuracy and F_1-score for distinguishing patients from healthy controls, with performance in many settings comparable to the whole-brain model. This finding is consistent with its role in memory circuits and early atrophy in Alzheimer's disease [52]. The parietal lobe ranked second in performance, a result that may reflect its role as association cortex within large-scale network connectivity [53]. In contrast, the frontal and central lobes showed weaker performance, despite their established links to executive and behavioral functions and to motor-related processing, respectively [54]. The occipital lobe, which is less involved in Alzheimer's disease and mainly associated with visual processing [53], exhibited the poorest performance and may therefore primarily contribute noise. In line with these observations, the whole-brain analysis in the HC vs FTD classification achieved an F_1-score of 0.88–0.89, representing only a 0.01–0.02 improvement over the temporal lobe (F_1: 0.86–0.88), which showed the best single-lobe performance.

By contrast, in the HC vs AD classification, whole-brain performance (F_1: 0.84–0.86) was comparable to, and in some cases slightly lower than, the temporal lobe alone (F_1: 0.85–0.88). This indicates that discriminative information for Alzheimer's disease classification is mainly concentrated in temporal regions, with the remaining lobes serving a complementary role.

On the other hand, the analysis of the role of the number of nodes in each lobe indicates that there is no linear or direct relationship between node count and classification performance. In other words, the superior performance of the temporal lobe is not merely a consequence of its number of nodes, but rather reflects the richer disease-related information captured by this region compared with other lobes. Likewise, the weaker performance of the frontal lobe, despite containing the largest number of nodes, demonstrates that simply increasing the number of nodes does not necessarily lead to improved disease discrimination. Instead, the functional distribution and informational relevance of the nodes appear to play a more influential role than their quantity. This finding is further supported by the quantitative results in Table 5.

Notably, although the whole-brain model integrates information from all lobes, its performance improvement over the temporal lobe remains limited to only a few percentage points. This analysis further confirms that the key regions containing crucial information for Alzheimer's disease diagnosis are predominantly concentrated in the temporal lobe. This finding is consistent with clinical evidence regarding the early and dominant involvement of temporal structures in Alzheimer's disease [55]. Accordingly, our model's results indicate that the temporal lobe plays a more influential role in diagnosing Alzheimer's disease than in diagnosing FTD. Nevertheless, given the limited number of EEG channels and electrode distribution in this study, future investigations with broader channel coverage may provide a more precise understanding of the actual contribution of each lobe.

Finally, the dual-band GNN model shows that whole-brain integration stabilizes the band-level variability observed in weaker lobes, reveals disease-specific spectral patterns, and largely preserves the temporal lobe's band pattern in the whole-brain representation (Table 5). Specifically, FTD showed a relatively balanced profile with a slight Alpha-band preference, whereas AD showed Beta-band dominance. These findings indicate that the stabilizing effect of the whole-brain model is mainly driven by the temporal lobe, further confirming it as the primary contributor to disease discrimination.

The use of Explainable AI (XAI) algorithms through the Captum library enabled investigation of the model's decision-making patterns from a whole-brain perspective. However, the results obtained from the three explainability methods, including SmoothGrad, DeepLIFT, and Integrated Gradients, did not show a high level of consistency with either the lobe-wise analysis or the patterns reported in clinical studies. By contrast, the lobe-wise analysis demonstrated stronger agreement with established clinical findings, suggesting that region-based attribution better captures neurobiologically meaningful patterns than the gradient-based XAI methods applied here. This discrepancy likely stems from how GNNs learn and represent information, because GNNs aggregate and propagate features across nodes through message passing.

Overall, combining lobe-wise analysis with conventional XAI methods highlights an important limitation: while lobe-wise analysis offers an interpretable and clinically consistent understanding of influential brain regions, the XAI methods tested here do not yet provide comparable reliability when applied to GNNs. Therefore, lobe-wise analysis remains the more clinically trustworthy approach for region-level interpretation in this context, while further methodological development and validation of XAI techniques specifically tailored to graph-based architectures are needed before they can be confidently applied in clinical settings. Nevertheless, to ensure the generalizability of these findings, future studies on independent datasets with a higher number of channels and different electrode configurations will be necessary.

5 Limitations and Future Work

A limitation of this study is that GNN algorithms rely on message-passing techniques that update feature vectors solely based on graph topology and node embeddings. As a result, although the Granger graphs contain informative edge features, we were not able to incorporate them into the model. To address this limitation, future work will focus on developing a model that leverages both node and edge features, which may lead to more accurate results. Moreover, although we constructed temporal graphs across four window sizes and overlapping scenarios, we were constrained by computational resources. Thus, we did not investigate other window lengths, which could have resulted in higher accuracy.

Declarations

- **Funding:** The authors received no specific funding for this work; all work was carried out by the authors themselves.

- **Competing interests:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
- **Ethics approval and consent to participate:** This study used a previously collected, publicly available, and fully anonymized EEG dataset (OpenNeuro accession ds004504) [38, 39]. Ethical approval for the original data collection was obtained from the Scientific and Ethics Committee of AHEPA University Hospital, Aristotle University of Thessaloniki (protocol no. 142/12-04-2023), in accordance with the Declaration of Helsinki, and written informed consent was obtained from all participants or their legal guardians. The privacy rights of all human subjects were observed, and all data used in this study were fully anonymized prior to public release. As the data analyzed in this study are fully anonymized and publicly available, no additional institutional ethics approval was required for this secondary analysis.
- **Consent for publication:** Not applicable.
- **Data availability:** The EEG dataset analyzed in this study is publicly available on OpenNeuro (accession number ds004504, https://doi.org/10.18112/openneuro.ds004504.v1.0.8), originally described in [38, 39].

- **Author contributions: Mahshad Faramarzi:** Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing. **Mohammad Fathi:** Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing. **Mohammadreza Gholami:** Software, Visualization, Writing – original draft, Writing – review & editing. **Ahmad Haj Reza Jafarabadi:** Resources. **Mir Mohsen Pedram:** Conceptualization, Supervision, Writing – review & editing.
- **Declaration of generative AI and AI-assisted technologies in the manuscript preparation process:** During the preparation of this work, the author(s) used ChatGPT for language and grammar editing. The author(s) reviewed and edited the output as needed and take full responsibility for the content of the published article.

References

[1] Safiri, S., Ghaffari Jolfayi, A., Fazlollahi, A., Morsali, S., Sarkesh, A., Daei Sorkhabi, A., Golabi, B., Aletaha, R., Motlagh Asghari, K., Hamidi, S., *et al.*: Alzheimer's disease: a comprehensive review of epidemiology, risk factors, symptoms diagnosis, management, caregiving, advanced treatments and associated challenges. Frontiers in Medicine **11**, 1474043 (2024)

[2] Lanctot, K.L., Hahn-Pedersen, J.H., Eichinger, C., Freeman, C., Clark, A., Tarazona, L., Cummings, J.: Burden of illness in people with Alzheimer's disease: a systematic review of epidemiology, comorbidities and mortality. The journal of prevention of Alzheimer's disease **11**(1), 97–107 (2024)

[3] Antonioni, A., Raho, E.M., Lopriore, P., Pace, A.P., Latino, R.R., Assogna, M., Mancuso, M., Gragnaniello, D., Granieri, E., Pugliatti, M., *et al.*: Frontotemporal dementia, where do we stand? a narrative review. International journal of molecular sciences **24**(14), 11732 (2023)

[4] Chen, X., Chen, Y., Ni, B., Huang, C.: Research trends and hotspots for frontotemporal dementia from 2000 to 2022: a bibliometric analysis. Frontiers in Neurology **15**, 1399600 (2024)

[5] Lin, C.-Y., Zhai, Y.-J., An, H.-H., Wu, F., Qiu, H.-N., Li, J.-B., Lin, J.-N.: Global trends in prevalence, disability adjusted life years, and risk factors for early onset dementia from 1990 to 2021. Scientific Reports **15**(1), 13488 (2025)

[6] Mostert, C.M., Udeh-Momoh, C., Winkler, A.S., McLaughlin, C., Eyre, H., Salama, M., Ranchod, K., Trepel, D., Vradenburg, G., Hynes, W., et al.: Broadening dementia risk models: building on the 2024 Lancet commission report for a more inclusive global framework. EBioMedicine **120** (2025)

[7] Nguyen, H.-D., Clément, M., Planche, V., Mansencal, B., Coupé, P.: Deep grading for MRI-based differential diagnosis of Alzheimer's disease and frontotemporal

dementia. Artificial intelligence in medicine **144**, 102636 (2023)

[8] Ma, D., Lu, D., Popuri, K., Wang, L., Beg, M.F., Initiative, A.D.N.: Differential diagnosis of frontotemporal dementia, Alzheimer's disease, and normal aging using a multi-scale multi-type feature generative adversarial deep neural network on structural magnetic resonance images. Frontiers in neuroscience **14**, 853 (2020)

[9] Cao, J., Yang, L., Sarrigiannis, P.G., Blackburn, D., Zhao, Y.: Dementia classification using a graph neural network on imaging of effective brain connectivity. Computers in Biology and Medicine **168**, 107701 (2024)

[10] Zhang, D., Zhu, C.: A dual path graph neural network framework for dementia diagnosis. Scientific Reports **15**(1), 23319 (2025)

[11] Wu, S., Zhan, P., Wang, G., Yu, X., Liu, H., Wang, W.: Changes of brain functional network in Alzheimer's disease and frontotemporal dementia: A graph-theoretic analysis. BMC neuroscience **25**(1), 30 (2024)

[12] Ehteshamzad, S.: Assessing the potential of EEG in early detection of alzheimer's disease: A systematic comprehensive review (2000–2023). Journal of Alzheimer's Disease Reports **8**(1), 1153–1169 (2024)

[13] Paitel, E.R., Otteman, C.B., Polking, M.C., Licht, H.J., Nielson, K.A.: Functional and effective EEG connectivity patterns in Alzheimer's disease and mild cognitive impairment: A systematic review. Frontiers in aging neuroscience **17**, 1496235 (2025)

[14] Benussi, A., Cantoni, V., Palacino, F., Altomare, D., Moretti, D.V., Manganotti, P., Borroni, B.: EEG network reorganization across Alzheimer's disease, frontotemporal dementia, and dementia with Lewy bodies. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring **18**(1), 70275 (2026)

[15] Li, J., Chen, X., Shen, F., Chen, J., Liu, Y., Zhang, D., Yuan, Z., Zhao, F., Li, M., Yang, Y.: Deep learning-powered electrical brain signals analysis: Advancing neurological diagnostics. IEEE Reviews in Biomedical Engineering (2025)

[16] Khan, A., Shim, V., Fernandez, J., Kasabov, N.K., Wang, A.: Review of deep learning models with spiking neural networks for modeling and analysis of multimodal neuroimaging data. Frontiers in Neuroscience **19**, 1623497 (2025)

[17] Balsabti, S.M., Al-Gburi, R.M., Mustafa, A., Ahmed, S.K., Issa, A.M., Al-Naimi, T.M., AlSaad, R., Elhenidy, A.M., *et al.*: Advances in deep learning for multimodal brain imaging: A comprehensive survey. Neuroscience Informatics **6**(1), 100252 (2026)

[18] Dong, C., Zhang, Z., Sun, D.: Multi-channel EEG-based neurological disorder classification using cross-dependency spatiotemporal interactive network.

Computer Methods and Programs in Biomedicine, 108982 (2025)

[19] Liu, Z., Zhao, J.: Leveraging deep learning for robust EEG analysis in mental health monitoring. Frontiers in neuroinformatics **18**, 1494970 (2025)

[20] Lyu, R.: Deep learning approaches for EEG-based healthcare applications: a comprehensive review. Frontiers in Human Neuroscience **19**, 1689073 (2026)

[21] Klepl, D., He, F., Wu, M., Blackburn, D.J., Sarrigiannis, P.: EEG-based graph neural network classification of Alzheimer's disease: An empirical evaluation of functional connectivity methods. IEEE Transactions on Neural Systems and Rehabilitation Engineering **30**, 2651–2660 (2022)

[22] Cao, J., Yang, L., Sarrigiannis, P.G., Blackburn, D., Zhao, Y.: Dementia classification using a graph neural network on imaging of effective brain connectivity. Computers in Biology and Medicine **168**, 107701 (2024)

[23] Shan, X., Cao, J., Huo, S., Chen, L., Sarrigiannis, P.G., Zhao, Y.: Spatial–temporal graph convolutional network for Alzheimer classification based on brain functional connectivity imaging of electroencephalogram. Human Brain Mapping **43**(17), 5194–5209 (2022)

[24] Coelli, S., Corda, M., Bianchi, A.M.: The time-varying brain: a comprehensive review of dynamic functional connectivity analysis in EEG and MEG. Journal of Neural Engineering **22**(5), 051005 (2025)

[25] Klepl, D., He, F., Wu, M., Blackburn, D.J., Sarrigiannis, P.G.: Cross-frequency multilayer network analysis with bispectrum-based functional connectivity: A study of Alzheimer's disease. Neuroscience **521**, 77–88 (2023)

[26] Yuan, Y., Zhao, Y.: The role of quantitative EEG biomarkers in Alzheimer's disease and mild cognitive impairment: Applications and insights. Frontiers in aging neuroscience **17**, 1522552 (2025)

[27] Tewari, N., Le, N.L.D., Liu, M., Ren, J., Xu, Z., Sarwar, T., Baths, V., Xia, F.: Explainable graph neural networks: Understanding brain connectivity and biomarkers in dementia. arXiv preprint arXiv:2509.18568 (2025)

[28] Mohammadi, H., Karwowski, W.: Graph neural networks in brain connectivity studies: Methods, challenges, and future directions. Brain Sciences **15**(1), 17 (2024)

[29] Ali, S., Piana, M., Pardini, M., Garbarino, S.: Graph neural networks in Alzheimer's disease diagnosis: a review of unimodal and multimodal advances. Frontiers in Neuroscience **19**, 1623141 (2025)

[30] Miyakoshi, M.: Artifact subspace reconstruction: a candidate for a dream solution

for EEG studies, sleep or awake. Oxford University Press US (2023)

[31] Cataldo, A., Criscuolo, S., De Benedetto, E., Masciullo, A., Pesola, M., Schiavoni, R., Invitto, S.: A method for optimizing the artifact subspace reconstruction performance in low-density EEG. IEEE Sensors Journal **22**(21), 21257–21265 (2022)

[32] Jiang, X., Bian, G.-B., Tian, Z.: Removal of artifacts from EEG signals: a review. Sensors **19**(5), 987 (2019)

[33] Pion-Tonachini, L., Kreutz-Delgado, K., Makeig, S.: The ICLabel dataset of electroencephalographic (EEG) independent component (IC) features. Data in brief **25**, 104101 (2019)

[34] Hamilton, W., Ying, Z., Leskovec, J.: Inductive representation learning on large graphs. Advances in neural information processing systems **30** (2017)

[35] Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International Conference on Machine Learning, pp. 3319–3328 (2017). PMLR

[36] Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: International Conference on Machine Learning, pp. 3145–3153 (2017). PMlR

[37] Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825 (2017)

[38] Miltiadous, A., Tzimourta, K.D., Afrantou, T., Ioannidis, P., Grigoriadis, N., Tsalikakis, D.G., Angelidis, P., Tsipouras, M.G., Glavas, E., Giannakeas, N., Tzallas, A.T.: A dataset of scalp EEG recordings of Alzheimer's disease, frontotemporal dementia and healthy subjects from routine EEG. Data **8**(6), 95 (2023) https://doi.org/10.3390/data8060095

[39] Miltiadous, A., Gionanidis, E., Tzimourta, K.D., Giannakeas, N., Tzallas, A.T.: DICE-net: a novel convolution-transformer architecture for Alzheimer detection in EEG signals. IEEe Access **11**, 71840–71858 (2023)

[40] Bullmore, E., Sporns, O.: Complex brain networks: graph theoretical analysis of structural and functional systems. Nature reviews neuroscience **10**(3), 186–198 (2009)

[41] Pelle, S., Piermaria, G., Magosso, E., Ursino, M.: On the virtues and limitations of Granger-causal brain connectivity estimate: Critical analysis using neural mass models. Network Neuroscience **10**(1), 1–24 (2026)

[42] Sanchez, L.A., Pandiaraju, S., Williams, A.O., Daraie, A.H., Onyike, C.U., Sarma, S.V.: Sink-index: a network-based EEG marker for frontotemporal dementia and

Alzheimer's disease. Brain Communications **7**(4), 259 (2025)

[43] Ghassemkhani, K., Saroka, K.S., Dotta, B.T.: Evaluating EEG complexity and spectral signatures in Alzheimer's disease and frontotemporal dementia: evidence for rostrocaudal asymmetry. npj Aging **11**(1), 50 (2025)

[44] Aljanabi, E., Türker, İ.: Connectogram-COH: A coherence-based time-graph representation for EEG-based Alzheimer's disease detection. Diagnostics **15**(11), 1441 (2025)

[45] Bunterngchit, C., Chearanai, T., Bunterngchit, Y.: Advanced EEG-based classification of alzheimer's disease using cnn-lstm-attention architecture. In: 2024 22nd International Conference on Research and Education in Mechatronics (REM), pp. 107–112 (2024). IEEE

[46] Jiang, R., Zheng, X., Sun, J., Chen, L., Xu, G., Zhang, R.: Classification for Alzheimer's disease and frontotemporal dementia via resting-state electroencephalography-based coherence and convolutional neural network. Cognitive Neurodynamics **19**(1), 46 (2025)

[47] Zheng, X., Wang, B., Liu, H., Wu, W., Sun, J., Fang, W., Jiang, R., Hu, Y., Jin, C., Wei, X., *et al.*: Diagnosis of Alzheimer's disease via resting-state EEG: integration of spectrum, complexity, and synchronization signal features. Frontiers in Aging Neuroscience **15**, 1288295 (2023)

[48] Singh, P., Kumar, L., Gandhi, T.K.: Exploring network topology-based methods to differentiate healthy and Alzheimer's cohorts: An EEG-based study. In: 2024 32nd European Signal Processing Conference (EUSIPCO), pp. 1576–1580 (2024). IEEE

[49] Goerttler, S., He, F., Wu, M.: Balancing spectral, temporal and spatial information for eeg-based alzheimer's disease classification. In: 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 1–4 (2024). IEEE

[50] Cruz, M.A.V.: Diagnosis of Alzheimer's disease, Parkinson's disease, frontotemporal dementia, and paranoid schizophrenia via complex network analysis of EEG data. Precision Nanomedicine **8**(2), 1464–1472 (2025)

[51] Rostamikia, M., Sarbaz, Y., Makouei, S.: EEG-based classification of Alzheimer's disease and frontotemporal dementia: a comprehensive analysis of discriminative features. Cognitive Neurodynamics **18**(6), 3447–3462 (2024)

[52] De Flores, R., Das, S.R., Xie, L., Wisse, L.E., Lyu, X., Shah, P., Yushkevich, P.A., Wolk, D.A.: Medial temporal lobe networks in Alzheimer's disease: structural and molecular vulnerabilities. Journal of Neuroscience **42**(10), 2131–2141 (2022)

[53] He, Y., Chen, Z., Evans, A.: Structural insights into aberrant topological patterns of large-scale cortical networks in Alzheimer's disease. Journal of Neuroscience **28**(18), 4756–4766 (2008)

[54] Ilchovska, Z.G., Lockwood, J., Proctor, E., Hosseini, A., Ralph, M.A.L., Jung, J.: The neural basis of frontotemporal dementia (FTD): insights from ALE meta-analyses of four FTD subtypes encompassing 8,057 patients. medRxiv (2025)

[55] Dickerson, B.C., Sperling, R.A.: Functional abnormalities of the medial temporal lobe memory system in mild cognitive impairment and Alzheimer's disease: insights from functional MRI studies. Neuropsychologia **46**(6), 1624–1635 (2008)