ST-BiT: Spatio-Temporal Bipartite Transformer Network for Interaction-Preserving EEG-Based Dementia Subtyping

Siddhant Ujjain, Vaibhav Kagathara, Pooja Singh, Tapan Gandhi and Sandeep Kumar
Indian Institute of Technology Delhi, New Delhi, India
{eez228484, eet242852, eez228470, tgandhi, ksandeep}@iitd.ac.in

Abstract

EEG-based dementia classification often degrades under clinically realistic subject-wise evaluation due to non-stationarity and large inter-subject variability. A key modeling limitation is *relation compression*: many EEG-GNN pipelines encode functional connectivity as scalar edge weights and blur interaction structure during message passing, while models may also exploit subject-specific cues. We propose **ST-BiT**, a **S**patio-**T**emporal **B**ipartite **T**ransformer that represents electrodes as *node tokens* and functional interactions as explicit, learnable *edge tokens*. Edge tokens preserve pairwise coupling patterns and are updated only from their endpoint electrodes via *incidence-masked* cross-attention, while electrode tokens aggregate information only from incident edges. Window-wise sparse graphs are constructed from time-sample correlations of band-limited signals to sparsify and initialize edge tokens. ST-BiT combines this interaction-preserving backbone with temporal self-attention, lightweight band attention, and domain-adversarial alignment to reduce subject bias. On OpenNeuro ds006036 (eyes-open with photic stimulation), using leakage-free subject-wise stratified 5-fold cross-validation with nested model selection, ST-BiT achieves 93.0% accuracy for CN vs. (AD+FTD) and 76.1% for CN/AD/FTD, outperforming classical ML and GNN baselines under identical folds. To assess robustness across recording states and align with prior work on this cohort, we also evaluate on the ds004504 (eyes-closed) dataset.

1 Introduction

Dementia is a growing clinical burden, yet scalable screening and reliable differential diagnosis remain limited outside specialized centers [Aranda *et al.*, 2021]. EEG offers an appealing path toward accessible assessment, due to its low cost and sensitivity to fast neural dynamics [Paitel *et al.*, 2025]; however, dementia classifiers often degrade under leakage-free subject-wise evaluation because inter-subject shifts and non-stationarity encourage reliance on subject- or session-specific cues [Varoquaux, 2018]. A second, more structural limitation lies in how many EEG graph pipelines represent interactions [Li *et al.*, 2023]. Dementia phenotypes are not only reflected in regional spectral changes but also in altered *functional coupling* across scalp regions [Jeong, 2004; Musaeus *et al.*, 2019]. Yet most EEG-GNN approaches encode connectivity as scalar edge weights and then compress relation information into node embeddings through message passing [Cui *et al.*, 2023]. This *relation compression* discards interaction-specific structure that may matter most for subtyping problems such as differentiating Alzheimer's disease (AD) from frontotemporal dementia (FTD), where scalp-level signatures can overlap and subtle coupling patterns may be discriminative.

We propose **ST-BiT** (see Figure 1), a **S**patio-**T**emporal **Bi**partite **T**ransformer that represents electrodes and electrode-electrode relations as distinct token types. ST-BiT models electrodes as node tokens and retained functional connections as learnable edge tokens, updated using incidence-masked cross-attention: each edge token is updated only from its two endpoint electrodes, and each electrode aggregates information only from its incident edges. This preserves interaction structure rather than collapsing relations into scalar adjacency weights. Window-wise sparse graphs are constructed from per-band time-sample correlations of band-limited signals and are used to sparsify connections and initialize edge tokens. The interaction-preserving spatial backbone is combined with temporal self-attention across windows, lightweight frequency-band attention, and domain-adversarial alignment to reduce subject-specific bias.

We evaluate primarily on OpenNeuro ds006036 (eyes-open with photic stimulation) from an 88-subject cohort (29 CN, 36 AD, 23 FTD) under leakage-free subject-wise stratified 5-fold cross-validation with nested model selection, a setting less studied than the eyes-closed benchmark for this cohort. To assess robustness across recording states and connect with prior literature, we additionally evaluate ST-BiT on OpenNeuro ds004504 (eyes-closed) under the same leakage-free subject-wise protocol.

Our contributions are as follows:

- **Interaction-preserving EEG modeling.** We introduce ST-BiT, a bipartite transformer that represents func-

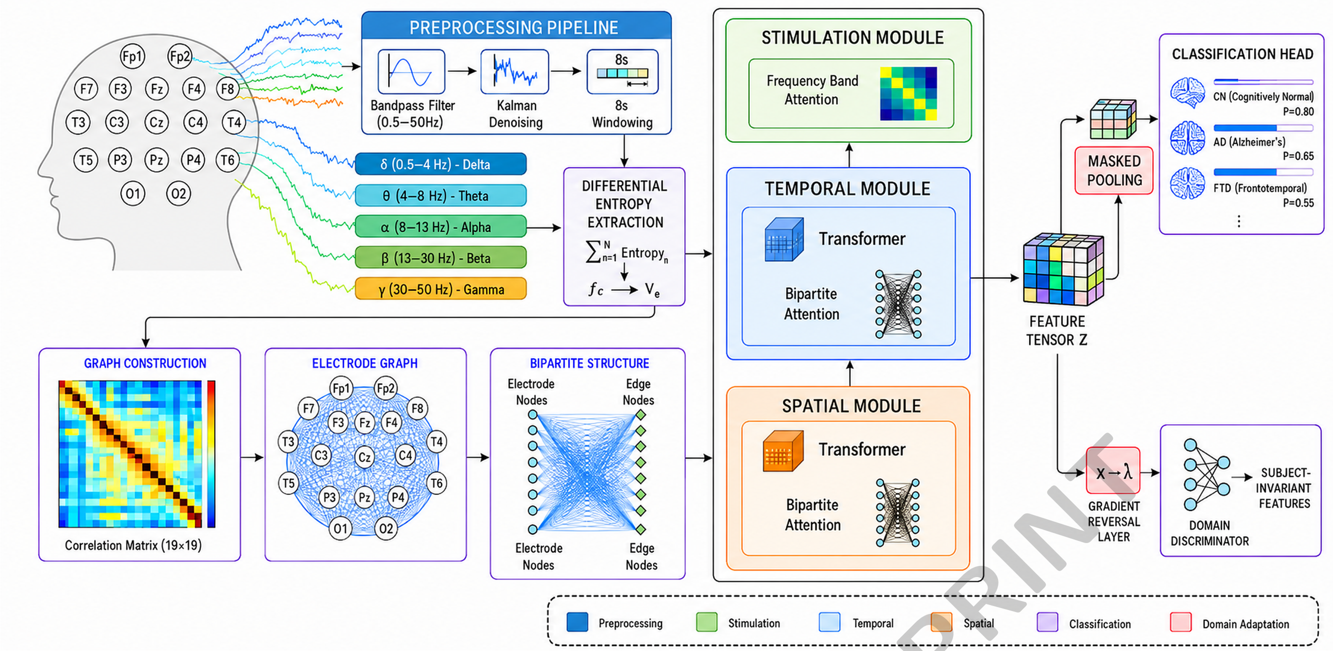


Figure 1: Overview of the proposed ST-BiT architecture. (1) **Preprocessing and features**: EEG is band-pass filtered (0.5–50 Hz), denoised with Kalman smoothing, and segmented into 8s window features via differential entropy over five bands (δ–γ). (2) **Graph Construction**: per-window functional connectivity is estimated using time-sample correlations of band-limited signals, then sparsified and converted to a bipartite incidence representation with electrode node tokens and explicit edge tokens. (3) **Backbone**: band attention, a temporal transformer for window-to-window dynamics, and incidence-masked bipartite cross-attention that exchanges information between node and edge tokens. (4) **Heads**: masked pooling and classification; an auxiliary domain discriminator supports adversarial subject-invariant learning.

tional connections as explicit edge tokens and updates node and edge tokens via incidence-masked node cross-attention, mitigating relation compression.

- **Robust spatio-temporal-spectral learning.** We couple the bipartite backbone with temporal self-attention, band attention, and domain-adversarial alignment to improve subject-wise robustness.
- **Comprehensive evaluation on eyes-open dementia subtyping.** We provide leakage-free nested subject-wise cross-validation on eyes-open EEG with ablations and sensitivity analyses, and we additionally evaluate on eyes-closed recordings to assess robustness across recording states and connect with prior work on this cohort.

2 Related Work

Early EEG-based dementia studies paired handcrafted spectral complexity descriptors with classical classifiers [Jeong, 2004]. Although effective in subject-dependent settings, performance often degrades under *subject-wise* evaluation due to inter-subject variability and recording-condition shift. Deep models (CNNs, time-frequency networks, transformers) have improved EEG-based dementia classification [Wang *et al.*, 2024; Stefanou *et al.*, 2025], but many pipelines still decompose space, time, and frequency into loosely coupled modules and provide limited support for explicitly modeling *inter-electrode interactions*. Graph neural networks (GNNs) incorporate functional connectivity by representing electrodes as nodes and connectivity estimates as edges [Klepl *et al.*, 2022]. In common EEG-GNN formulations, edges are treated as scalar weights and interaction information is propagated via message passing. This design can induce *relation compression*: edge-specific structure is absorbed into node embeddings, potentially blurring interaction patterns that matter for fine-grained phenotyping (e.g., AD vs. FTD) [Hasan *et al.*, 2024].

Heterogeneous and bipartite attention architectures mitigate relation compression by treating relations as explicit entities and alternating node→edge and edge→node updates, enabling learned relation representations rather than fixed adjacency scalars [Hu *et al.*, 2020; Kim *et al.*, 2022; Yun *et al.*, 2019]. Separately, domain-adversarial learning with gradient reversal reduces subject-specific bias by encouraging representations that are predictive of the label of interest while being uninformative of subject identity [Ganin *et al.*, 2016].

ST-BiT builds on these directions by modeling functional connections as explicit edge tokens updated using incidence-masked cross-attention between node and edge tokens, and by combining this interaction-preserving backbone with temporal self-attention, lightweight band attention, and domain-adversarial alignment to support robust subject-wise dementia screening and subtyping.

3 Dataset and Problem Formulation

We study EEG-based dementia *screening* and *subtyping* under *subject-wise* evaluation. All experiments are conducted on publicly available datasets to ensure reproducibility.

3.1 Datasets

We use two publicly available EEG datasets hosted on OpenNeuro, collected from a single clinical cohort of 88 subjects (29 cognitively normal (CN), 36 Alzheimer's disease (AD), and 23 frontotemporal dementia (FTD)) at the 2nd Department of Neurology, AHEPA General Hospital (Thessaloniki, Greece) [Ntetska *et al.*, 2025]. Participant demographics are summarized in Table 1. EEG was recorded using a 19-channel 10-20 montage at a sampling rate of 500 Hz. Each recording is denoted $\mathbf{X}^{(n)} \in \mathbb{R}^{T^{(n)} \times 19}$ with label $y^{(n)} \in \{0, 1, 2\}$ corresponding to CN, AD, and FTD.[1]

Group	N	Age (y)	MMSE	Rec. Dur. (min)
CN	29	67.9 ± 5.4	30.0 ± 0.0	6.43
AD	36	66.4 ± 7.9	17.8 ± 4.5	4.86
FTD	23	63.6 ± 8.2	22.2 ± 8.2	4.42

Table 1: Participant demographics for the 88-subject cohort (mean$\pm$SD). The same cohort was recorded in ds006036 (eyes-open photic stimulation) and ds004504 (eyes-closed resting-state).

Eyes-Open Dataset (ds006036). The ds006036 [Ntetska *et al.*, 2025] dataset, recorded under eyes-open conditions with incremental photic stimulation (5-30 Hz), is used as the primary evaluation dataset for training ST-BiT and for comparison against classical machine learning, graph-based, and deep-learning baselines. Representative grand-average denoised EEG waveforms highlighting cohort-level differences in amplitude and temporal variability are shown in Figure 2. Group-averaged band-power scalp topographies illustrating spatial distribution differences across cohorts are shown in Figure 3.

Why ds006036 (eyes-open) is challenging and under-benchmarked. Unlike the commonly studied eyes-closed resting-state setting, ds006036 contains eyes-open recordings with photic stimulation, which introduces stronger non-stationarity and amplifies inter-subject variability. To our knowledge, there is no established leakage-free *subject-wise* benchmark with standardized folds for this setting; we therefore treat ds006036 as the primary benchmark and report a unified subject-wise nested-CV protocol to support consistent comparison.

Eyes-Closed Dataset (ds004504). To evaluate robustness to recording conditions, we additionally assess ST-BiT on the ds004504 dataset [Miltiadous *et al.*, 2024], which consists of eyes-closed resting-state EEG recordings from the same cohort. This dataset is used exclusively to assess generalization and to enable comparison with prior works on the same cohort [Kim *et al.*, 2025; Stefanou *et al.*, 2025; Hasan *et al.*, 2024].

[1]For binary screening, AD and FTD are grouped into a single *dementia* class.

3.2 Tasks

We consider two classification tasks: (i) **binary screening** (CN vs. AD+FTD) and (ii) **three-way subtyping** (CN vs. AD vs. FTD).

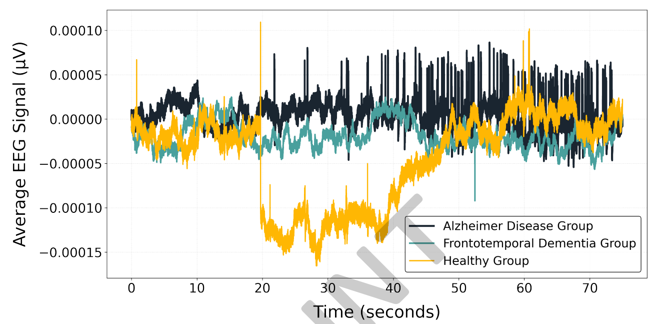


Figure 2: **Grand-average denoised EEG** waveforms (ds006036 eyes-open) over a 75 s interval for AD (charcoal), FTD (teal), and CN (gold). Cohorts show distinct offsets and amplitude fluctuations; AD exhibits higher temporal variability relative to CN.

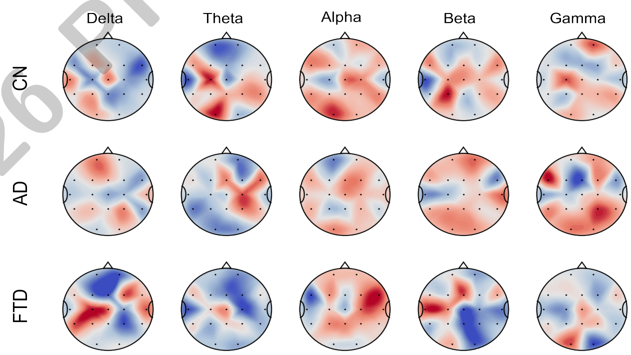


Figure 3: Group-averaged EEG scalp topographies (ds006036, eyes-open) of band-limited power across δ–γ for CN, AD, and FTD, illustrating qualitative differences in spatial power distribution across frequency bands.

4 Methodology

ST-BiT is a spatio-temporal bipartite transformer designed for clinically realistic *subject-wise* dementia screening and subtyping from EEG. The pipeline (see Figure 1) consists of: (i) preprocessing and window-level feature extraction, (ii) window-wise functional connectivity estimation and conversion to a bipartite incidence graph, and (iii) spatio-temporal modeling with explicit node-edge relational attention and subject-invariant learning.

4.1 Preprocessing and Feature Representation

We use a standard EEG preprocessing procedure that retains physiologically meaningful rhythms while suppressing broadband noise. Each channel is band-pass filtered to

0.5–50 Hz using a zero-phase IIR filter and Kalman smoothing [Nalatore *et al.*, 2009] to reduce measurement noise. We segment each recording into non-overlapping 8s windows $\{\mathbf{w}_\tau\}_{\tau=1}^{T_{\max}}$. For each window, we isolate the canonical EEG bands

$$\mathcal{B} = \{\delta(0.5\text{–}4),\, \theta(4\text{–}8),\, \alpha(8\text{–}13),\, \beta(13\text{–}30),\, \gamma(30\text{–}50)\}\,\text{Hz}, \tag{1}$$

using zero-phase filtering and compute differential entropy (DE) [Duan *et al.*, 2013] per electrode and band under a Gaussian assumption:

$$\mathrm{DE}(\mathbf{w}_{\tau,b}) = \tfrac{1}{2}\ln\!\big(2\pi e\,\sigma^2_{\tau,b}\big),\ \sigma^2_{\tau,b} = \tfrac{1}{L}\sum_{i=1}^{L}\big(\mathbf{w}_{\tau,b}[i] - \bar{\mathbf{w}}_{\tau,b}\big)^2. \tag{2}$$

Stacking DE features over $N{=}19$ electrodes and $|\mathcal{B}|{=}5$ bands yields $\mathbf{F} \in \mathbb{R}^{T_{\max}\times N\times 5}$. We apply per-recording z-score normalization for each channel-band series, pad sequences to $T_{\max}$, and use a binary mask $\mathbf{M} \in \{0,1\}^{T_{\max}}$ to ignore padded windows in attention and pooling. Minibatches are represented as $\mathbf{F}_{\text{batch}} \in \mathbb{R}^{B\times T_{\max}\times N\times 5}$ with masks $\mathbf{M}_{\text{batch}} \in \{0,1\}^{B\times T_{\max}}$.

4.2 Window-Wise Functional Connectivity & Bipartite Graph Construction

Dementia-related EEG alterations may manifest not only in regional activity but also in disrupted functional coupling across scalp regions. To expose interaction structure explicitly, we estimate functional connectivity *within each window* and convert the resulting electrode graph into a bipartite incidence representation, allowing electrode activity and interactions to be modeled as distinct token types. For each window τ and frequency band $b \in \mathcal{B}$, we compute Pearson correlation across time samples of the band-limited signals [Zheng *et al.*, 2025]:

$$\rho_{ij}^{(\tau,b)} = \mathrm{Corr}(x_{\tau,i,b}[t],\, x_{\tau,j,b}[t])\,, \quad t = 1,\ldots,L, \tag{3}$$

where $x_{\tau,i,b} \in \mathbb{R}^L$ denotes the band-pass filtered signal at electrode i within window τ. We convert band-wise correlations to nonnegative edge strengths using a contrast-enhancing power transform:

$$A_{ij}^{(\tau,b)} = \left|\rho_{ij}^{(\tau,b)}\right|^p, \quad p \geq 1, \tag{4}$$

where larger p increases the separation between strong and weak couplings. In all experiments we set $p \in \{1,2,3\}$ (selected on the validation folds), and then we sparsify each window graph by retaining the top-K edges per node based on $A_{ij}^{(\tau,b)}$. We aggregate bands by averaging retained $A^{(\tau,b)}$ across $b \in \mathcal{B}$ to form a single sparse adjacency per window. The retained edges define the interaction structure for window τ and initialize relation features for edge tokens.

We represent each retained electrode pair (i,j) as an explicit edge token e_{ij} and form a bipartite incidence graph $\mathcal{G}_B = (\mathcal{V}\cup\mathcal{E},\mathcal{C})$, where $\mathcal{V}$ are electrode nodes, $\mathcal{E}$ are edge tokens, and incidence links are

$$\mathcal{C} = \{\, i \leftrightarrow e_{ij},\ j \leftrightarrow e_{ij} \mid i < j \,\}. \tag{5}$$

This construction preserves interaction structure and avoids compressing functional relations into scalar adjacency weights during representation learning.

4.3 ST-BiT: Spatio-Temporal Bipartite Transformer

ST-BiT integrates three complementary dimensions of dementia-relevant EEG structure: (i) spatial topographies across electrodes, (ii) temporal dynamics across windows, and (iii) spectral variation across frequency bands. Its core architectural contribution is *incidence-masked node↔edge cross-attention*, which treats functional interactions as first-class entities and supports robust subject-wise learning when combined with domain-adversarial alignment.

Incidence-Masked Node↔Edge Cross-Attention. For each window τ, let $\mathbf{V}_\tau \in \mathbb{R}^{N\times d}$ denote electrode *node tokens* and $\mathbf{E}_\tau \in \mathbb{R}^{M_\tau\times d}$ denote *edge tokens* for the M_τ retained functional connections (Section 4.2). Let $\mathbf{I}_\tau \in \{0,1\}^{N\times M_\tau}$ be the incidence mask, where $(\mathbf{I}_\tau)_{i,m} = 1$ if node i is incident to edge m. ST-BiT alternates incidence-masked multi-head cross-attention (MHCA) updates:

$$\mathbf{E}_\tau \leftarrow \mathrm{MHCA}\big(\mathbf{Q}=\mathbf{E}_\tau,\ \mathbf{K}=\mathbf{V}_\tau,\ \mathbf{V}=\mathbf{V}_\tau;\ \text{mask}=\mathbf{I}_\tau^\top\big), \tag{6}$$

$$\mathbf{V}_\tau \leftarrow \mathrm{MHCA}(\mathbf{Q}=\mathbf{V}_\tau,\ \mathbf{K}=\mathbf{E}_\tau,\ \mathbf{V}=\mathbf{E}_\tau;\ \text{mask}=\mathbf{I}_\tau)\,. \tag{7}$$

Non-incident query-key pairs are suppressed by additive $-\infty$ masking of attention logits. Each MHCA block is followed by residual connections, layer normalization, and a position-wise feed-forward network. This mechanism ensures that each relation token is updated only from its endpoint electrodes, and each electrode aggregates information only from incident relations.

Spatial Module (Within-Window). For each window τ, we embed the 5-band DE features into d dimensions:

$$\mathbf{V}_\tau^{(0)} = \mathbf{F}_\tau\mathbf{W}_e + \mathbf{b}_e, \qquad \mathbf{F}_\tau \in \mathbb{R}^{N\times 5},\ \mathbf{V}_\tau^{(0)} \in \mathbb{R}^{N\times d}. \tag{8}$$

We apply L_s transformer encoder layers to capture within-window spatial structure, and then apply the bipartite node↔edge attention block (Eqs. 6–7) to integrate electrode activity and functional interactions.

Temporal Module (Across Windows) with Temporal Edges. Let $\widehat{\mathbf{V}}_\tau$ denote the spatially updated electrode tokens for window τ. We summarize each window into a window token and construct a masked sequence $\mathbf{S} \in \mathbb{R}^{T_{\max}\times d}$:

$$\mathbf{s}_\tau = \tfrac{1}{N}\sum_{i=1}^{N}\widehat{\mathbf{V}}_{\tau,i}, \qquad \mathbf{S} = [\mathbf{s}_1,\ldots,\mathbf{s}_{T_{\max}}]. \tag{9}$$

We apply L_t temporal transformer layers with the padding mask $\mathbf{M}$ to model non-stationary dynamics across windows. To mirror relation-centric modeling over time, we define sparse *temporal edge tokens* by connecting each window τ to its $\pm r$ neighbors and apply the same incidence-masked node↔edge attention between window tokens and temporal edge tokens using an analogous incidence mask.

Stimulation/Spectral Band Attention. In Figure 1, the "stimulation module" denotes the spectral band-attention component used to adaptively weight canonical EEG rhythms under the eyes-open photic-stimulation recording condition,

Model	Binary (CN vs. AD+FTD)					Multi-class (CN vs. AD vs. FTD)				
	ACC	F1	Prec.	Rec.	AUC	ACC	F1	Prec.	Rec.	AUC
Classical ML Models										
SVM [Jakkula, 2006]	79.48±5.9	0.72±0.11	0.84±0.07	0.72±0.09	0.86±0.07	45.10±15.2	0.34±0.12	0.37±0.11	0.40±0.14	0.59±0.14
Logistic Regression [Peng *et al.*, 2002]	76.14±9.2	0.72±0.13	0.73±0.11	0.73±0.13	0.85±0.05	51.96±12.9	0.50±0.12	0.50±0.12	0.52±0.13	0.66±0.12
KNN [Guo *et al.*, 2003]	74.97±9.3	0.69±0.13	0.72±0.13	0.69±0.11	0.71±0.11	47.39±13.6	0.42±0.11	0.46±0.17	0.45±0.13	0.63±0.13
Random Forest [Rigatti, 2017]	77.19±5.4	0.70±0.09	0.81±0.07	0.70±0.08	0.75±0.11	44.12±8.8	0.39±0.09	0.39±0.09	0.42±0.10	0.63±0.15
Gradient Boosting [Ke *et al.*, 2017]	66.01±5.7	0.60±0.08	0.61±0.09	0.60±0.07	0.68±0.05	41.83±11.9	0.39±0.10	0.41±0.11	0.41±0.10	0.58±0.12
AdaBoost [Schapire, 2013]	71.57±6.1	0.66±0.06	0.71±0.10	0.67±0.07	0.75±0.04	44.18±15.1	0.43±0.15	0.45±0.15	0.43±0.14	0.60±0.12
XGBoost [Chen and Guestrin, 2016]	69.48±7.9	0.65±0.11	0.66±0.10	0.66±0.10	0.74±0.07	43.01±11.5	0.39±0.10	0.38±0.10	0.41±0.11	0.61±0.12
MLP [Murtagh, 1991]	70.46±7.3	0.57±0.14	0.63±0.19	0.60±0.11	0.66±0.10	45.29±12.3	0.40±0.08	0.44±0.11	0.43±0.10	0.63±0.16
Graph Neural Networks										
GCN [Kipf and Welling, 2017]	71.57±5.2	0.53±0.12	0.63±0.25	0.58±0.08	0.78±0.07	43.07±4.8	0.27±0.07	0.32±0.15	0.37±0.06	0.60±0.20
GraphSAGE [Hamilton *et al.*, 2017]	72.75±6.4	0.59±0.12	0.72±0.21	0.61±0.08	0.76±0.04	46.47±13.9	0.35±0.18	0.40±0.22	0.42±0.15	0.68±0.15
GAT [Veličković *et al.*, 2018]	73.86±4.5	0.68±0.02	0.77±0.09	0.69±0.02	0.77±0.04	46.41±8.8	0.33±0.10	0.30±0.09	0.41±0.09	0.61±0.17
ST-BiT (Ours)	**93.01±6.87**	**0.93±0.06**	**0.97±0.05**	**0.90±0.10**	**0.95±0.05**	**76.14±2.05**	**0.75±0.03**	**0.80±0.03**	**0.76±0.04**	**0.82±0.04**

Table 2: Dementia-EEG classification performance across subject-wise stratified 5-fold cross-validation (Mean ± SD). Best results in each column are highlighted in soft yellow and bold. *Note: Experiments use OpenNeuro ds006036 eyes-open EEG with photic stimulation. For consistency with prior work, we report mean±SD across the five outer folds for* all *methods, including ST-BiT.*

rather than a separate external-stimulus input. Within each electrode-window, we apply lightweight attention over the five band embeddings to emphasize diagnostically informative spectral components. For electrode i in window τ, each scalar DE feature $f_{\tau,i,b} \in \mathbb{R}$ is independently projected to a d-dimensional band embedding:

$$\mathbf{g}_{\tau,i,b} = f_{\tau,i,b}\,\mathbf{w}_b + \mathbf{c}_b, \quad \mathbf{w}_b,\, \mathbf{c}_b \in \mathbb{R}^d, \tag{10}$$

where $\mathbf{w}_b$ and $\mathbf{c}_b$ are learnable parameters shared across electrodes and windows. We then compute attention over these band embeddings:

$$\alpha_{bb'} = \frac{\exp\!\big(\mathbf{g}_{\tau,i,b}\mathbf{g}_{\tau,i,b'}^{\top}/\sqrt{d}\big)}{\sum_{b''\in\mathcal{B}} \exp\!\big(\mathbf{g}_{\tau,i,b}\mathbf{g}_{\tau,i,b''}^{\top}/\sqrt{d}\big)}, \tag{11}$$

$$\tilde{\mathbf{g}}_{\tau,i,b} = \sum_{b'\in\mathcal{B}} \alpha_{bb'}\,\mathbf{g}_{\tau,i,b'}. \tag{12}$$

Masked Pooling & Prediction. Let $\mathbf{Z} \in \mathbb{R}^{T_{\max}\times N\times d}$ denote the final electrode representations. We apply masked mean pooling over valid windows and electrodes:

$$\mathbf{z} = \frac{1}{\sum_{\tau=1}^{T_{\max}} M_\tau} \sum_{\tau=1}^{T_{\max}} M_\tau \left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{Z}_{\tau,i}\right) \in \mathbb{R}^d, \tag{13}$$

and predict diagnosis using an MLP classifier $f_{\mathrm{cls}}(\mathbf{z})$ for binary screening or three-class subtyping.

Domain-Adversarial Subject-Invariant Learning. EEG exhibits strong inter-subject shifts that can cause models to overfit subject identity cues. To encourage subject-invariant representations, we adopt domain-adversarial learning with a gradient reversal layer (GRL). Using the pooled embedding $\mathbf{z}$ (Eq. 13), we predict diagnosis and subject identity:

$$\hat{y} = f_{\mathrm{cls}}(\mathbf{z}), \qquad \hat{s} = f_{\mathrm{dom}}(\mathrm{GR}_\lambda(\mathbf{z})), \tag{14}$$

where GR_λ acts as identity in the forward pass and multiplies gradients by $-\lambda$ during backpropagation. The min–max objective is

$$\min_{\theta_f,\theta_y} \max_{\theta_s} \; \mathcal{L}_{\mathrm{cls}}(y,\hat{y}) - \lambda\,\mathcal{L}_{\mathrm{dom}}(s,\hat{s}), \tag{15}$$

equivalently implemented as $\mathcal{L} = \mathcal{L}_{\mathrm{cls}} + \lambda\,\mathcal{L}_{\mathrm{dom}}$ with GRL. We use cross-entropy for both terms. We use DANN as subject-invariant adversarial regularization [Ganin *et al.*, 2016; Zhang *et al.*, 2023]: the discriminator is trained only on training-subject IDs within each fold (never test IDs) with a lightweight MLP, and λ is warmed up from 0 to 1 during early epochs for stability. Here, s denotes the subject identifier within the training fold (one domain per subject).

5 Experimental Setup

Evaluation Protocol. We evaluate cross-subject generalization using stratified 5-fold cross-validation [Varoquaux, 2018] with subject-wise splits: all windows from a subject are assigned to exactly one fold. In each outer fold, 80% of subjects are used for training and 20% are held out for testing while preserving class ratios. Hyperparameters are selected via nested cross-validation on training subjects only. We report mean±SD across the five outer folds for all experiments.

Prediction Granularity (Subject-Level). All metrics are computed at the subject level. Each recording is segmented into windows, processed by the model, and aggregated into a single embedding via masked pooling over valid windows (Eq. 13), yielding one prediction per subject. Classical ML baselines use the same evaluation unit by mean-pooling window features within each subject (using the same padding mask) under identical subject-wise folds.

Metrics. We report accuracy (ACC), macro-F_1, precision, recall, and AUC. For the three-class task, AUC is computed one-vs-rest and macro-averaged.

Implementation Details. All models are implemented in PyTorch and trained with AdamW (learning rate 10^{-3} with cosine decay, batch size 8, weight decay 10^{-4}, gradient clipping 1.0) with early stopping on validation accuracy. Inputs are DE features from non-overlapping 8s windows over five bands, forming tensors of shape $(B, T_{\max}, 19, 5)$ with padding masks; all baselines and ST-BiT use identical subject-wise folds and are tuned on training subjects only.

Method	ACC	F_1	Ref.
EEGNet (all) [Lawhern *et al.*, 2018]	44.44	0.44	[Kim *et al.*, 2025]
Informer [Zhou *et al.*, 2021]	48.45	0.46	[Wang *et al.*, 2024]
EEGNet [Lawhern *et al.*, 2018]($\delta+\alpha$)	50.00	0.40	[Kim *et al.*, 2025]
Crossformer [Zhang and Yan, 2023]	50.45	0.46	[Wang *et al.*, 2024]
Transformer [Vaswani *et al.*, 2017]	50.47	0.48	[Wang *et al.*, 2024]
SVM-RBF [Cortes and Vapnik, 1995]	51.35	0.51	[Hasan *et al.*, 2024]
Medformer [Wang *et al.*, 2024]	53.27	0.51	[Wang *et al.*, 2024]
CNN-Spec. [Stefanou *et al.*, 2025]	54.28	0.49	[Kim *et al.*, 2025]
LDA [Fisher, 1936]	55.69	0.56	[Hasan *et al.*, 2024]
Lin. SVC [Cortes and Vapnik, 1995]	57.54	0.58	[Hasan *et al.*, 2024]
EfficientNet [Tan and Le, 2019]	61.11	0.49	[Kim *et al.*, 2025]
SSPNet [Kim *et al.*, 2025]	72.22	**0.72**	[Kim *et al.*, 2025]
ST-BiT (Ours)	**73.92**	0.71	*Our work*

Table 3: Comparison on OpenNeuro ds004504 (eyes-closed) for 3-class CN/AD/FTD. Rows above ST-BiT are *reported* from the cited papers and are included for context; Best values are highlighted in soft yellow and bold.

Electrode tokens are initialized by a linear projection of 5-band DE features, while edge tokens use band-averaged $|\rho|^p$ connectivity strengths after top-K sparsification projected into d dimensions; temporal edge tokens connect each window to its $\pm r$ neighbors (with r selected on inner folds), and variable edge counts are handled by padding edge tokens to $M_{\max}$ with an edge mask jointly applied with the incidence mask. We tune $p \in \{1, 2, 3\}$, $K \in \{3, 5, 7\}$, $d \in \{64, 128, 256\}$, and heads $\in \{2, 4, 8\}$, via inner-CV on training subjects only.

6 Results & Analysis

We evaluate ST-BiT under leakage-free *subject-wise* cross-validation to reflect a clinically realistic deployment setting, where a model must generalize to unseen individuals and session conditions. We first compare against classical ML and GNN baselines on the primary eyes-open ds006036 benchmark, then analyze ablations, robustness across recording state, and interpretability.

Comparison with Baseline Models Table 2 reports performance on ds006036 (eyes-open with photic stimulation). ST-BiT achieves 93.01±6.87% accuracy for binary screening (CN vs. AD+FTD) and 76.14±2.05% accuracy for three-way subtyping (CN vs. AD vs. FTD), outperforming classical machine-learning baselines and GNN variants under identical subject-wise folds. In particular, ST-BiT improves substantially over the strongest classical model (SVM) and the best-performing GNN baseline (GAT), indicating that preserving interaction structure as explicit edge tokens and integrating it with temporal dynamics provides a measurable advantage beyond standard graph message passing. While classical methods remain competitive for binary screening in some folds, their performance degrades substantially in the multi-class setting, as subtyping requires richer feature structure beyond marginal spectral summaries. In contrast, ST-BiT maintains strong macro-F_1 and AUC across both tasks, supporting the central claim that *joint spatio-temporal-spectral modeling with explicit interaction tokens* improves cross-subject dementia classification.

Ablation Study. To isolate the effect of each component, Table 4 reports ablations on the ds006036 dataset under the same leakage-free subject-wise protocol. Removing the temporal module yields the largest degradation in three-way subtyping, indicating that non-stationary window-to-window dynamics are important for differentiating AD vs. FTD beyond static spectral summaries. Removing the spatial module also reduces performance, consistent with dementia-related topographic alterations across scalp regions. Importantly, removing the bipartite relation module (i.e., replacing incidence-masked node↔edge attention with node-only processing and eliminating explicit edge tokens) consistently lowers ACC/AUC in both tasks, supporting our central claim that explicit interaction tokens mitigate relation compression relative to standard message passing. Removing the stimulation/spectral band-attention module reduces performance, suggesting that adaptive emphasis over canonical EEG frequency bands improves robustness under the eyes-open photic-stimulation setting. Finally, removing domain-adversarial learning (DANN) hurts both screening and subtyping, indicating that encouraging subject-invariant representations is beneficial under strict subject-wise evaluation.

Eyes-Closed Evaluation & Generalizability. Our main results are on ds006036 (eyes-open with photic stimulation) under leakage-free subject-wise stratified 5-fold CV with nested model selection (Sec. 5), reported in Tables 2 and 4. To assess robustness across recording state and connect with prior work on this cohort, we additionally evaluate ST-BiT on ds004504 (eyes-closed resting-state) under the same leakage-free subject-wise protocol. ST-BiT achieves 73.92% accuracy for 3-class CN/AD/FTD classification, remaining competitive against prior work on ds004504 (Table 3).

Error Analysis. We analyze confusion matrices on ds004504 (Figure 4) to characterize residual failure modes. Binary screening shows high CN recall and high patient precision, indicating reliable separation of cognitively normal controls from dementia cases. In contrast, the 3-class setting shows most remaining errors arise from AD-FTD confusion, consistent with overlapping scalp-level EEG signatures and heterogeneity within clinical subtypes. This suggests that improved subtype discrimination may require richer interaction measures and/or complementary clinical metadata.

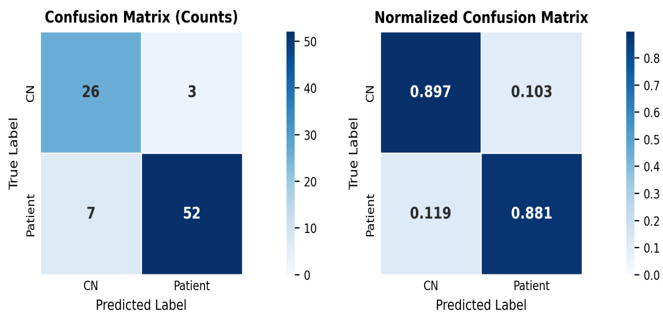
Window-Length Sensitivity. We vary the analysis window length (4s, 8s, 12s, 16s) under the same subject-wise protocol (Table 5). Performance peaks at 8s, balancing spectral resolution against effective sample count; shorter windows reduce frequency resolution and connectivity stability, while longer windows decrease training instances and may average over non-stationary dynamics. We use 8s windows throughout.

Sensitivity to the Number of Folds. Table 6 reports sensitivity to the number of subject-wise folds ($K \in \{3, 5, 7\}$). Performance is strongest at $K = 5$, providing a favorable balance between training-set size and stability for this cohort.

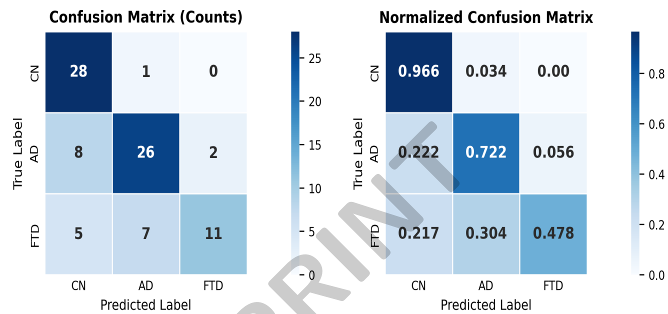
Interpretability & Clinical Plausibility. We summarize spatial attention over electrodes as a lightweight interpretability analysis. Figure 5 shows consistent emphasis on frontopolar sensors (Fp1/Fp2), followed by central/parietal/temporal electrodes (C4, P3, T5), compatible with prior reports linking dementia to frontal and fronto-temporal dysfunction [Jeong,

Configuration	Binary (CN vs. AD + FTD)					Multi-class (CN vs. AD vs. FTD)				
	ACC (%)	F1	Prec.	Recall	AUC	ACC (%)	F1	Prec.	Recall	AUC
Without Spatial	80.52±6.34	0.866±0.04	0.814±0.06	0.933±0.08	0.771±0.12	62.75±13.50	0.617±0.15	0.651±0.17	0.627±0.14	0.687±0.12
Without Temporal	79.41±5.15	0.854±0.03	0.828±0.09	0.898±0.10	0.778±0.10	55.88±8.06	0.507±0.09	0.498±0.11	0.559±0.08	0.640±0.09
Without Band Attention	84.97±9.74	0.887±0.08	0.888±0.08	0.898±0.12	0.917±0.07	61.57±14.96	0.608±0.16	0.659±0.16	0.616±0.15	0.698±0.13
Without Bipartite	86.21±8.13	0.903±0.05	0.863±0.05	**0.948±0.07**	0.889±0.08	67.06±7.14	0.655±0.08	0.713±0.07	0.671±0.07	0.763±0.07
Without DANN	86.21±7.94	0.899±0.06	0.897±0.09	0.917±0.11	0.859±0.08	68.24±4.48	0.673±0.06	0.720±0.07	0.682±0.04	0.761±0.05
Full Model (ST-BiT)	**93.01±6.87**	**0.93±0.06**	**0.97±0.05**	0.90±0.10	**0.95±0.05**	**76.14±2.05**	**0.75±0.03**	**0.80±0.03**	**0.76±0.04**	**0.82±0.04**

Table 4: Ablation study on eyes-open dataset: effect of removing each component (mean±SD). Best values in each column are highlighted in soft yellow and bold.



(a) Binary confusion matrix (CN vs. Patient)



(b) Three-way confusion matrix (CN vs. AD vs. FTD)

Figure 4: Confusion-matrix evaluation on ds004504 (eyes-closed). (a) CN vs. AD+FTD and (b) CN vs. AD vs. FTD. Left panels show fold-aggregated counts; right panels show row-normalized values (recall).

Window	Binary ACC	Multi-class ACC
4 s	92.1	70.3
8 s	**93.0**	**76.1**
12 s	88.8	67.9
16 s	88.5	68.3

Table 5: Window-length sensitivity. Best values are highlighted.

K	Task	ACC (%)	macro-F_1	AUC
$K = 3$	Binary	87.59	0.908	0.890
	Multi-class	67.05	0.645	0.736
$K = 5$	Binary	**93.01**	0.930	**0.948**
	Multi-class	**76.14**	**0.749**	**0.819**
$K = 7$	Binary	90.84	**0.937**	0.882
	Multi-class	71.61	0.652	0.772

Table 6: ST-BiT cross-validation sensitivity on ds006036 eyes-open. Best values within each task are highlighted.

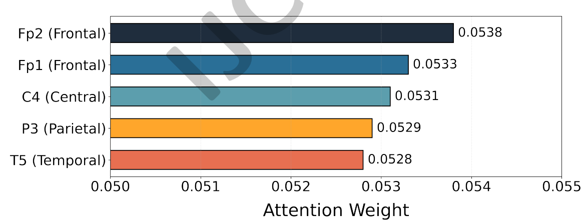


Figure 5: Top-5 Electrode Importance from ST-BiT's Spatial Attention on ds006036 eyes-open.

2004; Musaeus *et al.*, 2019]. Attention scores are descriptive hypothesis-generating signals, not causal biomarkers; clinical validation requires targeted studies with appropriate controls.

7 Conclusion

We introduced **ST-BiT**, a spatio-temporal bipartite transformer for EEG-based dementia screening and subtyping under strict *subject-wise* evaluation. The core idea is to model functional interactions as explicit *edge tokens* and update them using incidence-masked cross-attention between node and edge tokens, preserving interaction structure rather than compressing connectivity into scalar weights. Together with temporal modeling, lightweight band attention, and domain-adversarial alignment for subject-invariant learning, ST-BiT achieves strong performance on the eyes-open ds006036 benchmark under leakage-free nested cross-validation and remains competitive against representative classical, GNN, and recent deep learning baselines. Overall, the results highlight bipartite relation modeling as a practical inductive bias for robust EEG-based dementia phenotyping.

Despite these encouraging results, some challenges remain. The main remaining difficulty is AD–FTD separation in the three-class setting, which could benefit from richer interaction features and optional clinical context when available. The current edge initialization relies on within-window correlation; future work can incorporate phase-based or directed connectivity and explore alternative sparsification strategies to improve robustness. Finally, attention summaries should be interpreted as hypothesis-generating rather than clinically validated biomarkers, motivating follow-up validation with region-level aggregation, source localization, and links to cognitive measures or neuroimaging.

Ethics Statement

This work uses publicly available, de-identified EEG datasets and does not involve new human-subject data collection. The proposed model is intended for research on EEG-based dementia screening and subtyping, not for standalone clinical diagnosis. All results should be interpreted with caution because the cohort is small, single-center, and lacks external validation. Attention-based interpretations are hypothesis-generating and should not be treated as clinically validated biomarkers.

Contribution Statement

Siddhant Ujjain and Vaibhav Kagathara contributed equally to this work. Authors listed in alphabetical order of first name.

Acknowledgments

We gratefully acknowledge the support of Prof. Sandeep Kumar through the Core Research Grant (CRG, RP04820), which has been valuable in enabling this work.

References

[Aranda *et al.*, 2021] María P Aranda, Ian N Kremer, Ladson Hinton, Julie Zissimopoulos, Rachel A Whitmer, Cynthia Huling Hummel, Laura Trejo, and Chanee Fabius. Impact of dementia: health disparities, population trends, care interventions, and economic costs. *Journal of the American Geriatrics Society*, 69(7):1774–1783, 2021.

[Chen and Guestrin, 2016] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016.

[Cortes and Vapnik, 1995] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[Cui *et al.*, 2023] Weigang Cui, Yulan Ma, Jianxun Ren, Jingyu Liu, Guolin Ma, Hesheng Liu, and Y. Li. Personalized functional connectivity based spatio-temporal aggregated attention network for mci identification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:2257–2267, 2023.

[Duan *et al.*, 2013] Ruo-Nan Duan, Jia-Yi Zhu, and Bao-Liang Lu. Differential entropy feature for eeg-based emotion classification. In *2013 6th international IEEE/EMBS conference on neural engineering (NER)*, pages 81–84. IEEE, 2013.

[Fisher, 1936] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

[Ganin *et al.*, 2016] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

[Guo *et al.*, 2003] Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer. Knn model-based approach in classification. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 986–996. Springer, 2003.

[Hamilton *et al.*, 2017] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

[Hasan *et al.*, 2024] Wali Hasan, Sheraz Khan, and Abbas Sohrabpour. Classifying alzheimer's disease and dementia patients using non-invasive eeg biomarkers. *medRxiv*, pages 2024–10, 2024.

[Hu *et al.*, 2020] Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. Heterogeneous graph transformer. In *Proceedings of the web conference 2020*, pages 2704–2710, 2020.

[Jakkula, 2006] Vikramaditya Jakkula. Tutorial on support vector machine (svm). *School of EECS, Washington State University*, 37(2.5):3, 2006.

[Jeong, 2004] Jaeseung Jeong. Eeg dynamics in patients with alzheimer's disease. *Clinical neurophysiology*, 115(7):1490–1505, 2004.

[Ke *et al.*, 2017] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

[Kim *et al.*, 2022] Jinwoo Kim, Dat Nguyen, Seonwoo Min, Sungjun Cho, Moontae Lee, Honglak Lee, and Seunghoon Hong. Pure transformers are powerful graph learners. *Advances in Neural Information Processing Systems*, 35:14582–14595, 2022.

[Kim *et al.*, 2025] Ho-Jung Kim, Dogeun Park, Jeong-Woo Jang, Young-Gi Ju, and Dong-Ok Won. Sspnet: Towards feasible spatio-spectral portraits-based deep learning framework for neurodegenerative disease multi-classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2025.

[Kipf and Welling, 2017] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

[Klepl *et al.*, 2022] Dominik Klepl, Fei He, Min Wu, Daniel J. Blackburn, and Ptolemaios Sarrigiannis. 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.

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

[Li *et al.*, 2023] Fanshi Li, Zhihui Wang, Yifan Guo, Congcong Liu, Yanjie Zhu, Yihang Zhou, Jun Li, Dong Liang, and Haifeng Wang. Dynamic dual-graph fusion convolutional network for alzheimer's disease diagnosis. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 675–679, 2023.

[Miltiadous *et al.*, 2024] Andreas Miltiadous, Katerina D. Tzimourta, Theodora Afrantou, Panagiotis Ioannidis, Nikolaos Grigoriadis, Dimitrios G. Tsalikakis, Pantelis Angelidis, Markos G. Tsipouras, Evripidis Glavas, Nikolaos Giannakeas, and Alexandros T. Tzallas. "a dataset of eeg recordings from: Alzheimer's disease, frontotemporal dementia and healthy subjects", 2024.

[Murtagh, 1991] Fionn Murtagh. Multilayer perceptrons for classification and regression. *Neurocomputing*, 2(5-6):183–197, 1991.

[Musaeus *et al.*, 2019] Christian Sandøe Musaeus, Knut Engedal, Peter Høgh, Vesna Jelic, Morten Mørup, Mala Naik, Anne-Rita Oeksengaard, Jon Snaedal, Lars-Olof Wahlund, Gunhild Waldemar, et al. Oscillatory connectivity as a diagnostic marker of dementia due to alzheimer's disease. *Clinical Neurophysiology*, 130(10):1889–1899, 2019.

[Nalatore *et al.*, 2009] Hariharan Nalatore, Mingzhou Ding, and Govindan Rangarajan. Denoising neural data with state-space smoothing: method and application. *Journal of neuroscience methods*, 179(1):131–141, 2009.

[Ntetska *et al.*, 2025] Aimilia Ntetska, Andreas Miltiadous, Markos G. Tsipouras, Katerina D. Tzimourta, Theodora Afrantou, Panagiotis Ioannidis, Dimitrios G. Tsalikakis, Konstantinos Sakkas, Emmanouil D. Oikonomou, Nikolaos Grigoriadis, Pantelis Angelidis, Nikolaos Giannakeas, and Alexandros T. Tzallas. A complementary dataset of scalp eeg recordings featuring participants with alzheimer's disease, frontotemporal dementia, and healthy controls, obtained from photostimulation eeg. *Data*, 10(5), 2025.

[Paitel *et al.*, 2025] Elizabeth Paitel, Christian Otteman, Mary Polking, Henry Licht, and Kristy Nielson. Functional and effective eeg connectivity patterns in alzheimer's disease and mild cognitive impairment: a systematic review. *Frontiers in Aging Neuroscience*, 17, 02 2025.

[Peng *et al.*, 2002] Chao-Ying Joanne Peng, Kuk Lida Lee, and Gary M Ingersoll. An introduction to logistic regression analysis and reporting. *The journal of educational research*, 96(1):3–14, 2002.

[Rigatti, 2017] Steven J Rigatti. Random forest. *Journal of insurance medicine*, 47(1):31–39, 2017.

[Schapire, 2013] Robert E Schapire. Explaining adaboost. In *Empirical inference: festschrift in honor of vladimir N. Vapnik*, pages 37–52. Springer, 2013.

[Stefanou *et al.*, 2025] Konstantinos Stefanou, Katerina D Tzimourta, Christos Bellos, Georgios Stergios, Konstantinos Markoglou, Emmanouil Gionanidis, Markos G Tsipouras, Nikolaos Giannakeas, Alexandros T Tzallas, and Andreas Miltiadous. A novel cnn-based framework for alzheimer's disease detection using eeg spectrogram representations. *Journal of Personalized Medicine*, 15(1):27, 2025.

[Tan and Le, 2019] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[Varoquaux, 2018] Gaël Varoquaux. Cross-validation failure: Small sample sizes lead to large error bars. *Neuroimage*, 180:68–77, 2018.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Veličković *et al.*, 2018] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

[Wang *et al.*, 2024] Yihe Wang, Nan Huang, Taida Li, Yujun Yan, and Xiang Zhang. Medformer: A multi-granularity patching transformer for medical time-series classification. *Advances in Neural Information Processing Systems*, 37:36314–36341, 2024.

[Yun *et al.*, 2019] Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang, and Hyunwoo J Kim. Graph transformer networks. *Advances in neural information processing systems*, 32, 2019.

[Zhang and Yan, 2023] Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The eleventh international conference on learning representations*, 2023.

[Zhang *et al.*, 2023] YiFan Zhang, Xue Wang, Jian Liang, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. Free lunch for domain adversarial training: Environment label smoothing. In *International Conference on Learning Representations (ICLR)*, 2023.

[Zheng *et al.*, 2025] Huang Zheng, Han Xiao, Yinan Zhang, Haozhe Jia, Xing Ma, and Yiqun Gan. Time-frequency functional connectivity alterations in alzheimer's disease and frontotemporal dementia: An eeg analysis using machine learning. *Clinical Neurophysiology*, 170:110–119, 2025.

[Zhou *et al.*, 2021] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11106–11115, 2021.