University of Padova

Department of Information Engineering

Master Thesis in Bioengineering

Integrating EEG and Clinical Data with Multimodal Deep Learning for Dementia Diagnosis

Supervisor
Prof. Manfredo Atzori
University of Padova

Dr. Andrea Zanola
University of Padova

Master Candidate
Rachele Costantini

Student ID
2130634

Academic Year
2024-2025

Alla mia stella lassù,
che mi ha ispirata a trasformare l'ingegno in dono per il prossimo,
dedico l'amore messo in questo lavoro e tutto il resto.

Abstract

Accurate dementia classification remains a major challenge in clinical neuroscience. Conventional diagnostic tools often provide only partial insights rather than a comprehensive picture of the patient, limiting their ability to distinguish between Alzheimer's disease (AD), frontotemporal dementia (FTD), and cognitively normal (CN) individuals. Multimodal learning offers a promising alternative by integrating heterogeneous biomedical data to construct richer and more informative patient representations.

In this work, we combine electroencephalography (EEG) and clinical data (CD) from the publicly available BrainLat dataset to explore both unimodal and multimodal learning strategies. EEG signals were converted into ROI-based spectrograms to emphasize spatially localized spectral dynamics, while clinical variables were standardized into compact feature vectors. Building on these inputs, we implemented both unimodal architectures and multimodal fusion strategies (late and middle fusion), refined through specialized training strategies, regularization techniques, and structured hyperparameter search.

Our results confirm EEG as the most discriminative unimodal modality, with ResNet3D outperforming its 2D counterpart through volumetric spatio-temporal modeling. More importantly, multimodal fusion consistently surpassed unimodal baselines: middle fusion achieved the highest balanced accuracy and lowest variability, while lightweight late fusion improved performance without increasing complexity.

These findings underscore the clinical value of EEG and CD, the advantages of spatio-temporal representations, and the robustness of moldel complexity. Key strengths include the custom ROI-based EEG feature extraction pipeline, which enhanced biological plausibility, and the extensive benchmarking framework covering a wide range of architectures and strategies. The principal limitation remains the relatively small cohort. Future research should refine EEG feature extraction to reflect disease-specific patterns and explore advanced middle-fusion mechanisms such as gating with projection layers.

Taken together, this work highlights the potential of multimodal learning to enhance the accuracy, robustness, and clinical applicability of dementia diagnosis.

La classificazione accurata delle demenze rimane una sfida cruciale nelle neuroscienze cliniche. Gli strumenti diagnostici convenzionali forniscono spesso soltanto informazioni parziali, senza restituire un quadro complessivo del paziente, limitandone l'efficacia nel distinguere tra malattia di Alzheimer (AD), demenza frontotemporale (FTD) e soggetti cognitivamente normali (CN). L'apprendimento multimodale rappresenta un'alternativa promettente, poiché integra dati biomedici eterogenei per costruire rappresentazioni del paziente più ricche e informative.

In questo lavoro vengono combinati elettroencefalografia (EEG) e dati clinici (CD) provenienti dal dataset pubblico BrainLat, al fine di esplorare strategie di apprendimento unimodale e multimodale. I segnali EEG sono stati trasformati in spettrogrammi ROI-based, così da enfatizzare le dinamiche spettrali localizzate nello spazio, mentre le variabili cliniche sono state standardizzate in vettori di caratteristiche compatti. Su questi input sono state implementate sia architetture unimodali sia strategie di fusione multimodale (late e middle fusion), affinate tramite strategie di addestramento specifiche, tecniche di regolarizzazione e una ricerca strutturata degli iperparametri.

I risultati confermano l'EEG come la modalità unimodale più discriminativa, con ResNet3D che supera la sua controparte 2D grazie alla modellazione spaziotemporale volumetrica. Ancora più rilevante, la fusione multimodale ha superato costantemente le baseline unimodali: la middle fusion ha ottenuto la massima accuratezza bilanciata e la minore variabilità, mentre la late fusion ha migliorato le prestazioni senza aumentare la complessità architetturale.

Questi risultati sottolineano il valore clinico di EEG e CD, i vantaggi delle rappresentazioni spaziotemporali e la solidità delle strategie di fusione. Tra i punti di forza emergono la pipeline personalizzata di estrazione delle feature EEG basata su ROI, che ne ha incrementato la plausibilità biologica, e l'ampio framework di benchmarking, che ha coperto numerose architetture e strategie. La principale limitazione resta la dimensione relativamente ridotta della coorte. Futuri studi dovranno affinare ulteriormente l'estrazione delle feature EEG in funzione dei pattern patologici specifici ed esplorare meccanismi avanzati di middle fusion, come il gating con layer di proiezione.

Nel complesso, questo lavoro evidenzia il potenziale dell'apprendimento multimodale nel migliorare accuratezza, robustezza e applicabilità clinica della diagnosi di demenza.

Contents

Listing of figures

Listing of tables

List of acronyms

AD . . . . . Alzheimer's Disease

FTD . . . . . Frontotemporal Dementia

MCI . . . . . Mild Cognitive Impairment

EEG . . . . . Electroencephalogram

ML . . . . . Machine Learning

DL . . . . . Deep Learning

PET . . . . . Positron Emission Tomography

MRI . . . . . Magnetic Resonance Imaging

CSF . . . . . Cerebrospinal Fluid

LSTM . . . . . Long Short-Term Memory

NPVs . . . . . Neuropsychological Variables

SVM . . . . . Support Vector Machine

APOE . . . . . Apolipoprotein E

Aβ . . . . . Amyloid-Beta plaques

NFT . . . . . Neurofibrillary tangles

bvFTD . . . . . behavioral variant Frontotemporal Dementia

CN . . . . . Cognitively Normal

ATN . . . . . Amyloid deposition Tau patology Neurodegeneration

MoCA . . . . . Montreal Cognitive Assessment

IFS . . . . . INECO Frontal Screening

MMSE . . . . . Mini-Mental State Examination

PDD . . . . . Parkinson's disease dementia

FAB . . . . . Frontal Assessment Battery

AUC . . . . . Area Under the Curve

ROC . . . . . Receiver Operating Characteristic

FERT . . . . . Facial Emotion Recognition Test

ToM . . . . . Theory of Mind

ROI . . . . . Region of Interest

CNN . . . . . Convolutional Neural Network

XGBoost . . . . . eXtreme Gradient Boosting

MLP . . . . . Multilayer Perceptron

ReLU . . . . . REctified Linear Unit

GELU . . . . . Gaussian Error Linear Unit

ResNet . . . . . Residual Network

IQR . . . . . Interquartile Range

1 Introduction

1.1 Multimodal Learning for Precision Medicine in Dementia

Multimodal learning plays a pivotal role in advancing precision medicine [1]. Recent developing in biomedical technologies, together with the increasing availability of heterogeneous data, have made it feasible to combine information across multiple modalities [1]. Such integration enables to leverage the complementary properties of complex biomedical phenomena, thereby enhancing the supporting more effective medical strategies [2]. The rationale behind this line of research is that unimodal data sources provide limited insights of a clinical condition. Their integration, by contrast, allows the construction of multidimensional and information-rich representations, offering a more comprehensive profile of the patient [1].

1.1.1 Multimodal Learning Strategies

Multimodal fusion techniques are typically categorized according to the stage of integration within the modeling pipeline [1, 3, 4, 5]. Three main strategies can be distinguished: Early fusion, Middle fusion, and Late fusion [3, 4, 5]. Each of these approaches is characterized by specific methodological choices as well as advantages and limitations [2].

Early fusion, or feature-level fusion, takes place before model training, when features extracted from the various modalities are combined into a single feature vector that serves as input to the model [4, 5]. Integration may be accomplished through concatenation, pooling procedures, or gated-based units [5]. The primary advantage of Early fusion lies in its capacity to capture complementary information across modalities. However, this approach presupposes well-aligned and compatible feature sets, and it is susceptible to issues of high dimensionality and noise accumulation [1, 3].

Middle fusion, also referred to as joint or intermediate fusion, is performed during the training process, typically within deep learning model, where feature representations generated at intermediate layers are merged and then forwarded to an head classifier [4, 5]. A defining characteristic of Middle fusion is that error signals are propagated back not only to the classifier but also to the modality-specific feature extractors, thereby enabling end-to-end optimization of the

entire architecture [4, 5]. Middle fusion offers the advantage of producing more expressive and task-adapted modality representations [3]; however, it involves greater computational cost and requires careful architectural design to ensure appropriate alignment across modalities.

Late fusion, frequently termed decision-level fusion, occurs after independent models have been trained on each modality, where the separate predictions are aggregated to produce the final decision [4, 5]. Aggregation can be achieved through simple rules such as averaging, majority voting, weighted voting, or more advanced mechanisms such as meta-classifiers [5]. The attractiveness of Late fusion derives from its flexibility and relative ease of implementation, as new modalities can be incorporated without retraining the entire pipeline. Its main limitation, however, lies in its inability to model fine interactions between modalities [1].

Massimo Salvi *et al.*(2023) [3] discuss that selection among Early, Middle, and Late fusion involves a trade-off, yet it is also strongly shaped by the specific clinical question being addressed and the availability of patient data [1]. To capture emerging trends in this research field, the literature has introduced a dual-axis framework that classifies multimodal learning approaches along two complementary dimensions. The first dimension concerns the clinical domain, which can be grouped into nine categories: cognitive impairment, mental disorders, sleep health, cardiovascular diseases, COVID-19, oncology, ophthalmology, pediatric disorders, and other studies. The second dimension relates to the type of data modality, which can be divided into biomedical imaging, biosignals, and mixed modalities. Within this framework, our focus is directed toward the cognitive impairment domain, with particular attention to approaches that rely on mixed modalities, where biosignals, and clinical data are jointly integrated.

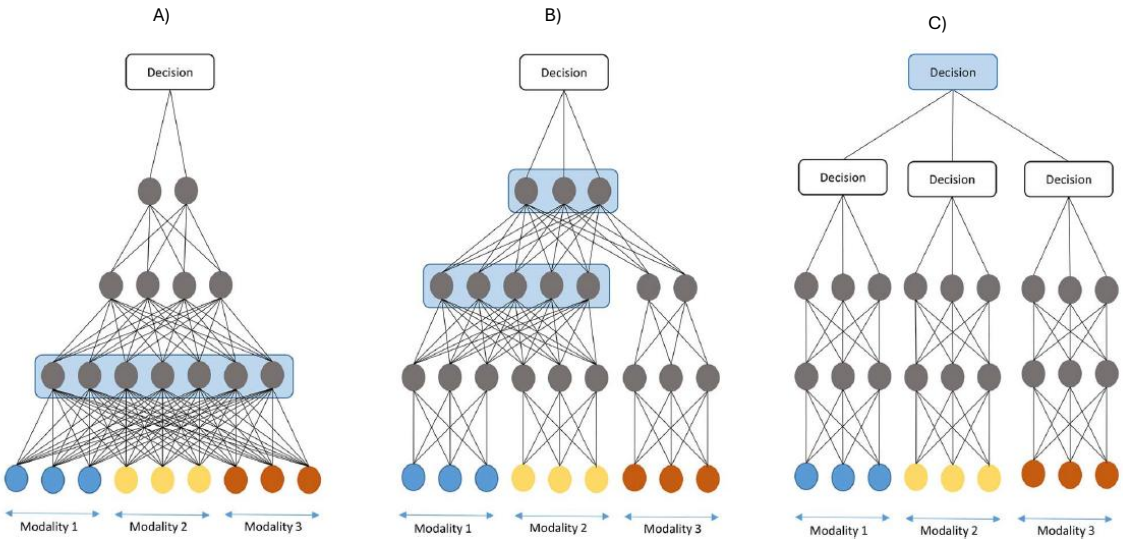

Figure 1.1: Multimodal Fusion Techniques
Illustration of three common strategies for Multimodal Learning: **(A) Early Fusion**, takes place before model training, when features extracted from the various modalities are combined into a single feature vector that serves as input to the model; **(B) Middle Fusion**, is performed during the training process, typically within deep learning model, where feature representations generated at intermediate layers are merged and then forwarded to an head classifier; **(C) Late Fusion**, occurs after independent models have been trained on each modality, where the separate predictions are aggregated to produce the final decision.
Reproduced with permission from [1].

1.1.2 Multimodal Learning for Dementia Classification

Cognitive impairment, including dementia, Mild Cognitive Impairment (MCI), and Alzheimer's Disease (AD), has emerged as a central application domain for multimodal learning. The multifactorial nature of these disorders makes them particularly suited to approaches that integrate heterogeneous data sources. Neuroimaging modalities, such as structural and functional Magnetic Resonance Imaging (MRI) or Positron Emission Tomography (PET), provide spatially precise insights into brain morphology and metabolism. Genetic and molecular biomarkers, including APOE4 genotype, amyloid/tau concentrations, or Cerebrospinal Fluid (CSF) profiles, act as static indicators of underlying biological mechanisms. Clinical and demographic assessments offer contextual information, while biosignals such as EEG capture dynamic, high–temporal-resolution measures of neural activity [1] .

Research on dementia classification has accordingly explored diverse multimodal strategies [1]. A first line of work has focused on combining imaging with clinical assessments and molecular biomarkers. Structural MRI and PET have been fused with cognitive tests to support diagnosis and progression modeling [6], while longitudinal frameworks have incorporated genetic and clinical information to predict individual symptom trajectories [7]. Other strategies have fused MRI with demographic and neuropsychological data within deep learning architectures, supporting both Alzheimer's diagnosis and prediction of MCI [8]. Recurrent neural architectures such as Long short-term memory (LSTM) have also been employed to merge sociodemographic features, APOE genotype, and MRI data, significantly improving the detection of amnestic MCI [9]. Multi-kernel and deep learning methods have been used to merge MRI, PET, CSF profiles, capturing complementary aspects of disease biology and outperforming single-modality models [10]. Attention-based models allow to capture complex cross-modal interactions between imaging,genetic, and clinical data demonstrating advantages over unimodal baselines [11]. A second direction has emphasized the contribution of EEG. When integrated with imaging and CSF biomarkers, EEG adds temporal information that enhances early diagnostic sensitivity [12]. Beyond imaging, studies have also combined EEG with clinical and cognitive metrics. For instance, models comparing EEG with MRI and Mini-Mental State Examination (MMSE) demonstrated that while MRI achieved superior classification accuracy, EEG retains value as a low-cost and portable alternative [13]. Other work has shown that EEG connectivity, particularly when combined with neuropsychological assessments, improves discrimination between bvFTD and AD, with connectivity biomarkers often providing stronger disease-specific signatures than behavioral metrics alone [14]. Language and speech data represent a further promising modality. Deep learning models integrating transcripts with neuropsychological tests have been used to detect dementia, and predict MMSE scores [15]. Within neuroimaging itself, advanced architectures such as stacked autoencoders, polynomial networks, and dense convolutional models have been designed to learn joint representations across MRI and PET, with some methods also imputing missing PET data to enhance diagnostic accuracy [16, 17, 18]. Ensemble and hybrid frameworks have also gained prominence for their flexibility in aggregating diverse modalities [19, 20]. These include random forest models integrating up to eleven different data sources with explainability components, as well as ensemble learning strategies that combine

heterogeneous base learners through stacking.

1.1.3 Multimodal Learning with EEG and Clinical Data

As discussed in Subsection 1.1.2, mixed modalities, where biosignals are integrated with clinical information, have emerged as a central direction in multimodal dementia classification [1] . Within this line of research, particular attention has been given to some studies that combine electroencephalography (EEG) with clinical and neuropsychological assessments.

Farina *et al.*(2020) [13] investigated the ability of different modalities, including resting-state EEG, structural MRI, and cognitive assessments in classifying AD, MCI, and Cognitively Normal (CN). Classification performance was evaluated using penalized logistic regression with Elastic Net regularization to prevent overfitting and support feature selection. Results showed that structural MRI achieved near-perfect discrimination of AD from CN and from MCI, while resting-state EEG provided lower accuracy. For distinguishing MCI from healthy aging, both modalities showed only moderate performance, with limited sensitivity. Combining EEG with MMSE scores offered no additional benefit beyond MMSE alone. Overall, this study highlighted the superior performance of MRI but also the potential role of EEG as a more accessible, low-cost alternative.

Dottori *et al.*(2017) [14] explored the integration of EEG connectivity measures with Neuropsychological Variables (NPVs) in distinguishing patients with bvFTD, AD, and CN. Using a novel information-sharing method and Support Vector Machine classification (SVM), the analysis revealed a characteristic hypoconnectivity pattern in frontotemporal networks specific to behavioral variant Frontotemporal Dementia (bvFTD), which was not observed in AD. Importantly, these alterations were detectable even with low-density EEG, highlighting clinical feasibility. While neuropsychological tests alone provided reasonable discrimination between bvFTD and CN, combining them with connectivity features further enhanced classification accuracy. Moreover, connectivity-based models outperformed neuropsychological measures in distinguishing bvFTD from AD, emphasizing their disease-specific discriminative value.

Together, these findings highlight both the potential and the limits of integrating EEG with clinical assessment, suggesting that while behavioral data can complement EEG in some contexts, connectivity biomarkers retain strong disease-specific discriminative value. Importantly, the studies considered so far have relied on a single fusion strategy, leaving open the possibility that alternative fusion approaches, whether at the feature, intermediate, or decision level, could further enhance performance and provide deeper insights into the effectiveness of multimodal integration.

1.2 Three Dementia Groups

This section introduces the clinical domain of the present work, which addresses on a specific use case: differentiating between three forms of dementia: Alzheimer's disease (AD), Frontotemporal dementia (FTD), and Cognitively Normal (CN). AD and FTD represent two of the most prevalent and clinically relevant forms of dementia, each defined by distinct symptomatology, affected brain regions, pathological mechanisms, risk factors, and diagnostic approaches.

Cognitively Normal (CN) subjects, by contrast, consist of healthy individuals whose EEG and cognitive profiles are representative of normal aging, thereby serving as a critical baseline for comparison. Taken together, these three groups establish a clinically meaningful framework for comparative classification studies, enabling the distinction between pathological and non-pathological aging.

1.2.1 Alzheimer's Disease

Alzheimer's disease (AD) is the most prevalent form of dementia, responsible for 60–80% of cases in older adults [21]. Clinically, it presents as a progressive decline in memory and cognitive function [22, 23]. AD is marked by a selective vulnerability of cortical neurons, with particular involvement of the medial temporal lobe (hippocampus and entorhinal cortex), the posterior parietal lobe (precuneus), the occipital lobe, and the posterior cingulate cortex. Neurons with long, poorly myelinated axons are especially susceptible to degeneration. Pathological hallmarks include the extracellular accumulation of amyloid-beta (Aβ) plaques and intracellular neurofibrillary tangles (NFTs) of hyperphosphorylated tau, leading to impaired synaptic transmission and widespread neuronal loss [22, 23]. Disease progression can be described through established schemes. The ATN framework classifies AD according to biomarkers of amyloid deposition (A), tau pathology (T), and neurodegeneration (N) [22]. Braak's [24] staging provides a topographic model, describing six progressive phases:

- Early stages (I–II) with entorhinal cortex and hippocampal involvement and associated with mild episodic memory impairment;
- Intermediate stages (III–IV) with spread to limbic and temporal cortices, leading to mild cognitive impairment (MCI)
- Advanced stages (V–VI) with neocortical extension, corresponding to severe dementia and loss of independence.

Risk factors encompass aging, genetic predispositions such as the APOE ε4, and vascular contributors. Treatment strategies remain largely symptomatic, with ongoing trials exploring disease-modifying therapies and lifestyle interventions [22, 23]. Diagnosis has evolved from purely cognitive assessments to multimodal strategies, integrating neuroimaging (MRI, PET), genetic and molecular biomarkers, clinical evaluations, and biosignals such as EEG [22].

1.2.2 Frontotemporal Dementia

Frontotemporal dementia (FTD) represents another major cause of dementia, accounting for 5–10% of cases, often with an earlier onset [21]. Warren et al. (2013) [25] emphasized that FTD predominantly affects behavior and language, unlike AD, which primarily impairs memory and cognition. The behavioral variant (bvFTD) is characterized by profound disturbances in interpersonal and executive functioning, accompanied by altered emotional responsivity and the emergence of abnormal behaviors such as apathy, disinhibition, obsessions, rituals, and stereotypies. The bvFTD is associated with degeneration in the anterior frontal lobes, including the orbitofrontal cortex and anterior cingulate cortex, and the anterior temporal regions, such as the

anterior insula and temporal cortex. Atrophy typically appears asymmetrically between the two hemispheres and tends to spare posterior cortical areas in the early stages. Neuropathologically, FTD is defined by the abnormal accumulation of aggregated proteins, most frequently phosphorylated tau or TDP-43, and less commonly FUS. These pathological inclusions drive progressive disruption of frontotemporal cortical circuits. Genetic factors contribute substantially, with 10–20% of cases exhibit an autosomal dominant inheritance pattern. Mutations in MAPT, GRN, and C9orf72 are the most common, while rarer cases involve genes such as VCP, often in association with motor syndromes like inclusion body myositis or Paget's disease of bone. Current treatments remain supportive, as no disease-modifying therapies are available for FTD, and management is limited to symptom control and comprehensive support. Non-pharmacological strategies, such as environmental adaptations, structured activities, and speech therapy, are central, while pharmacological options offer only modest or limited benefits. Diagnostic strategies rely on neuropsychological testing and neuroimaging, which often reveal focal atrophy in frontal and temporal lobes. Research efforts are increasingly directed toward biomarker discovery and the development of genetic or molecularly targeted therapies [25].

1.3 Data Modalities for Dementia Classification

This section outlines the modalities considered in the present work, electroencephalography (EEG) and clinical assessments, which provide complementary perspectives on brain function and cognitive performance. In addition to their diagnostic value, both approaches are cost-effective and time-efficient, making them particularly suitable for routine clinical use as well as for large-scale research studies. EEG recordings typically last between 10 and 20 minutes, whereas widely adopted cognitive assessments such as the Montreal Cognitive Assessment (MoCA) and the INECO Frontal Screening (IFS) generally require only about 10 minutes to administer.

1.3.1 Electroencephalography

Electroencephalography (EEG) is a low-cost and noninvasive diagnostic technique that records rhythmic brain activity with high temporal resolution, providing real-time insights into neuronal dynamics. For this reason, EEG is increasingly investigated as a promising tool for the early detection of AD and other dementias [26, 27]. EEG signals are highly variable depending on cognitive state, level of alertness, and sleep phases. In healthy individuals, delta activity (<4 Hz) corresponds to high-amplitude slow waves typical of deep sleep but may appear abnormally in awake adults with encephalopathy. Theta activity (4–8 Hz) is normal in children and during drowsiness, yet in awake adults it usually reflects subcortical or metabolic dysfunction. Alpha rhythms (8–12 Hz) dominate during relaxed wakefulness with closed eyes but are consistently reduced in AD and related conditions. Beta activity (13–30 Hz), linked to active thinking and attention, is typically diminished in cortical dysfunction and influenced by sedative drugs [28]. In AD [29], EEG recordings reveal a general slowing of brain rhythms, reduced complexity of electrical activity, and impaired synchronization across regions. Spectral analysis demonstrates increased delta and theta power alongside decreased posterior alpha and beta power compared

with healthy controls. Early alterations are often characterized by heightened theta activity and reduced beta power, followed later by alpha suppression [30]. More recent studies [31] confirm a marked increase of occipital delta activity and a significant decrease of parieto-occipital α_1 (8.5–10 Hz) and temporal α_2 (10–12 Hz) power. In contrast, patients with FTD show EEG profiles largely similar to controls, with the exception of a diffuse increase in theta power. The diagnostic potential of EEG has been strengthened by advanced analytical methods. Preprocessing includes filtering, artifact removal, either manual or automatic such as independent component analysis, and referencing. Linear approaches, such as averaging, Fourier analysis, and wavelet transforms, are often combined with nonlinear methods, including entropy measures, fractal dimension, detrended fluctuation analysis, Hjorth parameters, and the Hurst exponent [28].

Looking ahead, future research should emphasize the integration of linear and nonlinear analyses, the combination of EEG with other biomarkers, and the investigation of sleep-related EEG patterns, since sleep disturbances are among the earliest manifestations of AD and Parkinson's disease dementia (PDD) [32].

1.3.2 Clinical Data

Clinical assessments represent essential instruments for the screening and diagnosis of dementia. Among them, the Mini-Mental State Examination (MMSE) [33] and the Montreal Cognitive Assessment (MoCA) [34] are the most widely used. Both tests evaluate multiple cognitive domains, including orientation, memory, attention, language, and executive functions, yet differ in their sensitivity to specific stages of decline. Evidence shows that the MoCA demonstrates greater accuracy in detecting MCI and the early phases of AD, whereas the MMSE remains more reliable in assessing advanced dementia and in tracking disease progression, particularly in PDD [35]. For FTD, the INECO Frontal Screening (IFS) [36] was specifically designed to overcome the shortcomings of older tools. Unlike the Frontal Assessment Battery (FAB) [37], the IFS enhances discriminative ability between AD and FTD by focusing more precisely on frontal and executive functions [36]. Diagnostic accuracy relies on sensitivity and specificity across defined cut-off scores. However, these cut-offs must be adapted to educational background to avoid false positives or negatives [38]. Recent guidelines recommend weighted averages of sensitivity and specificity, Receiver Operating Characteristic (ROC) curve analysis, and area under the curve (AUC) as robust metrics for evaluating diagnostic accuracy [39]. Beyond screening, clinical scales play a crucial role in monitoring disease progression, supporting care planning, and complementing biomarker-based approaches.

Ongoing research advocates for the development of unified and standardized tools combining selected subitems from MMSE and MoCA scales, which could further improve diagnostic performance [40].

Cognitive function		MMSE (No. of points/trials)	MoCA (No. of points/trials)
Orientation		10 tasks (10 points)	6 tasks (6 points)
Memory	Learning	Learning of 3 words (3points/1 trial allowed)	learning of 5 words (no points/2 trials allowed)
	Delayed recall	3 words (3 points)	5 words (5points)
	Cued recall (optional)	not present	5 words (no points)
	Recognition (optional	not present	5 words (no points)
Naming		2 items (2 points)	3 pictures (3 points)
Visuospatial functions		copy of pentagons (1 point)	copy of cube (1 point) clock drawing (3 points)
Comprehension		3-stage command (3)	not present
Vigilance		not present	tapping with hand at letter A (1 point)
Language		repetition of sentence (1 point)	Repetition of 2 sentences (2 points)
Reading		Sentence (1 point)	not present
Abstract thinking		not present	similarities (2 points)
Writing		patient's sentence (1 point)	not present
Alternating Trial Making		not present	1 trial (1 points)

Figure 1.2: Comparison of MMSE and MoCA
Side-by-side overview of the **Mini-Mental State Examination (MMSE)** and the **Montreal Cognitive Assessment (MoCA)**, highlighting cognitive domains, number of items/trials, and scoring criteria.
Reproduced with permission from [35]

1.4 Learning Approaches for Dementia Classification

This section details the primary objective of the present work, which is to design and evaluate effective strategies for dementia classification through a rigorous and unbiased comparative analysis of diverse Machine Learning (ML) and Deep Learning (DL) models. We begin by establishing unimodal baselines, where cognitive and EEG data are analyzed independently to assess their predictive capacity. Building on these results, we investigate multimodal architectures, systematically comparing them with unimodal benchmarks. Particular attention is given to Middle fusion and Late fusion, which are critically assessed to determine their effectiveness in capturing the complex signatures of dementia.

1.4.1 Unimodal Learning

In the unimodal setting, each modality was processed and analyzed independently, providing a critical baseline for evaluating the additional benefits of multimodal integration.

For the Clinical Data (CD) modality, two models were employed: XGBoost and Multilayer Perceptron (MLP) [41].

For the Electroencephalogram (EEG) modality, we adopted ResNet-based architectures [42, 43]. This model choice reflects the suitability of convolutional deep learning methods for handling high-dimensional time-series data.

Collectively, these unimodal experiments establish the reference performance against which the multimodal approaches are evaluated.

1.4.2 Multimodal Learning

In the multimodal scenario, the two modalities, CD and EEG, were integrated in order to exploit complementary sources of information and better capture the heterogeneous nature of dementia-related patterns. Two categories of fusion strategies were investigated: late fusion and middle fusion.

Late fusion was explored in two ways. In the Late Separate Average Fusion scheme, the predictions generated by unimodal models were averaged to produce a combined decision. In the Late Meta Fusion scheme, the unimodal predictions were instead used as inputs for a meta-classifier.

Middle fusion, by contrast, was implemented by merging intermediate feature representations from each modality, allowing the model to learn joint latent embeddings and capture inter-modality dependencies before the final classification step.

These multimodal strategies were systematically compared both with each other and with the unimodal baselines, providing a comprehensive evaluation of the relative advantages of data integration in dementia classification tasks.

2 Methods

2.1 Dataset and Methodological Framework

2.1.1 Dataset Description

The data employed in this study originate from a publicly available dataset hosted on Synapse under the name The Latin American Brain Health Institute (BrainLat)[44]. * This dataset represents a multicenter Latin American initiative aimed at collecting high-density resting-state electroencephalography (EEG) and comprehensive clinical assessments from patients with neurodegenerative disorders as well as healthy controls. EEG recordings were acquired using a 128-channel BioSemi ActiveTwo system, ensuring high temporal and spatial resolution. In addition to demographic variables such as sex, age, and years of education, the dataset incorporates clinical scores obtained from standardized neurocognitive assessments. Specifically, three cognitive instruments were included: the Montreal Cognitive Assessment (MoCA), the INECO Frontal Screening (IFS), and the Mini-Social Cognition and Emotional Assessment (mini-SEA). The MoCA evaluates six core cognitive subdomains: visuospatial abilities, attention, language, abstraction, memory, and orientation. The IFS is designed to assess executive functioning across eight tasks, including motor series, conflicting instructions, motor inhibition, digit span, month sequencing, visual working memory, proverbs, and verbal inhibition. The mini-SEA targets socio-emotional processes and is composed of two measures: the Facial Emotion Recognition Test (FERT) and the Faux-Pas Test, which evaluates Theory of Mind (ToM). The laterality feature was discarded because it showed no discriminative relevance. The current release comprises data from 35 patients with Alzheimer's disease (AD), 19 patients with frontotemporal dementia (FTD), and 32 cognitively normal controls (CN).

*Data availability statement `https://www.synapse.org/Synapse:syn51549340/wiki/624187`

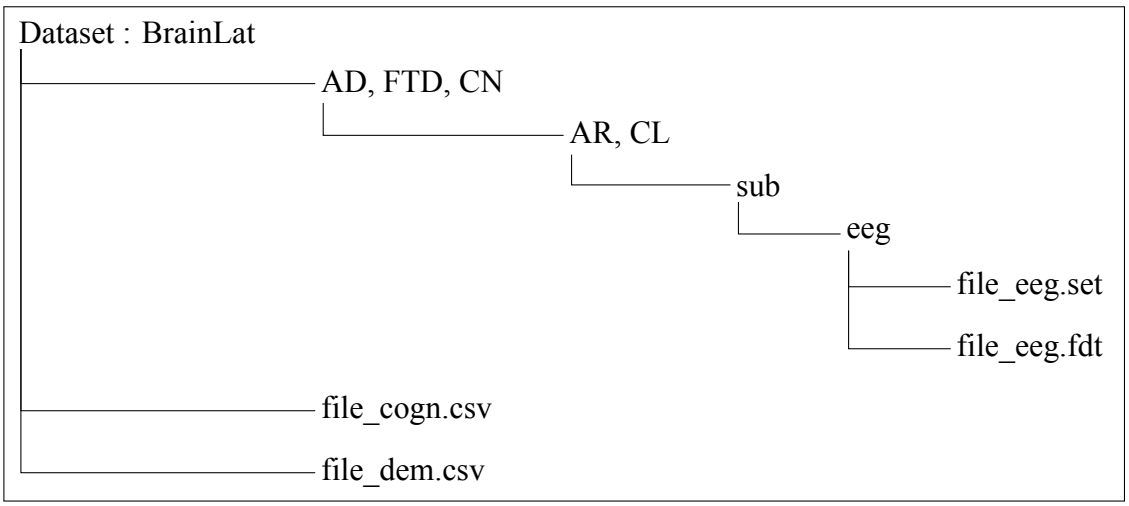

Figure 2.1: BrainLat dataset directory tree
The**BrainLat dataset** is structured into three diagnostic groups **Alzheimer's Disease (AD), Frontotemporal Dementia (FTD), Cognitively Normal controls (CN)**. Within each group, data are organized by acquisition site **(AR, CL)** and subject folders **(sub)**. EEG data are stored in **.set** and **.fdt** files, while complementary cognitive and demographic information is provided in file**_cogn.csv** and file**_dem.csv**.

2.1.2 Data-Level and Model-Level Methods

The methodological design of this study was organized into several key stages.

At the data level, the workflow focused on model input processing. The procedure began with data selection, followed by a series of steps designed to transform raw EEG signals and clinical information into suitable inputs for for subsequent modeling. This stage encompassed preprocessing as well as feature extraction.

At the model level, the workflow was oriented toward the model development and optimization of effective learning approaches. This involved defining the network architecture and incorporating EEG-specific design choices tailored to the characteristics of the signals. The models were subsequently trained with carefully designed training strategies, complemented by regularization and optimization techniques to enhance generalization and efficiency. To further refine performance, systematic parameter tuning and hyperparameter search were performed.

2.1.3 Sofware and Hardware Configuration

The experimental workflow was implemented in PyTorch framework and executed on an NVIDIA® A30 GPU device.

2.2 Model Input Processing

2.2.1 Data Selection

Data selection was performed to retain subjects with complete data across modalities.

As outlined in subsection 2.1.1 , the initial dataset consisted of 35 patients diagnosed with Alzheimer's Disease (AD), 19 with Frontotemporal Dementia (FTD), and 32 Cognitively Normal controls (CN). For the clinical modality, subjects presenting missing clinical features were excluded from the analysis. Similarly, for the EEG modality, participants lacking the corresponding raw signal files were removed.

After applying these exclusion criteria, the final study cohort consisted of 58 participants, including 28 AD patients, 17 CN controls, and 13 FTD patients. This refined dataset was retained for all subsequent analyses. For clarity, Figure 2.2 reports both the absolute counts and the relative distribution of subjects across diagnostic groups.

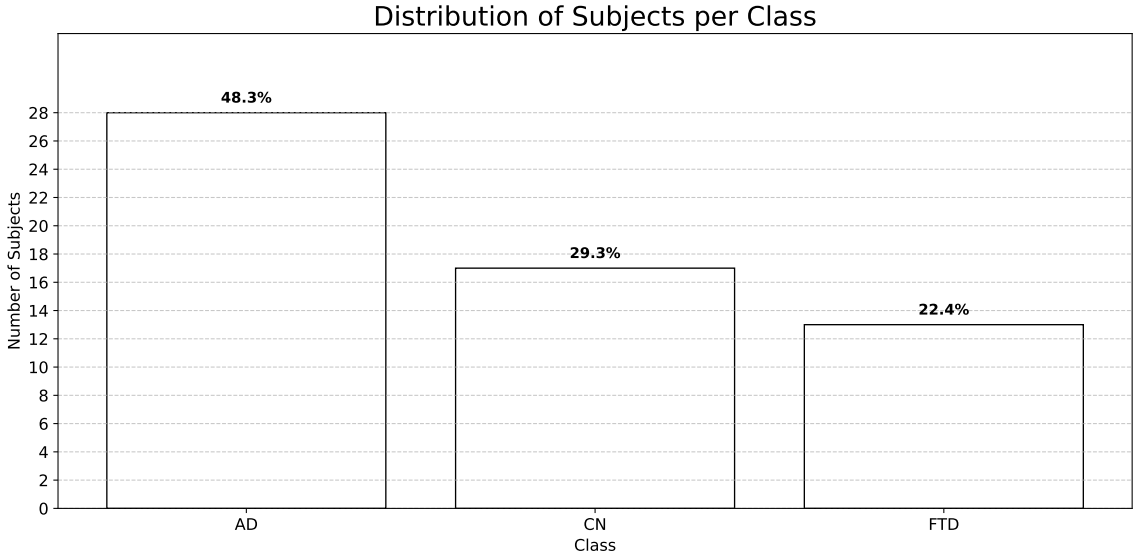

Figure 2.2: Class distribution of the analysis dataset
The dataset includes 28 subjects with Alzheimer's Disease (AD) , 17 cognitively normal controls (CN), and 13 patients with Frontotemporal Dementia (FTD). The bar plot shows the **absolute counts** with corresponding **percentages** for each class.

2.2.2 Data Preprocessing

Data preprocessing and feature extraction were performed to transform data into a learnable format. For the clinical data, normalization was applied to standardize the input features. The resulting representation was a one-dimensional vector of shape (20,1), where 20 corresponds to the total number of selected features. Among the available approaches, min–max scaling was adopted, as formalized in Equation (2.1). This method is widely used in data preprocessing due to its simplicity, interpretability, and effectiveness. By rescaling feature values into a fixed range, typically [0,1], min–max normalization ensures comparability across variables.

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{2.1}$$

where:

- x = original value of the feature,
- x' = normalized value,
- $x_{\min}$ = minimum value of the feature,
- $x_{\max}$ = maximum value of the feature.

This method rescales the feature values into the range $[0, 1]$.

2.2.3 Feature Extraction

For the EEG modality, a Custom feature extraction step was performed, designed to capture time–frequency characteristics of the signal while simultaneously preserving its spatial information. The final EEG data structure was a 3D tensor with shape (11, 89, 83). The Rationale behind design choises for ROI-Based Spectrogram Representation can be summarised as follows, while more through descriptions of each point are presented below the list.

- Compromise strategy via ROI mapping: Using all 128 channels preserves the highest spatial resolution but comes at the cost of substantial computational demands. In contrast, single-channel spectrograms, obtained by averaging across channels, drastically reduce computational load but lose the spatial variability intrinsic to EEG. As a compromise, the 128 electrodes were mapped into a reduced set of 11 regions of interest (ROIs), following a structured anatomical criterion that balances spatial resolution with computational efficiency
- ROI-based spectrogram construction: Each spectrogram was generated with $C = 11$ channels, where each channel corresponds to one ROI. The resulting spectrograms are represented as (C, F, T), with $C = 11$ ROIs, F frequency bins, and T time steps.
- Managing variable sequence lengths: A challenge in preprocessing arose from inter-subject variability in the number of epochs, producing spectrograms of unequal temporal length. To standardize dimensions while preserving spectral information, zero-padding was applied along the temporal axis.
- Group-level inspection: To qualitatively evaluate the feature extraction pipeline, representative spectrograms were visualized for one subject per diagnostic group (AD, FTD, CN).

Compromise strategy via ROI mapping

The EEG setup used in this study consisted of 128 electrodes, as illustrated in the montage image provided Figure 2.3. When generating spectrogram representations, two extreme strategies can be considered:

1. Full-channel spectrograms (128 channels): This approach preserves the maximum spatial resolution, as information from all electrodes is retained. However, the computational burden is extremely high, making it less practical for large-scale analyses.
2. Single-channel spectrograms (median across channels): In this case, the spatial dimension is collapsed by averaging signals across electrodes. This drastically reduces the computational cost but also eliminates the spatial variability of brain activity, which is often critical for distinguishing between diagnostic groups.

To overcome these limitations, a compromise strategy was adopted. The mapping criterion was grounded in the assumption that some principal differences between diagnostic groups (AD,

FTD, CN) are associated with the specific brain regions affected, as outlined in Section 1.2. Accordingly, EEG channels were grouped by both hemisphere (Left, Medial, Right) and lobe (Frontal, Central, Temporal, Parietal, Occipital), following the mapping kindly suggested by Dr. Antonio and provided in the Appendix A. To facilitate interpretation, the mapping is summarized in Table 2.1 and visually represented through an Interactive 3D Topomap. This topomap highlights the spatial organization of electrodes across the scalp, providing an intuitive link between the defined ROIs and their underlying anatomical distribution. Figure 2.4 shows some screenshots of the interactive 3D topomap, while the fully interactive version is available in the Appendix A).

In practice, the 128 electrodes were aggregated into 11 anatomically defined ROIs, encompassing the frontal, parietal, temporal, and occipital lobes, further subdivided into left, right, and medial regions. This mapping preserved clinically meaningful spatial information while substantially reducing computational demands, thus providing a balanced solution that retains biological plausibility without incurring the prohibitive costs of full-channel analysis.

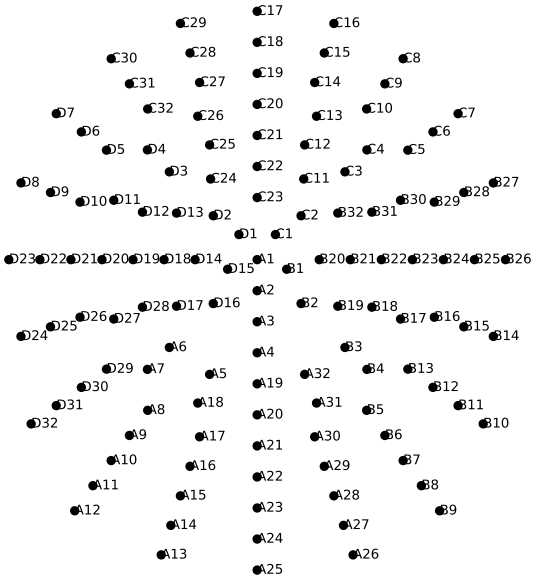

Figure 2.3: Channels montage
Electrode montage of the **128 EEG channels** used in the BrainLat dataset. Electrodes are distributed across frontal, central, parietal, temporal, and occipital regions, providing comprehensive scalp coverage for signal acquisition.

ROI (11)	Channels (128)
Frontal Medial	C17, C18, C19, C20, C21, C22, C23, A1, A2
Frontal Right	C16, C15, C14, C13, C12, C11, C2, C1, B1, C8, C9, C10, C4, C3, C7, C6, C5, B32, B31, B30, B29, B20, B21, B22, B23, B24
Frontal Left	C29, C28, C27, C26, C25, C24, C30, C31, C32, D7, D6, D5, D4, D3, D2, D1, D10, D11, D12, D13, D14, D15, D21, D20, D19, D18
Parietal Medial	A3, A4, A19, A20, A21, A22
Parietal Right	B2, A32, A31, A30, A29, B3, B4, B5, B6, B19, B18, B17, B16, B13, B12
Parietal Left	D26, D27, D28, D17, D16, D30, D29, A9, A8, A7, A6, A16, A17, A18, A5
Temporal Right	B28, B27, B25, B26, B15, B14, B11, B10
Temporal Left	D8, D9, D23, D22, D24, D25, D32, D31
Occipital Medial	A23, A24, A25
Occipital Right	A28, A27, A26, B7, B8, B9
Occipital Left	A12, A11, A10, A13, A14, A15

Table 2.1: Mapping Channels-ROI
Channels are grouped according to anatomical **lobes (Frontal, Parietal, Temporal, Occipital) and hemispheres (Medial, Left, Right)**.

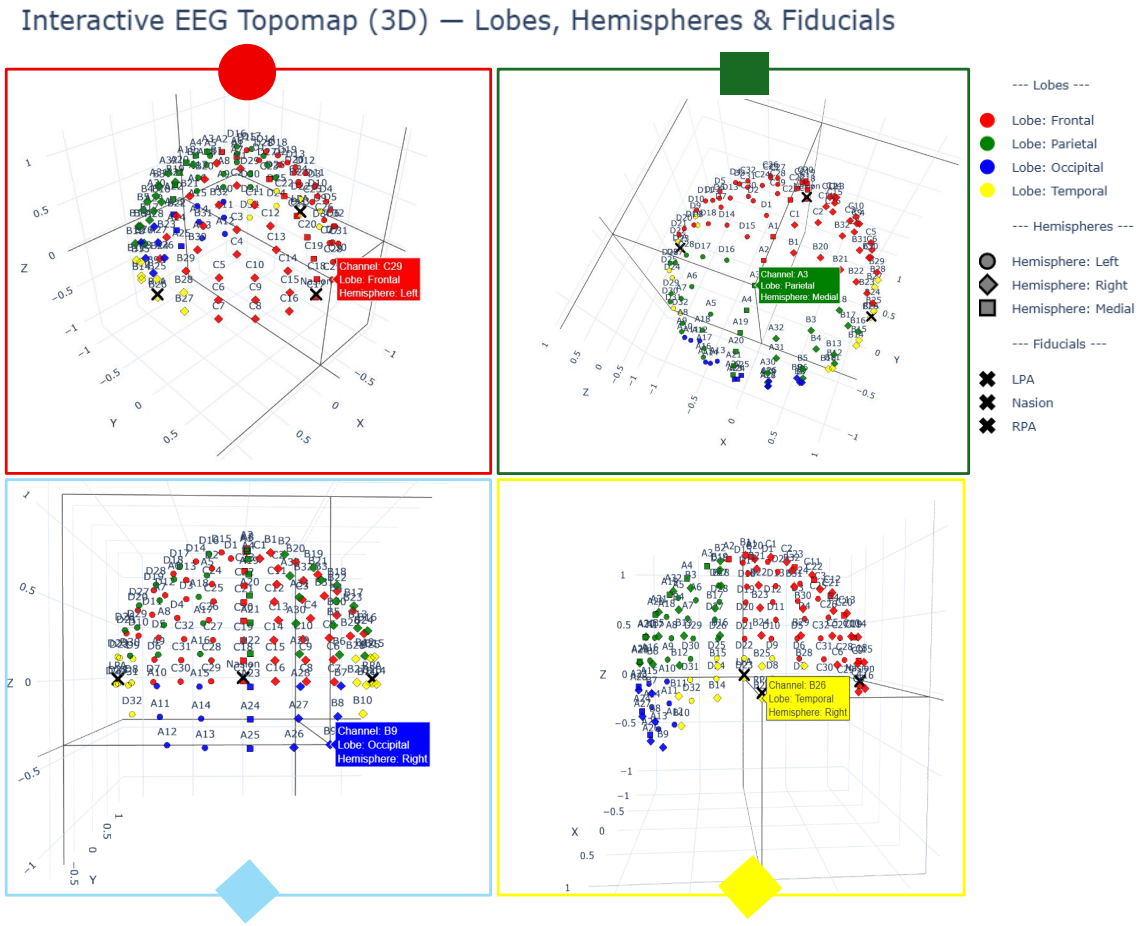

Figure 2.4: Interactive EEG Topomap (3D) - Lobe, Hemispheres, Fiducials
Screenshots of the interactive EEG topomap, where electrodes are color-coded according to **Lobes** (red = Frontal, green = Parietal, blue = Occipital, yellow = Temporal) and shaped according to **Hemispheres** (circle = Left, square = Right, diamond = Medial). Black crosses indicate anatomical **Fiducials** (LPA = left preauricular, RPA = right preauricular, and nasion). The four panels provide different 3D perspectives of the same montage, highlighting the ROI mapping from the original 128 electrodes into lobe–hemisphere groupings. The interactive version of this visualization is available in the Supplementary Material A.

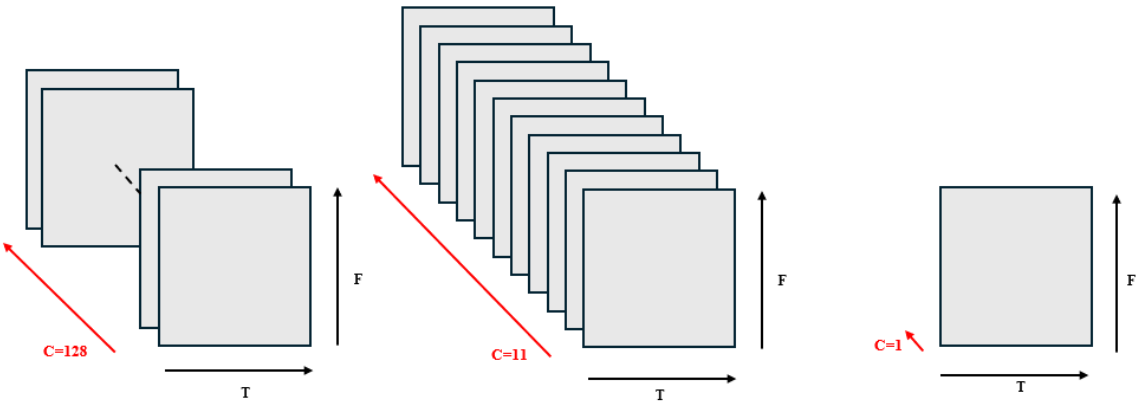

Figure 2.5: Comparison of spectrogram strategies based on the number of channels. From left to right: (1) **Full-channel strategy** ($C = 128$), which preserves the full spatial resolution of the original EEG montage but results in a very high computational load; (2) **ROI-based strategy** ($C = 11$), where electrodes are mapped into 11 regions of interest (ROIs), providing a compromise between spatial resolution and computational efficiency; (3) **Single-channel strategy** ($C = 1$), based on the median spectrogram across all channels, which drastically reduces computational cost but at the expense of losing spatial information. In all cases, the spectrograms retain the same temporal dimension (T) and frequency resolution (F).

ROI-based spectrogram construction

In this study, raw EEG signals were segmented into epochs and subsequently transformed into spectrograms by estimating the power spectral density (PSD) with the `P-Welch` method. First, the 128 EEG channels were averaged into 11 predefined regions of interest (ROIs). For each recording of length L, the signal was partitioned into windows of $W = 16$ s with an overlap of $o = 50\%$. The total number of windows extracted from a recording was defined as :

$$N_{\mathrm{w}} = \left\lfloor \frac{L - W}{W(1 - o)} \right\rfloor + 1. \tag{2.2}$$

The PSD of each window was then estimated using the P-Welch method, with a Hanning window, 50% overlap, and a frequency resolution of $\Delta f = 0.5$ Hz (NFFT $= 2f_s = 1024$). Since the data were band-pass filtered between 1 Hz and 45 Hz, each PSD vector had length $N_{\mathrm{PSD}} = 89$. Consequently, the spectrogram for each subject was represented as a tensor with dimensions $(\mathrm{ROIs}, N_{\mathrm{PSD}}, N_{\mathrm{w}})$.

Managing variable sequence length

A major challenge in EEG preprocessing was the variability in the number of epochs across subjects, which led to differences in temporal dimensions of the spectrograms. To standardize input sizes while preserving spectral information, zero-padding was applied to match the longest

sequence, corresponding to $N_w = 83$ (see Figure 2.6). The final EEG input tensor, denoted X_{EEG}, thus had dimensions $(11, 89, 83)$ (see Figure 2.7).

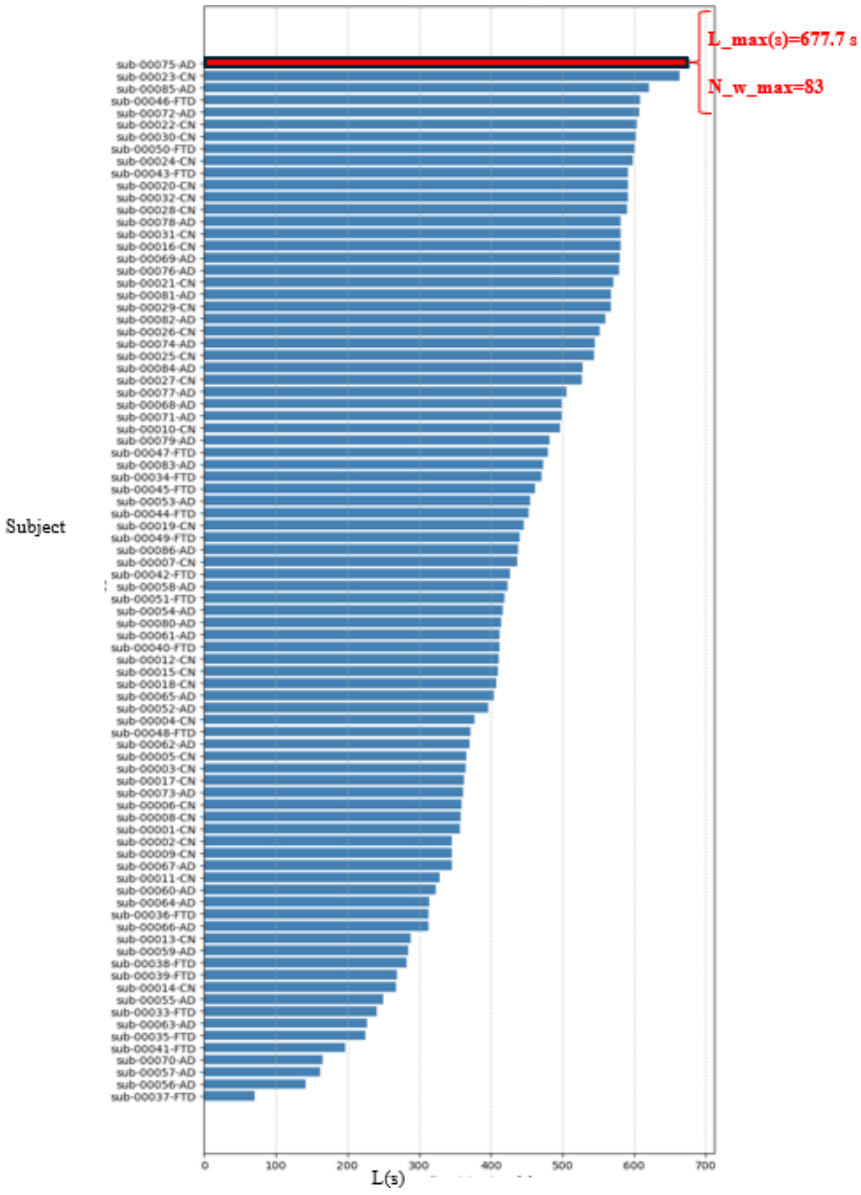

Figure 2.6: Lenghts across subjects

The bar plot shows the signal lengths across subjects. The maximum length was found to be **L_max=677.7 s** , then **N_w_max=83**, which was selected as the reference length for zero-padding all signals to ensure uniform input dimensions.

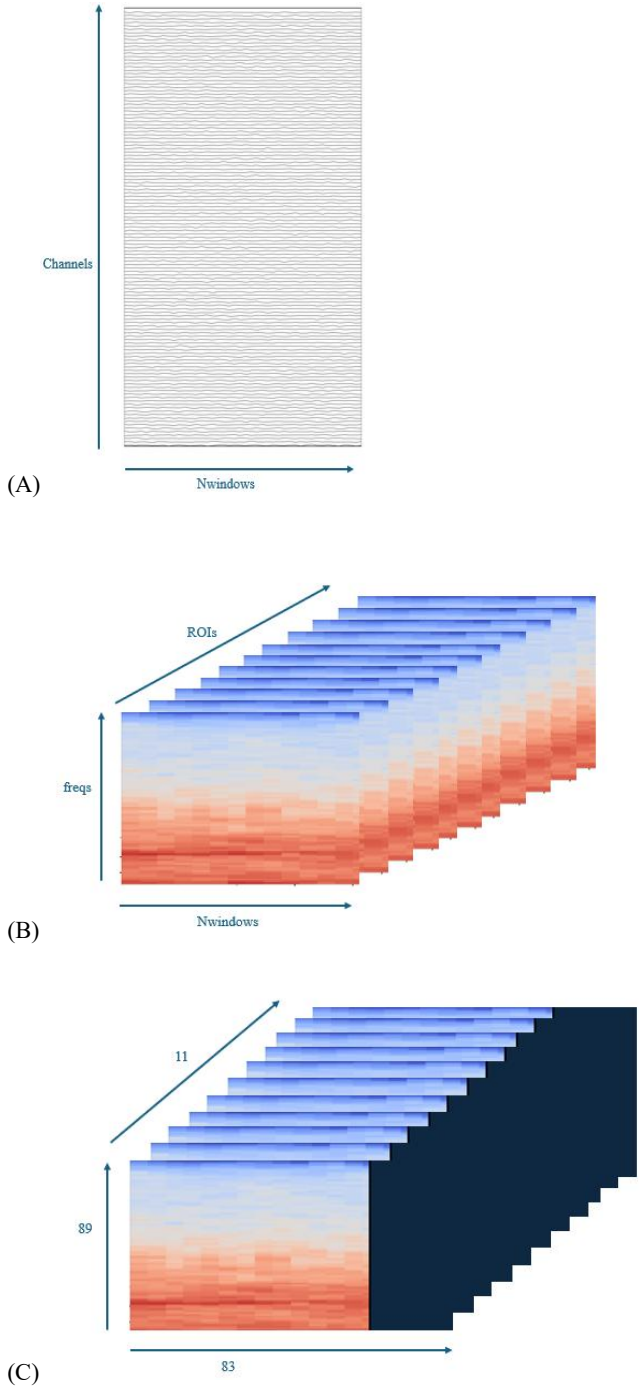

Figure 2.7: ROI-based spectrogram with zero-padding
Panel (A) shows the raw EEG signal across multiple channels and time. Panel (B) illustrates the spectrogram obtained after grouping the 128 channels into 11 regions of interest (ROIs), computing the power spectral density (PSD) for each window, and stacking them along the temporal dimension. Panel (C) presents the final spectrogram tensor with dimensions $(11, 89, 83)$ after applying zero-padding to standardize temporal length across subjects. In this example, the subject under consideration did not have the maximum recording length, hence the padded dark region.

Group-level inspection

In this paragraph, we present illustrative examples of EEG spectral features for one representative subject from each diagnostic group: Alzheimer's disease (AD), frontotemporal dementia (FTD), and cognitively normal controls (CN). For each subject, we first show two-dimensional spectrograms across the 11 predefined ROIs, which provide a time–frequency representation of brain activity. We then summarize the same information in terms of mean band power distributions, visualized through boxplots across the canonical EEG frequency bands (delta, theta, alpha, beta, gamma). This dual representation allows us to both appreciate the temporal dynamics of spectral content (via spectrograms) and capture aggregated statistical properties of each frequency band (via mean band power), thereby highlighting distinctive spectral signatures across diagnostic groups.

In AD, the spectrograms reveal a generalized slowing of brain activity, characterized by elevated delta and theta power alongside reduced alpha, beta, and gamma activity. These changes are most pronounced in posterior regions, reflecting the classical EEG slowing associated with AD. In FTD, slowing appears more focal, with increased delta and theta power predominantly in frontal and temporal regions, consistent with the disease's anterior cortical involvement. In contrast, the CN subject displays preserved EEG organization, with a dominant posterior alpha rhythm, low power in slow frequencies (delta and theta), and stable beta and gamma activity—patterns aligned with normal brain function.

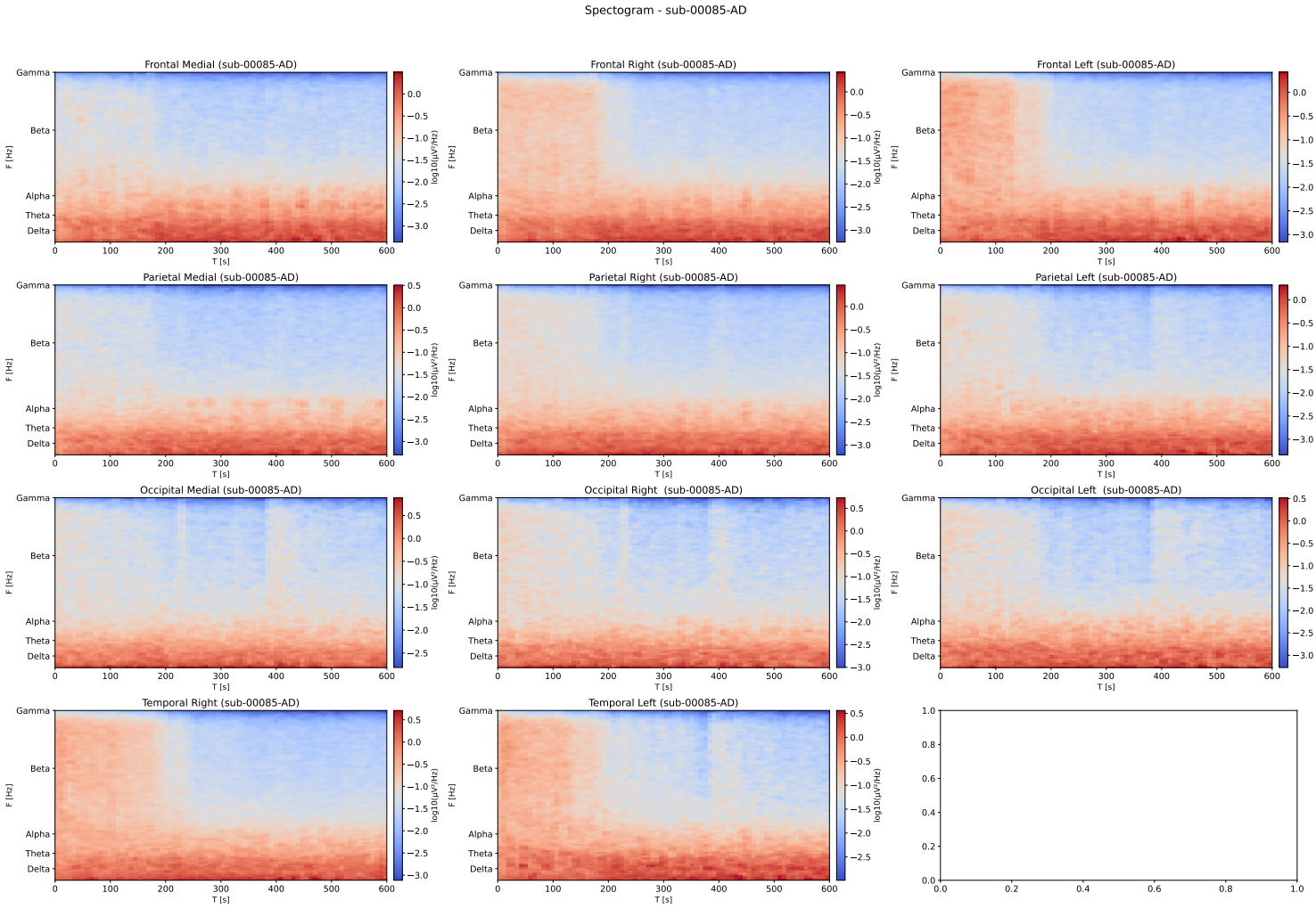

Figure 2.8: Spectrogram of a AD subject

Two-dimensional spectrogram representations for each of the 11 ROIs (frontal, parietal, occipital, and temporal lobes, subdivided by hemisphere and medial regions). Each panel shows the log-scaled power spectral density ($\log_{10}(\mu V^2/Hz)$) across frequency bands (delta, theta, alpha, beta, gamma)as a function of time. The figure shows a generalized slowing of brain activity, with increased delta and theta power and reduced alpha, beta, and gamma activity. These alterations are most evident in posterior regions, reflecting the characteristic EEG slowing typical of AD.

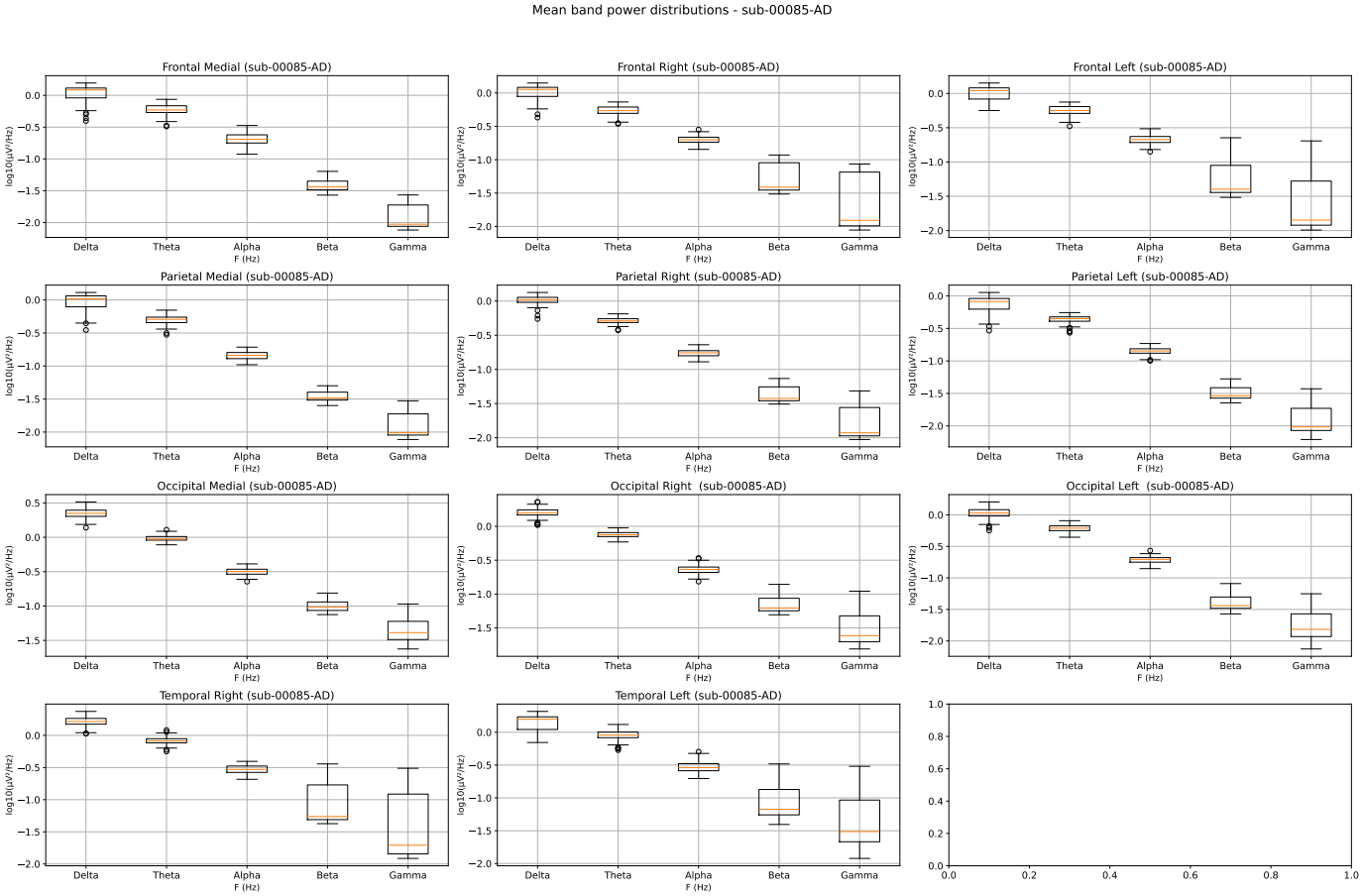

Figure 2.9: Mean band power distributions of a AD subject

Boxplots of log-scaled power spectral density ($\log_{10}(\mu V^2/Hz)$) across the five canonical EEG frequency bands (delta, theta, alpha, beta, gamma) for each of the 11 ROIs. Each panel corresponds to a different ROI, highlighting the variability of spectral power within and across brain regions.

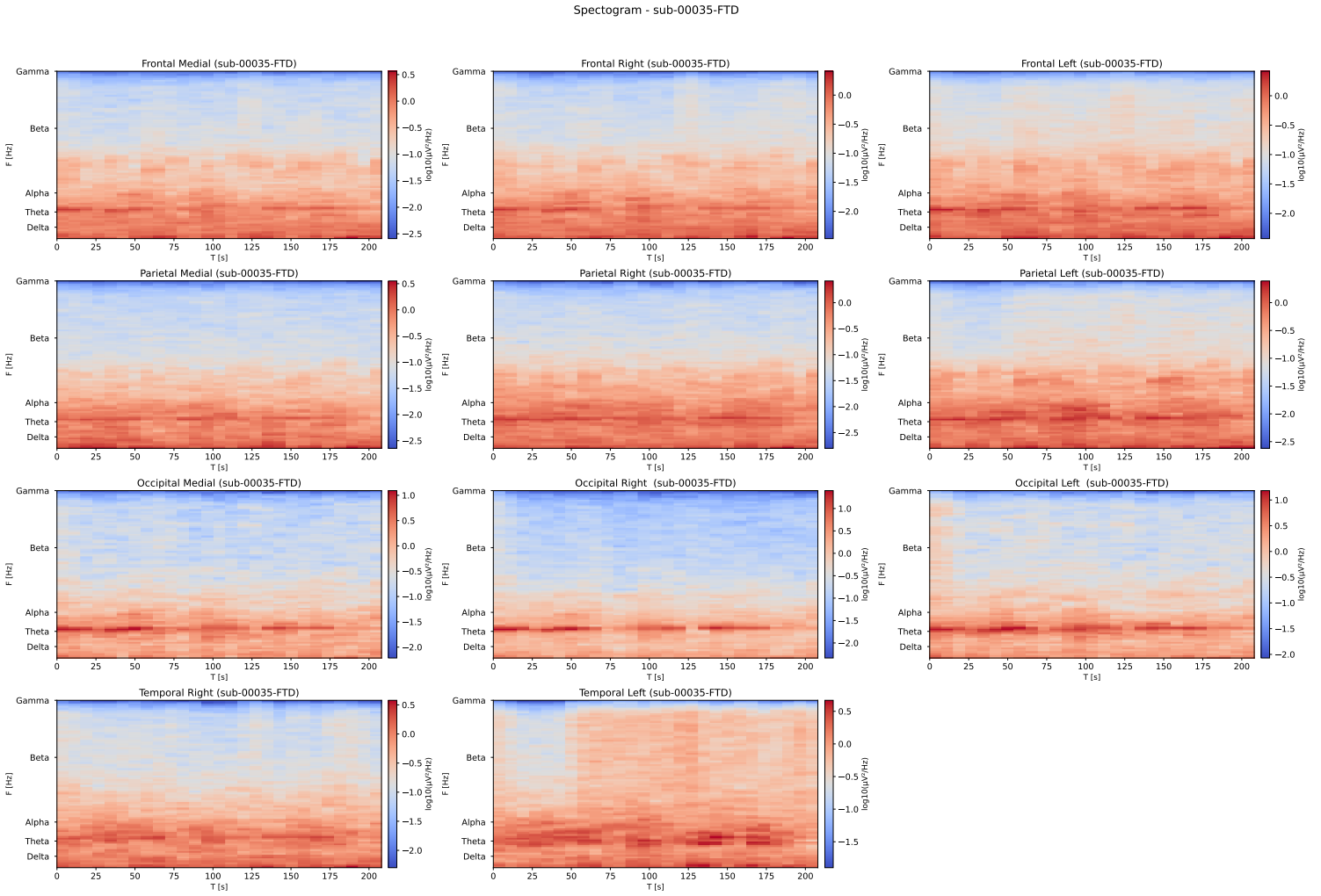

Figure 2.10: Spectrogram of a FTD subject

Two-dimensional spectrogram representations for each of the 11 ROIs (frontal, parietal, occipital, and temporal lobes, subdivided by hemisphere and medial regions). Each panel shows the log-scaled power spectral density ($\log_{10}(\mu V^2/Hz)$) across frequency bands (delta, theta, alpha, beta, gamma) as a function of time. The figure shows focal slowing predominantly localized to frontal and temporal regions, marked by increased delta and theta in anterior areas.

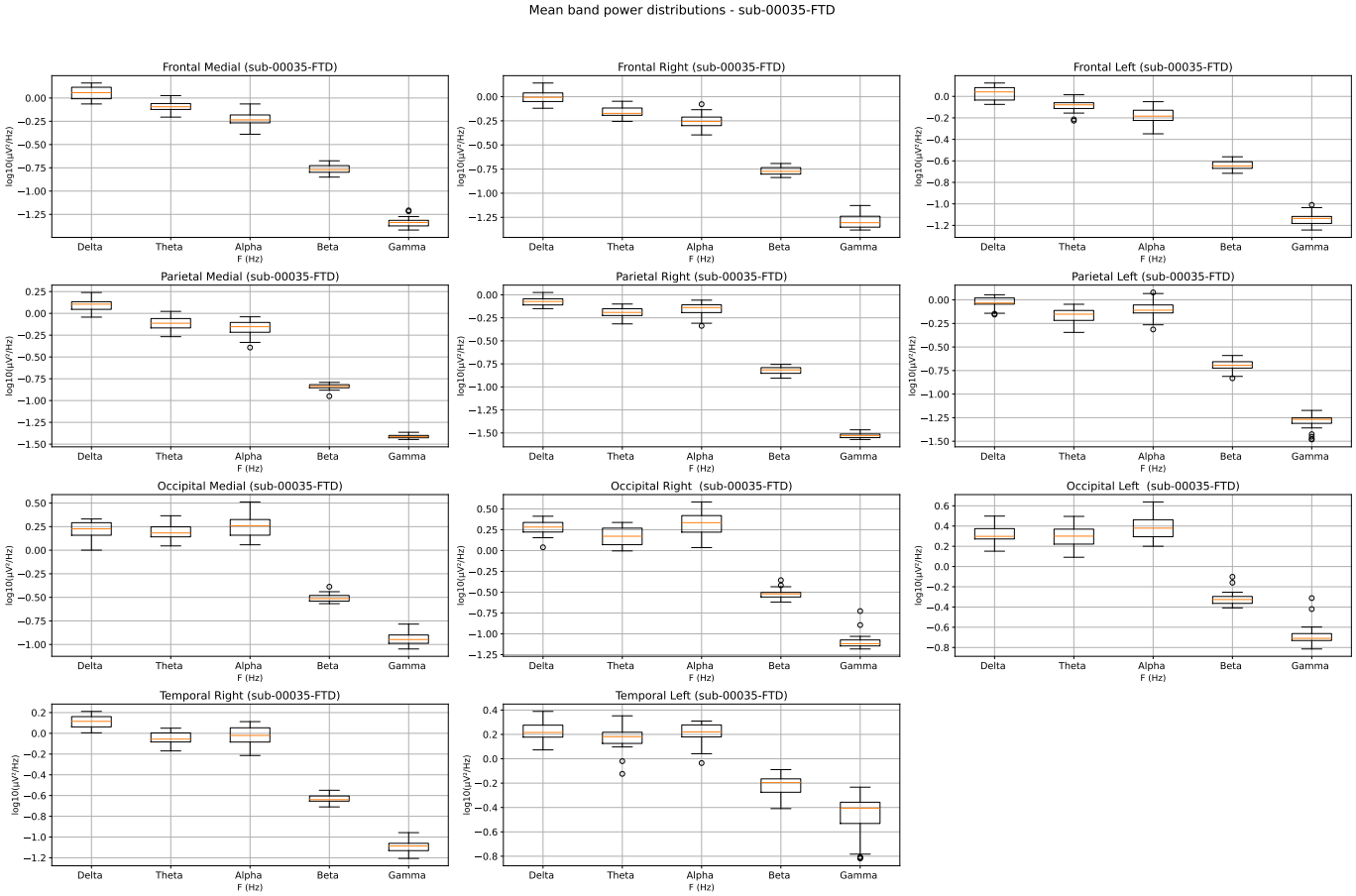

Figure 2.11: Mean band power distributions of a FTD subject

Boxplots of log-scaled power spectral density ($\log_{10}(\mu V^2/Hz)$) across the five canonical EEG frequency bands (delta, theta, alpha, beta, gamma) for each of the 11 ROIs. Each panel corresponds to a different ROI, highlighting the variability of spectral power within and across brain regions.

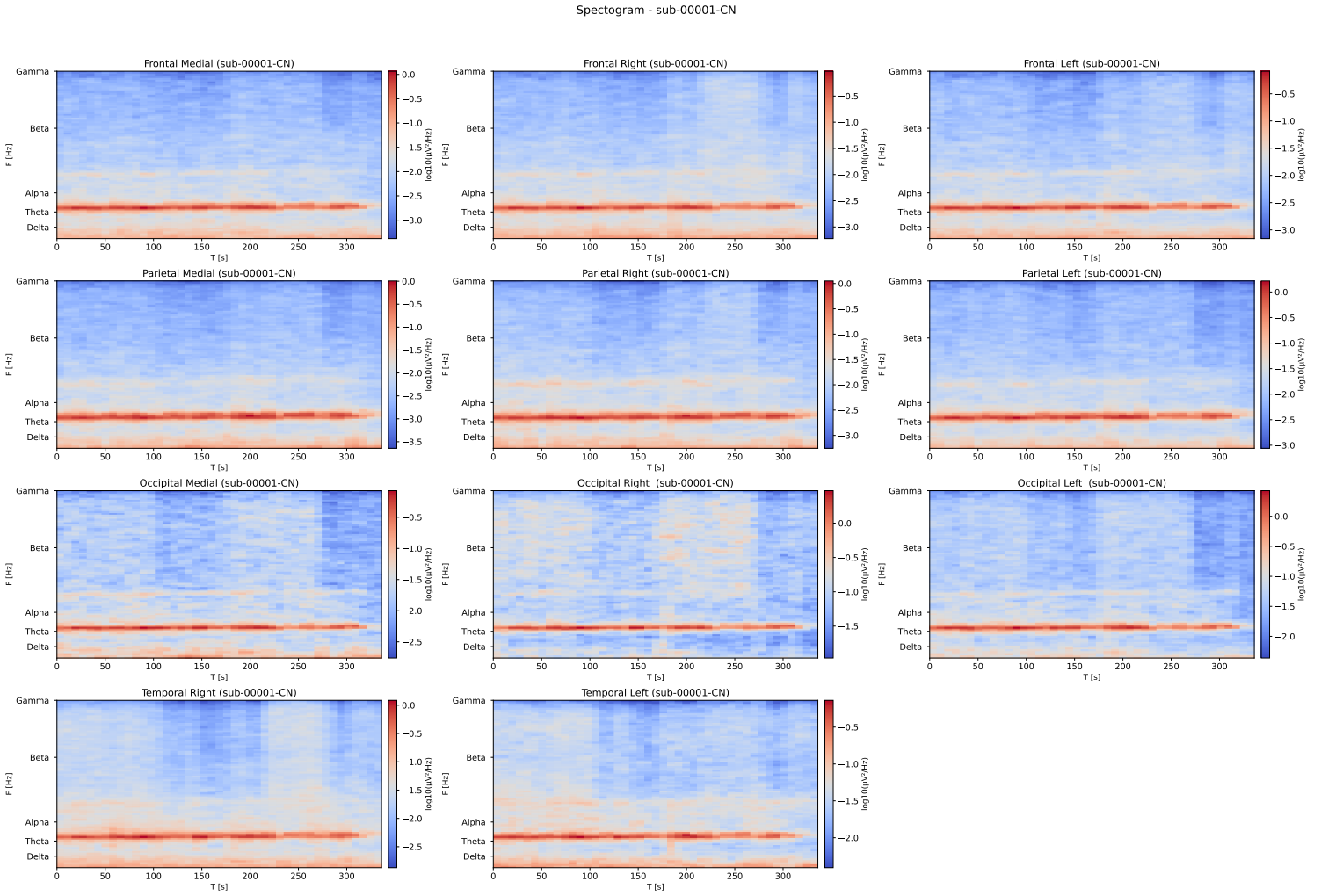

Figure 2.12: Spectrogram of a CN subject

Two-dimensional spectrogram representations for each of the 11 ROIs (frontal, parietal, occipital, and temporal lobes, subdivided by hemisphere and medial regions). Each panel shows the log-scaled power spectral density ($\log_{10}(\mu V^2/Hz)$) across frequency bands (delta, theta, alpha, beta, gamma)as a function of time. The figure shows preserved EEG organization with a dominant posterior alpha rhythm, low power in delta and theta bands, and relatively stable beta and gamma activity, consistent with normal brain dynamics.

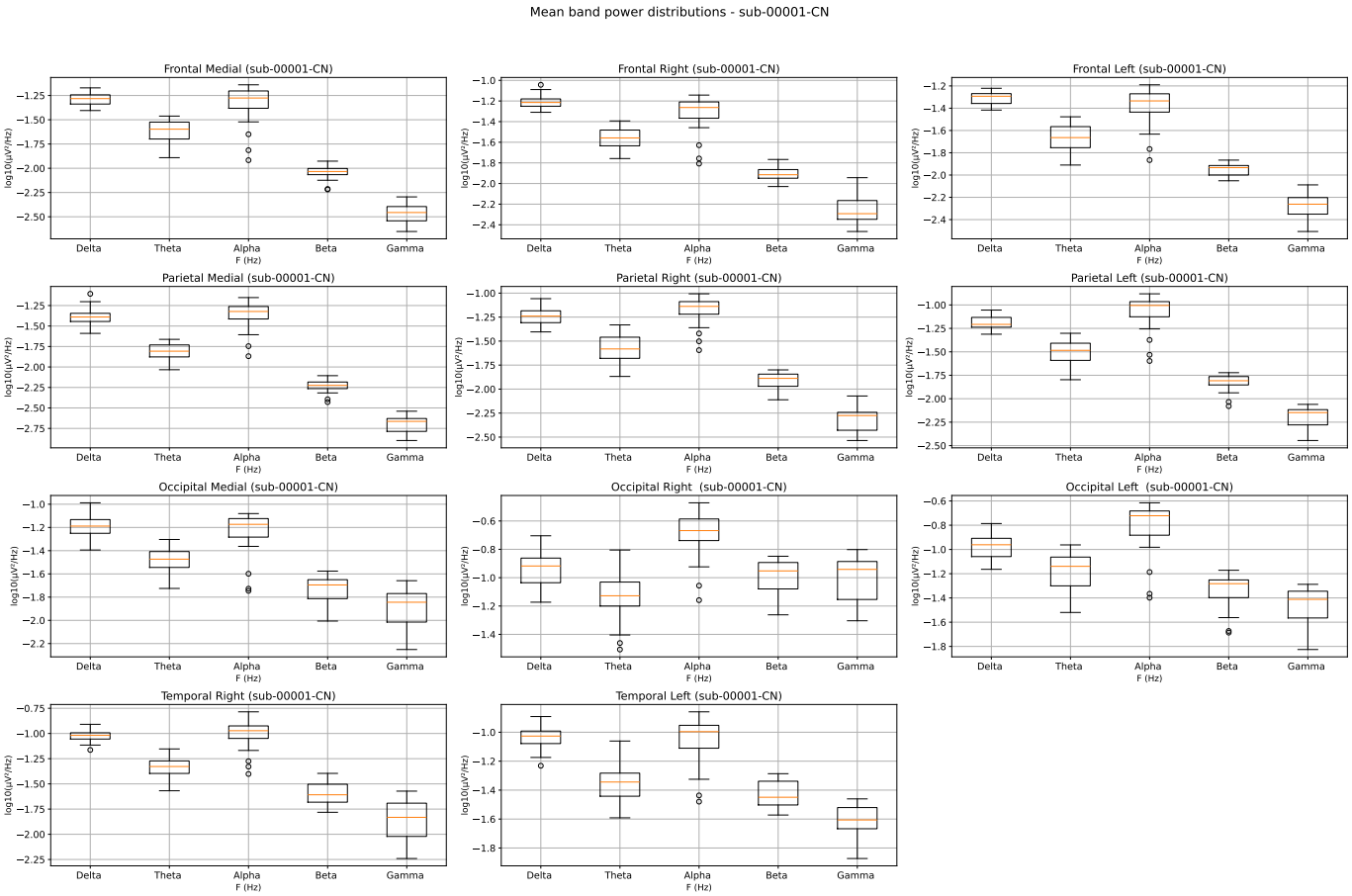

Figure 2.13: Mean band power distributions of a CN subject

Boxplots of log-scaled power spectral density ($\log_{10}(\mu V^2/Hz)$) across the five canonical EEG frequency bands (delta, theta, alpha, beta, gamma) for each of the 11 ROIs. Each panel corresponds to a different ROI, highlighting the variability of spectral power within and across brain regions.

2.3 Model Development and Optimization

In this section, we present the methodological framework adopted for each learning approach. The general strategies, such as design choices, training protocols, and optimization procedures, are described here to provide a common foundation. In the subsequent sections dedicated to each specific model, the corresponding architectural insights are outlined in greater detail.

2.3.1 Architecture

Established machine learning frameworks were combined with the design of customized architectures, specifically tailored to capture the distinctive characteristics and complexity of the dataset. On one side, widely adopted methods such as XGBoost were employed as robust baselines, offering interpretability and competitive performance without the need for extensive architectural tuning. On the other hand, custom deep learning architectures like ResNet variants were developed to more effectively capture the underlying structure and complexity of the data. This hybrid approach, balancing established methods with domain-specific architectural design, allowed us to both benchmark against state-of-the-art models and explore novel solutions optimized for the clinical problem at hand.The class definitions used for the models are provided in the AppendixA.

2.3.2 Training strategy

The adopted training strategy was designed to ensure both robust model development and fair evaluation of performance. Its core lies in the definition of a systematic data partitioning scheme that guarantees reproducibility while preserving class balance across all subsets.

Data partitioning was carried out using a stratified repeated hold-out strategy, ensuring that the relative proportions of Alzheimer's disease (AD), frontotemporal dementia (FTD), and control (CN) subjects were maintained in each split. The dataset was divided a priori into three disjoint subsets: training, validation, and test. The test set size was fixed at $p_{\text{test}} = 20\%$, a compromise chosen to balance two competing requirements: on the one hand, having a sufficiently large test set to yield statistically reliable performance estimates; on the other hand, reserving enough data for model training to capture the underlying patterns effectively. From the remaining portion $(1 - p_{\text{test}})$, a validation fraction of $p_{\text{val}} = 20\%$ was allocated, while the rest was used for training. Reproducibility was ensured by fixing the random seed to 123.

Within this framework, the training and validation phases were dedicated to parameter optimization and overfitting control, while the test phase served as an independent benchmark to assess generalization (see Results Chapter). To strengthen reliability, the entire process was repeated for $N = 100$ independent random splits, ensuring that each subject appeared in the test set at least once. This procedure provided a comprehensive evaluation of model performance, independent of any single partition.

Subset	Total	AD	FTD	CN
Training	37 (63.4%)	18	8	11
Validation	9 (16.6%)	4	2	3
Test	12 (20%)	6	3	3
Overall	58 (100%)	28 (48.3%)	13 (22.4%)	17 (29.3%)

Table 2.2: Data partitioning with stratified repeated hold-out strategy
Data partitioning was carried out using a **stratified repeated hold-out strategy** with $p_{\text{Test}} = 20\%$, $p_{\text{Val}} = (1 - p_{\text{Test}}) \times 20\%$. The class distribution is reported in Subsection2.2.1and includes a total of 58 participants, 28 AD patients (48.3%), 17 CN controls (29.3%), and 13 FTD patients (22.4%).

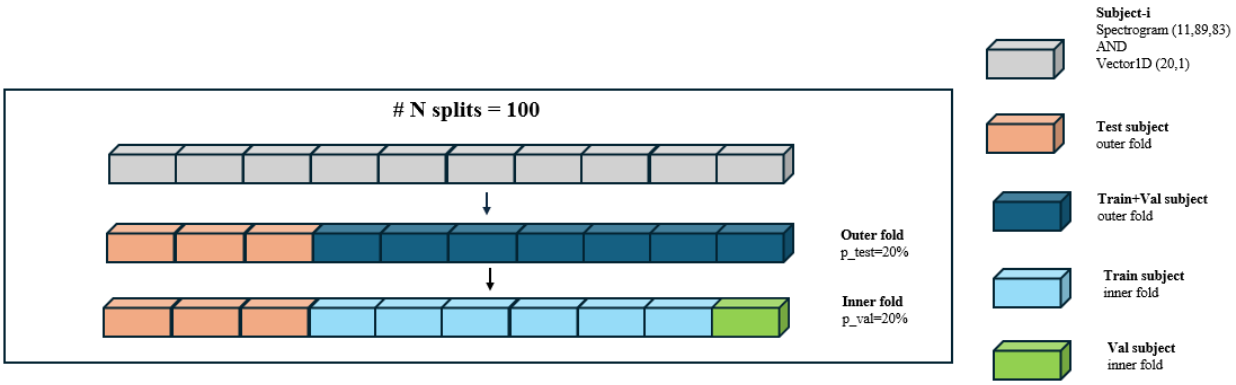

Figure 2.14: Training strategy.
The procedure was repeated for $N = 100$ **independent random splits**. Data partitioning followed a stratified repeated hold-out strategy with fixed p_{test} and p_{val}. The grey block represents a subject with its spectrogram and clinical vector. In the outer fold, orange blocks indicate test samples, while blue blocks indicate the training+validation subset. Within the inner fold, light-blue blocks correspond to the training subset, and green blocks to the validation subset.

2.3.3 Regularization and Optimization

Training of the neural networks was carried out under a carefully designed optimization and regularization framework. Mini-batches of size 37 were used during training. Optimization was performed using the Adam algorithm, initialized with a learning rate of 0.001. To further prevent overfitting, an L2 weight decay penalty of 1×10^{-5} was applied. The loss function was defined as cross-entropy, augmented with class-specific weights to account for the imbalance between diagnostic categories. To adaptively refine the learning process, a ReduceLROnPlateau scheduler was employed. Whenever validation performance stopped improving, the learning rate was reduced by a factor of 0.1, with a patience of 5 epochs before triggering. Finally, early stopping was incorporated. Training was allowed to proceed for up to 250 epochs, but would terminate earlier if the validation loss failed to improve by at least 1×10^{-4} over 15 consecutive epochs. Together, these strategies formed a robust protocol, combining adaptive optimization, regularization, and dynamic learning rate control to ensure both stability during training and improved generalization on the test data.

2.3.4 Parameter tuning and Hyperparameter search

Hyperparameter tuning often dominates training time, as many models are highly sensitive to optimizer settings and architectural choices. To ensure fair comparisons across models, both parameter and hyperparameter searches were systematically conducted. An exhaustive grid search was adopted, implemented via sklearn.model_selection.ParameterGrid [45]. Given a predefined grid of parameters, each with a discrete set of candidate values, ParameterGrid deterministically iterates over all possible parameter combinations using Python's built-in iterator. This approach is straightforward, deterministic, and trivially parallelizable, providing full coverage of small- to medium-sized search spaces without the computational overhead of more sophisticated search schemes.The Parameter and Hyperparameter spaces used for the models are provided in the Appendix A.

2.4 Unimodal Learning with Clinical Data

2.4.1 XGBoost

As first line of work for clinical modality, we adopted XGBoost[46], a powerful and scalable ML framework for gradient tree boosting. Widely recognized as a state-of-the-art method, XGBoost combines a rigorous theoretical basis in decision tree ensembles with outstanding computational efficiency and robust performance across numerous domains. Its success is largely attributed to its exceptional scalability, which allows it to achieve execution speeds more than an order of magnitude faster than many conventional solutions on a single machine, while also supporting large-scale problems with billions of instances in distributed or resource-constrained environments. Moreover, XGBoost is released as an open-source package, ensuring broad accessibility and seamless integration into modern data science pipelines.

Parameter Tuning And Hyperparameter Search

For XGBoost, we explored a grid over both capacity and stochasticity controls. Model complexity was varied via maximum tree depth (2 or 5) and the minimum child weight (1 or 5). Randomness was introduced through subsampling ratios (0.5 or 1.0) and column subsampling per tree (0.2 or 0.5). Finally, the learning rate was tuned at three levels (0.05, 0.1, 0.2), with the number of boosting rounds fixed at 100. Other hyperparameters, gamma, reg_alpha (ℓ_1 regularization) and reg_lambda (ℓ_2 regularization), objective, number of classes, evaluation metric, and tree method, were held constant for comparability and reproducibility. This yields a total of 48 XGBoost configurations, covering a conservative-to-moderate capacity regime.

In terms of hyperparameters, all runs followed the same protocol: a 20% test split, min–max normalization, two alternative class-weighting schemes (exponential vs. inverse), an inner validation split of 20% for model selection, and 100 train/val/test splits under a random seed. This procedure was consistently applied across all clinical unimodal approaches to ensure comparability of results.

2.4.2 MLP

As a second approach for the clinical modality, we employed the Multilayer Perceptron (MLP) [41], a neural network architecture widely adopted in ML and often described as a global approximator in multidimensional spaces.Its structure is organized in layers: an input layer that receives the features, one or more hidden layers that perform non-linear transformations, and an output layer that produces the final prediction. In its classical form, the MLP employs sigmoidal activation functions for the neurons, although other non-linearities (e.g., REctified Linear Unit (ReLU), Gaussian Error Linear Unit (GELU)) are commonly used in modern implementations. Connections are restricted to adjacent layers, and the architecture is strictly feedforward, meaning that information flows from the input to the output without recurrence.

Architecture

The multilayer perceptron (MLP) designed in this study is a neural network that receives as input a tensor of shape (37, 20), corresponding to 37 samples, each represented by 20 features.

Its architecture is organized into two hidden layers followed by an output layer. The first hidden layer projects the 20 input features into a 64-dimensional representation through a fully connected (Linear) layer. A GELU activation introduces non-linearity, and a Dropout layer is applied to mitigate overfitting. The second hidden layer reduces this representation to 32 units via another Linear transformation, again followed by GELU activation and Dropout regularization. Finally, the output layer maps the 32 hidden units to 3 output nodes through a Linear layer, producing the final class predictions. From a computational perspective, the model is highly compact, comprising a total of 3,523 trainable parameters.

A detailed summary of the architecture is presented in Table 2.3 for ease of reference.

Layer (type)	Input Shape	Output Shape	Kernel Shape	Param #
MLP	$[37, 20]$	$[37, 3]$	–	–
+ Linear	$[37, 20]$	$[37, 64]$	–	1,344
+ GELU	$[37, 64]$	$[37, 64]$	–	–
+ Dropout	$[37, 64]$	$[37, 64]$	–	–
+ Linear	$[37, 64]$	$[37, 32]$	–	2,080
+ GELU	$[37, 32]$	$[37, 32]$	–	–
+ Dropout	$[37, 32]$	$[37, 32]$	–	–
+ Linear	$[37, 32]$	$[37, 3]$	–	99
			Total params:	3,523
			Trainable params:	3,523

Table 2.3: MLP architecture summary

Parameter Tuning And Hyperparameter Search

For the MLP, we explored a grid of architectural and regularization choices. The first hidden layer was set to either 32 or 64 units, and the second to either 16 or 32 units. Each configuration was combined with two levels of dropout regularization (moderate at 0.2 or strong at 0.8) applied to both layers. The activation function was fixed to GELU throughout. This design results in 16 distinct MLP configurations, allowing us to systematically trade off model capacity against regularization while keeping the nonlinearity constant.

In line with the clinical unimodal framework, MLP adhered to the common hyperparameter protocol described for XGBoost.

2.5 Unimodal Learning with EEG

For the EEG modality in the unimodal setting, we adopted convolutional neural networks (CNNs) with both 2D and 3D kernels. Among the CNN architectures, we selected a Residual Neural Network (ResNet) [42, 43], chosen for its effectiveness in addressing the limitations of very deep networks. ResNet introduces residual (skip) connections, which allow the direct propagation of information across layers, thereby stabilizing the training process. The core building block of this architecture is the residual unit, composed of convolutional layers, batch normalization, and non-linear activation functions (typically ReLU). Within each block, the input x is added to the transformed output $F(x)$, yielding:

$$y = F(x) + x \tag{2.3}$$

This formulation enhances gradient flow during backpropagation, alleviating the vanishing gradient problem and enabling the training of substantially deeper models. A ResNet architecture is obtained by stacking multiple residual blocks, preceded by an initial stem stage that performs the first transformation of the input data. The depth of the network is defined by the number N of residual blocks, which characterizes the specific variant (e.g., ResNet-2, ResNet-10, ResNet-50).

2.5.1 Customized ResNet Architecture

In our experiments, we implemented a two-block ResNet variant, which offered an effective balance between model complexity and predictive performance. This lightweight architecture outperformed deeper alternatives while relying on a substantially reduced number of parameters. The modifications introduced in this variant include the removal of the initial stem stage, the addition of a pooling layer after the first residual unit, and the use of a stride of 1 in the first unit.

2.5.2 ResNet2-2D

Convolutional neural networks with two-dimensional (2D) kernels represent a natural baseline when dealing with inputs that can be represented in an image-like form, such as spectrograms, where the different "RGB" channels correspond to the various regions of interest (ROIs) considered. In this configuration, each ROI acts as an independent input channel, allowing the network to capture spatially localized spectral dynamics while maintaining a compact representation. A

further advantage of 2D kernels is that they considerably reduce the number of trainable parameters compared to higher-dimensional convolutional schemes. This property is particularly relevant in the present context, where the input size is limited and the risk of overfitting must be carefully controlled.

Architecture

The ResNet2D model processes inputs of shape (37, 11, 89, 83), where 37 represents the batch dimension.

The network is composed of two residual blocks, each built with 2D convolutional layers, batch normalization, and ReLU activations. The first residual block begins with a projection layer to align dimensions, followed by a 3×3 convolution . These layers are batch-normalized and passed through ReLU activations, then further processed with another 3×3 convolution and normalization. A max-pooling operation reduces spatial dimensions. The second residual block follows a similar structure but increases the number of filters to 32. An adaptive average pooling layer reduces the spatial features to [1, 1], and finally, a fully connected (Linear) layer maps the resulting 32-dimensional representation into 3 output classes. Overall, the model contains 18,691 trainable parameters.

For reference, a detailed overview of the architecture is provided in Table 2.4.

Layer (type)	Input Shape	Output Shape	Kernel Shape	Param #
ResNet2-2D	[37, 11, 89, 83]	[37, 3]	–	–
+ BasicBlock-2D	[37, 11, 89, 83]	[37, 16, 89, 83]	–	–
+ Conv2d	[37, 11, 89, 83]	[37, 16, 89, 83]	[1, 1]	176
+ Conv2d	[37, 11, 89, 83]	[37, 16, 89, 83]	[3, 3]	1,584
+ BatchNorm2d	[37, 16, 89, 83]	[37, 16, 89, 83]	–	32
+ ReLU	[37, 16, 89, 83]	[37, 16, 89, 83]	–	–
+ Conv2d	[37, 16, 89, 83]	[37, 16, 89, 83]	[3, 3]	2,304
+ BatchNorm2d	[37, 16, 89, 83]	[37, 16, 89, 83]	–	32
+ ReLU	[37, 16, 89, 83]	[37, 16, 89, 83]	–	–
+ MaxPool2d	[37, 16, 89, 83]	[37, 16, 44, 41]	2	–
+ BasicBlock-2D	[37, 16, 44, 41]	[37, 32, 44, 41]	–	–
+ Conv2d	[37, 16, 44, 41]	[37, 32, 44, 41]	[1, 1]	512
+ Conv2d	[37, 16, 44, 41]	[37, 32, 44, 41]	[3, 3]	4,608
+ BatchNorm2d	[37, 32, 44, 41]	[37, 32, 44, 41]	–	64
+ ReLU	[37, 32, 44, 41]	[37, 32, 44, 41]	–	–
+ Conv2d	[37, 32, 44, 41]	[37, 32, 44, 41]	[3, 3]	9,216
+ BatchNorm2d	[37, 32, 44, 41]	[37, 32, 44, 41]	–	64
+ ReLU	[37, 32, 44, 41]	[37, 32, 44, 41]	–	–
+ AdaptiveAvgPool2d	[37, 32, 44, 41]	[37, 32, 1, 1]	–	–
+ Linear	[37, 32]	[37, 3]	–	99

Total params: 18,691
Trainable params: 18,691

Table 2.4: ResNet2_2D architecture summary

Parameter Tuning And Hyperparameter Search

In the case of CNNs, the architectures were kept fixed throughout these experiments. As a result, the parameter search for this branch was reduced to a placeholder, without introducing additional combinations into the overall search space.

Across runs, the procedure hyperparameters were held constant: a 20% test split, two alternative class-weighting schemes (exponential vs. inverse), an inner validation split of 20% for model selection, and 100 train/val/test splits under a random seed.

2.5.3 ResNet2-3D

Convolutional neural networks with three-dimensional (3D) kernels extend the 2D formulation by introducing an additional dimension in the convolutional filters, making them particularly suited for tensor-structured inputs such as the EEG representation employed in this study. Unlike 2D kernels, which only capture spatial patterns within individual slices or channels, 3D kernels operate simultaneously across the ROI, frequency, and time dimensions. This property introduces the minimal inductive bias required to learn directly from volumetric representations,

avoiding the need for extensive feature engineering.

Architecture

The ResNet3D model operates on volumetric inputs of shape (37, 1, 11, 89, 83), where 37 denotes the batch size.

The network shares the same overall structure as the 2D variant, but replaces all convolutional kernels with their 3D counterparts to capture spatio-temporal dependencies. Overall, the model includes 49,635 trainable parameters.

For reference, a detailed overview of the architecture is provided in Table 2.5.

Layer (type)	Input Shape	Output Shape	Kernel Shape	Param #
ResNet2-3D	[37, 11, 89, 83]	[37, 3]	–	–
+ BasicBlock-3D	[37, 1, 11, 89, 83]	[37, 16, 11, 89, 83]	–	–
+ Conv3d	[37, 1, 11, 89, 83]	[37, 16, 11, 89, 83]	[1, 1, 1]	16
+ Conv3d	[37, 1, 11, 89, 83]	[37, 16, 11, 89, 83]	[3, 3, 3]	432
+ BatchNorm3d	[37, 16, 11, 89, 83]	[37, 16, 11, 89, 83]	–	32
+ ReLU	[37, 16, 11, 89, 83]	[37, 16, 11, 89, 83]	–	–
+ Conv3d	[37, 16, 11, 89, 83]	[37, 16, 11, 89, 83]	[3, 3, 3]	6,912
+ BatchNorm3d	[37, 16, 11, 89, 83]	[37, 16, 11, 89, 83]	–	32
+ ReLU	[37, 16, 11, 89, 83]	[37, 16, 11, 89, 83]	–	–
+ MaxPool3d	[37, 16, 11, 89, 83]	[37, 16, 5, 44, 41]	2	–
+ BasicBlock-3D	[37, 16, 5, 44, 41]	[37, 32, 5, 44, 41]	–	–
+ Conv3d	[37, 16, 5, 44, 41]	[37, 32, 5, 44, 41]	[1, 1, 1]	512
+ Conv3d	[37, 16, 5, 44, 41]	[37, 32, 5, 44, 41]	[3, 3, 3]	13,824
+ BatchNorm3d	[37, 32, 5, 44, 41]	[37, 32, 5, 44, 41]	–	64
+ ReLU	[37, 32, 5, 44, 41]	[37, 32, 5, 44, 41]	–	–
+ Conv3d	[37, 32, 5, 44, 41]	[37, 32, 5, 44, 41]	[3, 3, 3]	27,648
+ BatchNorm3d	[37, 32, 5, 44, 41]	[37, 32, 5, 44, 41]	–	64
+ ReLU	[37, 32, 5, 44, 41]	[37, 32, 5, 44, 41]	–	–
+ AdaptiveAvgPool3d	[37, 32, 5, 44, 41]	[37, 32, 1, 1, 1]	–	–
+ Linear	[37, 32]	[37, 3]	–	99

Total params: 49,635
Trainable params: 49,635

Table 2.5: ResNet2_3D architecture summary

2.6 Multimodal Learning with Late-Fusion

The multimodal learning frameworks, which integrate EEG and Clinical data, build upon the unimodal approaches. This section examines late fusion strategies, wherein independently trained models for each modality provide predictions that are subsequently aggregated into a final decision. A key property of late fusion is that it operates on probability distributions that are already normalized, meaning that the softmax function has been applied within each unimodal model.

Formally, the late fusion approach seeks to learn a function f that combines the individual distributions into a joint one, expressed as:

$$p_{CD+EEG} = f(p_{CD}, p_{EEG}) \tag{2.4}$$

2.6.1 Late Separate Average Fusion

Architecture

The probability outputs of the base models were treated as input features, and their integration was achieved via three alternative schemes: averaging, logarithmic averaging, and entropy-based fusion.
The first is the use of the arithmetic mean on the set of probabilities.

$$\mathrm{p}^{i}_{\mathrm{LSAF}} = \frac{1}{M}\left(\mathrm{p}^{i}_{\mathrm{CD}} + \mathrm{p}^{i}_{\mathrm{EEG}}\right), \quad i \in \{1, \ldots, C\} \tag{2.5}$$

The second is the use of the arithmetic mean on the logarithm of the probabilities.

$$\mathrm{p}^{i}_{\mathrm{LSAF}} = \sqrt[M]{\mathrm{p}^{i}_{\mathrm{CD}} \cdot \mathrm{p}^{i}_{\mathrm{EEG}}}, \quad i \in \{1, \ldots, C\} \tag{2.6}$$

The third consists in weighting the two sets of probabilities, by the entropy counterpart.

$$\mathrm{p}^{i}_{\mathrm{LSAF}} = \frac{H_{\mathrm{EEG}}}{H_{\mathrm{EEG}} + H_{\mathrm{CD}}}\mathrm{p}^{i}_{\mathrm{CD}} + \frac{H_{\mathrm{CD}}}{H_{\mathrm{EEG}} + H_{\mathrm{CD}}}\mathrm{p}^{i}_{\mathrm{EEG}}, \quad i \in \{1, \ldots, C\} \tag{2.7}$$

where C=3 classes, M=2 modalities, and the entropy is defined as

$$H = -\sum_{i=1}^{C} \mathrm{p}^{i} \log \mathrm{p}^{i}$$

Training Strategy

The training strategy for late separate averaging strategy does not introduce any additional learning stage. Training is therefore limited to the unimodal models, which are optimized independently.

Parameter Tuning And Hyperparameter Search

In the late separate averaging strategy, no additional parameter tuning or hyperparameter search was required, since the final prediction relied solely on the probability averaging of independently trained unimodal models. Consequently, optimization was confined to the unimodal branches, with no extra search space introduced at the fusion stage.

2.6.2 Leta Meta Fusion

Architecture

Within the late fusion framework, three alternative meta-classifiers were explored to integrate the predictions of the base models: XGBoost, AttentionFusion and GatedFusion. By leveraging the probability outputs of the base models as input features, the integration was carried out through an XGBoost meta-classifier.

The AttentionFusion block operated on concatenated probability vectors. A small sequential MLP was employed to refine these inputs. The first linear projection expanded the representation to 32 dimensions , followed by dropout regularization and a GELU activation . A second linear layer mapped the features to a single output unit. Overall, this architecture comprised 161 trainable parameters and the gating formulation is:

$$p^i_{\text{LMF}} = \frac{e^{\alpha_{\text{CD}}}}{e^{\alpha_{\text{CD}}} + e^{\alpha_{\text{EEG}}}} p^i_{\text{CD}} \; + \; \frac{e^{\alpha_{\text{EEG}}}}{e^{\alpha_{\text{CD}}} + e^{\alpha_{\text{EEG}}}} p^i_{\text{EEG}}, \quad i \in \{1, \ldots, C\}. \tag{2.8}$$

The GatedFusion block followed a complementary design. Both DL modules take as input the concatenated probability vectors of two base models, but differ in the fusion mechanism: AttentionFusion employs learned attention weights, whereas GatedFusion uses a gating strategy with sigmoid activations. In total, the GatedFusion module contained 257 trainable parameters and in the gating formulation is:

$$p^i_{\text{LMF}} = \alpha\, p^i_{\text{CD}} + (1 - \alpha)\, p^i_{\text{EEG}}, \quad i \in \{1, \ldots, C\}, \tag{2.9}$$

where C=3 classes, and the scalar $\alpha \in [0, 1]$ is a learnable parameter.

For reference, a detailed overview of the architectural blocks is provided in Tables 2.6, 2.7.

Layer (type)	Input Shape	Output Shape	Kernel Shape	Param #
AttentionFusion	[37, 3]	[37, 3]	–	–
+ Sequential	[37, 2, 3]	[37, 2, 1]	–	–
+ Linear	[37, 2, 3]	[37, 2, 32]	–	128
+ Dropout	[37, 2, 32]	[37, 2, 32]	–	–
+ GELU	[37, 2, 32]	[37, 2, 32]	–	–
+ Linear	[37, 2, 32]	[37, 2, 1]	–	33

Total params: 161
Trainable params: 161

Table 2.6: AttentionFusion architecture summary

Layer (type)	Input Shape	Output Shape	Kernel Shape	Param #
GatedFusion	$[37, 3]$	$[37, 3]$	–	–
+ Sequential	$[37, 6]$	$[37, 1]$	–	–
+ Linear	$[37, 6]$	$[37, 32]$	–	224
+ Dropout	$[37, 32]$	$[37, 32]$	–	–
+ GELU	$[37, 32]$	$[37, 32]$	–	–
+ Linear	$[37, 32]$	$[37, 1]$	–	33
+ Sigmoid	$[37, 1]$	$[37, 1]$	–	–

Total params: 257
Trainable params: 257

Table 2.7: GatedFusion architecture summary

Training Strategy

The training strategy for late meta-fusion introduces an additional learning stage beyond unimodal training. In this setup, the unimodal models are first trained independently to convergence, after which their probability outputs serve as input features for a second-level learner. This learner is then optimized to integrate the complementary decision patterns of the unimodal predictors, effectively decoupling branch-specific optimization from the fusion step while enhancing predictive integration.

Parameter Tuning And Hyperparameter Search

In the late meta-fusion, when the fusion head was implemented as a classical meta-learner, two alternatives were considered. For XGBoost, we adopted the same 48-point grid described in Section 2.4.1, allowing the meta-learner to span from shallow and strongly regularized to moderately deeper configurations, with stochastic subsampling and a small sweep over the learning rate. For learnable gating blocks, a compact four-point grid was explored by varying the hidden layer width (16 or 32 units) and the dropout rate (0.2 or 0.8), while keeping the activation fixed to GELU. This small, four-point grid is enough to toggle the gate between a compact/regularized setting and a slightly more expressive one.

Across runs, the hyperparameters were held constant: a 20% test split, log-probability fusion , two alternative class-weighting schemes (exponential vs. inverse), an inner validation split of 20% for model selection, and 100 train/val/test splits under a random seed.

2.7 Multimodal Learning with Middle-Fusion

As discussed in Section 2.6, multimodal learning builds upon unimodal approaches. In the middle fusion strategy, intermediate representations are extracted from both models, the MLP applied to clinical data and the CNN applied to EEG signals, at predefined layers. These feature vectors, which capture complementary aspects of the data, are then merged into a single joint representation through simple concatenation. Formally, the middle fusion approach seeks to learn

a function g that combines the feature representations extracted from the individual modalities into a joint embedding, expressed as:

$$\mathbf{z} = g(\mathbf{a}, \mathbf{b}) = [\mathbf{a}; \mathbf{b}], \tag{2.10}$$

where $\mathbf{a}$ and $\mathbf{b}$ denote the features extracted from clinical data and EEG, respectively. The fused vector $\mathbf{z}$ is subsequently passed to a shared classification head, which learns to exploit the complementary information across modalities.

Architecture

The Middle Fusion Model integrates multimodal information by combining a multilayer perceptron (MLP) branch with a convolutional backbone (ResNet2D/3D), both corresponding to the unimodal models, followed by a dedicated fusion head. The MLP branch receives as input a tensor of shape $(37, 20)$, whereas the ResNet3D branch processes spectrogram inputs of shape $(37, 11, 89, 83)$. At the fusion stage, the outputs of both branches are concatenated and passed through a sequential classification head composed of two fully connected layers, with GELU activation, batch normalization, and dropout applied between them. The full model has 59,689 trainable parameters.

For reference, a detailed overview of the architecture is provided in Table 2.8.

Layer (type)	Input Shape	Output Shape	Kernel Shape	Param #
Middle Fusion Model	[37, 20]	[37, 3]	–	–
+ MLP	[37, 20]	[37, 3]	–	–
+ Linear	[37, 20]	[37, 64]	–	1,344
+ GELU	[37, 64]	[37, 64]	–	–
+ Dropout	[37, 64]	[37, 64]	–	–
+ Linear	[37, 64]	[37, 32]	–	2,080
+ GELU	[37, 32]	[37, 32]	–	–
+ Dropout	[37, 32]	[37, 32]	–	–
+ Linear	[37, 32]	[37, 3]	–	99
+ ResNet2_3D	[37, 11, 89, 83]	[37, 3]	–	–
+ BasicBlock_3D	[37, 1, 11, 89, 83]	[37, 16, 11, 89, 83]	–	–
+ Conv3d	[37, 1, 11, 89, 83]	[37, 16, 11, 89, 83]	[1, 1, 1]	16
+ Conv3d	[37, 1, 11, 89, 83]	[37, 16, 11, 89, 83]	[3, 3, 3]	432
+ BatchNorm3d	[37, 16, 11, 89, 83]	[37, 16, 11, 89, 83]	–	32
+ ReLU	[37, 16, 11, 89, 83]	[37, 16, 11, 89, 83]	–	–
+ Conv3d	[37, 16, 11, 89, 83]	[37, 16, 11, 89, 83]	[3, 3, 3]	6,912
+ BatchNorm3d	[37, 16, 11, 89, 83]	[37, 16, 11, 89, 83]	–	32
+ ReLU	[37, 16, 11, 89, 83]	[37, 16, 11, 89, 83]	–	–
+ MaxPool3d	[37, 16, 11, 89, 83]	[37, 16, 5, 44, 41]	2	–
+ BasicBlock_3D	[37, 16, 5, 44, 41]	[37, 32, 5, 44, 41]	–	–
+ Conv3d	[37, 16, 5, 44, 41]	[37, 32, 5, 44, 41]	[1, 1, 1]	512
+ Conv3d	[37, 16, 5, 44, 41]	[37, 32, 5, 44, 41]	[3, 3, 3]	13,824
+ BatchNorm3d	[37, 32, 5, 44, 41]	[37, 32, 5, 44, 41]	–	64
+ ReLU	[37, 32, 5, 44, 41]	[37, 32, 5, 44, 41]	–	–
+ Conv3d	[37, 32, 5, 44, 41]	[37, 32, 5, 44, 41]	[3, 3, 3]	27,648
+ BatchNorm3d	[37, 32, 5, 44, 41]	[37, 32, 5, 44, 41]	–	64
+ ReLU	[37, 32, 5, 44, 41]	[37, 32, 5, 44, 41]	–	–
+ AdaptiveAvgPool3d	[37, 32, 5, 44, 41]	[37, 32, 1, 1, 1]	–	–
+ Linear	[37, 32]	[37, 3]	–	99
+ Sequential	[37, 96]	[37, 3]	–	–
+ Linear	[37, 96]	[37, 64]	–	6,208
+ GELU	[37, 64]	[37, 64]	–	–
+ BatchNorm1d	[37, 64]	[37, 64]	–	128
+ Dropout	[37, 64]	[37, 64]	–	–
+ Linear	[37, 64]	[37, 3]	–	195
	Total params:	59,689		
	Trainable params:	59,689		

Table 2.8: MiddleFusionModel architecture summary

Training Strategy

The training strategy for the middle fusion models required an end-to-end approach, in which a dedicated fusion module was optimized jointly with the unimodal branches. Unlike late fusion,

which combines already trained decision outputs, this strategy enabled the model to learn cross-modal interactions directly at the representation level. During training, the feature extractors for EEG and clinical data were updated simultaneously with the fusion head, ensuring that the intermediate representations from each modality were not only individually informative but also complementary when concatenated.

Parameter Tuning And Hyperparameter Search

In the middle fusion, the fusion block was predefined; therefore, no grid search at the model level was required.

Across runs, the procedure hyperparameters were held constant: a 20% held-out test split, min–max normalization , two alternative classweighting schemes (exponential vs. inverse), an inner validation split of 20% for model selection, and 100 train/val/test splits under a random seed.

3 Results

3.1 Evaluation Framework

The evaluation framework was structured to rigorously measure the performance of the proposed learning strategies on the held-out test set. After completing the training and validation phases, each model was applied to the test data to generate predicted class labels as described in Subsection 2.3.2. Performance was primarily measured using balanced accuracy, chosen for its robustness in the presence of class imbalance.

In addition, the evaluation included a systematic comparison between unimodal and multimodal approaches, the results of which are summarized in detail in the Appendix.A. While unimodal baselines relied exclusively on either EEG or clinical data, the multimodal models integrated both modalities, thereby exploiting complementary sources of information. Fusion strategies, Late and Middle fusion, were benchmarked against unimodal counterparts. This comparative analysis quantified the added value of multimodal integration and tested whether fusion mechanisms yielded statistically significant improvements in predictive accuracy.

3.2 Model Outputs and Performance Metrics

The predicted class labels represent the direct outputs of the models. These outputs served as an immediate indicator of predictive capability and served as the basis for quantitative assessment.

Performance assessment was centered on balanced accuracy[47], a metric particularly suitable in the presence of class imbalance. To obtain a robust estimate across repeated experiments, performance was summarized using the median balanced accuracy calculated over 100 random stratified splits.

In addition, the interquartile range (IQR) was reported to capture variability, offering an indication of the stability of model performance across different partitions of the dataset. Together, the median balanced accuracy and IQR provided complementary insights into both the central

tendency and consistency of the results.

$$\text{Balanced Accuracy} = \frac{1}{C}\sum_{i=1}^{C}\frac{TP_i}{TP_i + FN_i} \tag{3.1}$$

where $C = 3$ classes , TP_i = number of true positives for class-i, FN_i = number of false negatives for class-i.

$$\widetilde{\text{BA}} = \text{median}\left(\text{BA}_1, \text{BA}_2, \ldots, \text{BA}_{100}\right) \tag{3.2}$$

$$\text{IQR}_{\text{BA}} = Q_3\left(\text{BA}_{1..100}\right) - Q_1\left(\text{BA}_{1..100}\right) \tag{3.3}$$

where Q_1 = first quartile (25th percentile) of the balanced accuracy distribution, Q_3 = third quartile (75th percentile) of the balanced accuracy distribution.

3.3 Unimodal Learning Results

3.3.1 Unimodal with EEG

To refine the choice of the most suitable convolutional network for EEG analysis, we compared the performance of two ResNet variants a 2D version and its 3D extension. Each architecture was evaluated in both its reference configuration and a modified version incorporating lightweight adjustments. Specifically, the variants include: ResNet2-2D_reference, ResNet2-2D_modified, ResNet2-3D_reference, ResNet2-3D_modified.

The boxplots in Figure 3.1 summarize the distributions of balanced accuracy across models, highlighting medians and interquartile ranges (IQRs).

The results demonstrate consistent improvements for both architectures when moving from the reference to the modified configuration, underscoring the benefits of architectural refinements. For this reason, the modified versions are adopted as the unimodal EEG baselines in all subsequent analyses and referred to as ResNet2-2D and ResNet2-3D. More importantly, ResNet2-3D_modified clearly outperforms its 2D counterpart across all comparisons. This evidence strongly supports the adoption of ResNet2-3D as the backbone for subsequent analyses. By explicitly modeling the spatio-temporal structure of the EEG spectrograms, the 3D convolutional approach better captures the underlying signal dynamics, making it the most effective unimodal deep learning strategy for the classification of AD, FTD, and CN subjects.

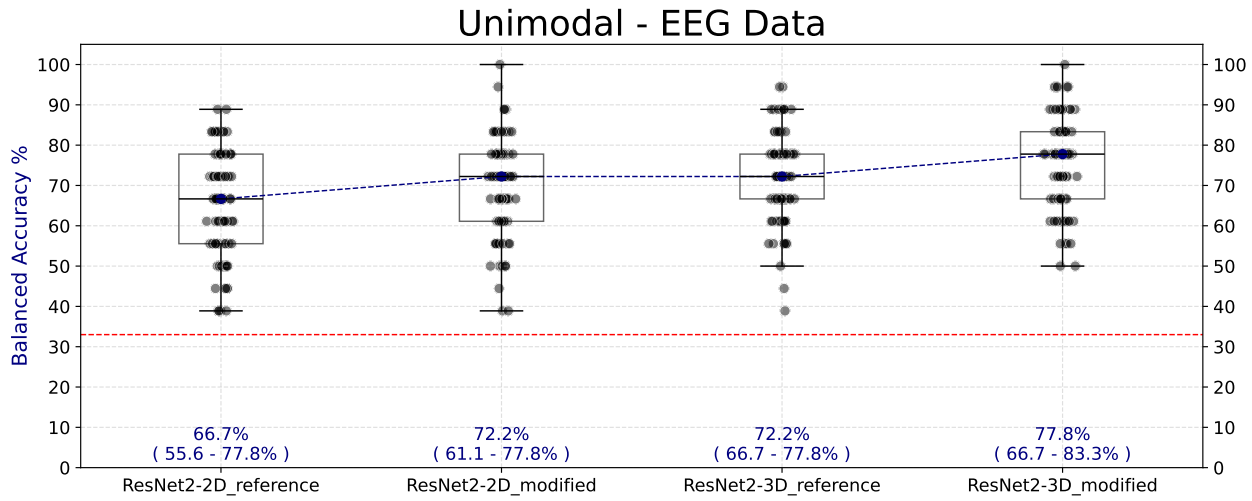

Figure 3.1: Unimodal with EEG Data Performance Results
Balanced accuracy distributions for ResNet2-2D and ResNet2-3D, each in reference vs modified configuration. Black dots are the balanced accuracies obtained across 100 train/validation/test splits inside a stratified repeated hold-out strategy; the boxplots summarize their distribution (median and interquartile range, with whiskers); the blue dashed line connects the medians across methods; the red dashed line indicates the 33% chance level for a 3-class problem. The values printed below each box report the median (top) and the IQR (bottom). Both models improve from reference to modified, and 3D consistently outperforms 2D, with **ResNet2-3D_modified** yielding the highest median balanced accuracy.

3.3.2 Unimodal with CD or EEG

The evaluation of unimodal models provides the first insight into the relative contribution of each data modality to the classification task. Four learning strategies were tested independently on EEG and clinical data: XGBoost (CD), MLP (CD), ResNet2-2D (EEG), and ResNet2-3D (EEG).

The boxplots in Figure 3.2 summarize the distributions of balanced accuracy across models, highlighting medians and interquartile ranges (IQRs).

Results indicate that EEG-based models consistently outperform those relying on clinical features alone, highlighting the stronger discriminative power of electrophysiological markers compared to cognitive and demographic measures. Among the EEG-based approaches, the ResNet2-3D architecture achieves the highest median balanced accuracy. In contrast, models trained exclusively on clinical data achieve lower performance, confirming their utility as a baseline but underscoring the added value of EEG in capturing disease-specific signatures. This comparative analysis between EEG and clinical unimodal baselines thus establishes a clear hierarchy of modalities, where electrophysiological information contributes more strongly to the classification of AD, FTD, and CN subjects. These results provide a reference point for subsequent multimodal experiments, where the integration of complementary sources of information aims to further enhance classification performance.

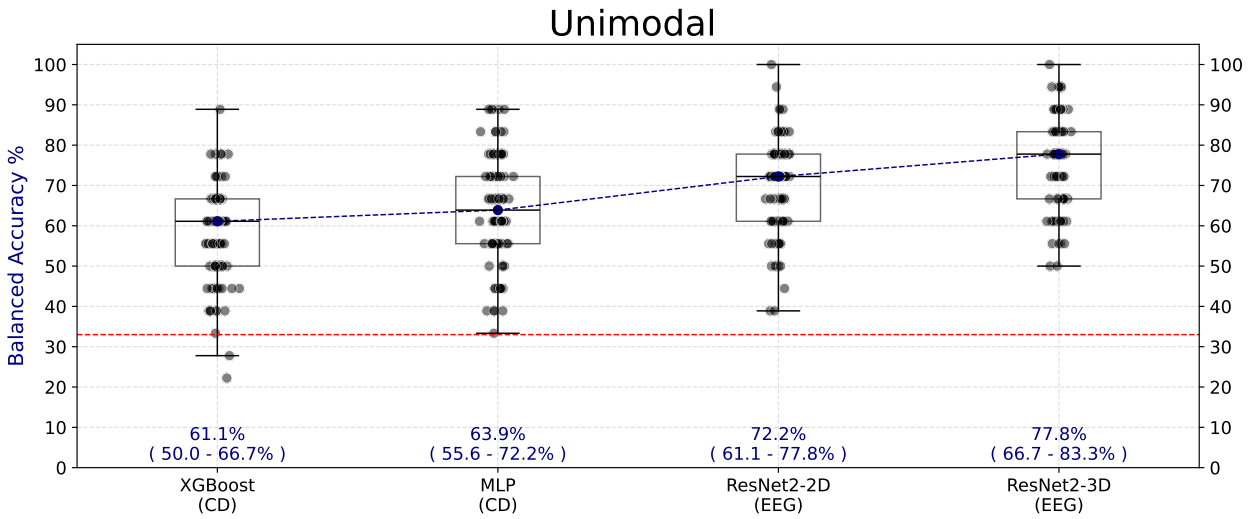

Figure 3.2: Unimodal Performance Results

Balanced accuracy distributions for four unimodal models: XGBoost, MLP, CNN2D, and CNN3D. Black dots are the balanced accuracies obtained across 100 train/validation/test splits inside a stratified repeated hold-out strategy; the boxplots summarize their distribution (median and interquartile range, with whiskers); the blue dashed line connects the medians across methods; the red dashed line indicates the 33% chance level for a 3-class problem. The values printed below each box report the median (top) and the IQR (bottom). Overall, **ResNet2-3D** yields the highest median balanced accuracy, and EEG modalities outperforms the clinicals.

3.4 Late-Fusion Results

3.4.1 Late Separate Average Fusion

The first Late fusion strategy investigated was the Separate Average approach, applied in three alternative variants: Average, Log-Average, and Entropy-based.

The boxplots in Figure 3.3 summarize the distributions of balanced accuracy across models, highlighting medians and interquartile ranges (IQRs).

The results highlight two key trends. First, MLP-based fusion outperforms XGBoost-based fusion, achieving higher median balanced accuracy across both 2D and 3D feature sets. Second, models leveraging 3D CNN features surpass those based on 2D CNN features, confirming the added discriminative power of spatiotemporal representations. Importantly, all multimodal fusion configurations substantially exceed chance performance. Among models with comparable performance, the best trade-off between accuracy and stability is achieved by the MLP combined with ResNet2-3D predictions under the Log-Average fusion scheme. This configuration yields both the highest median balanced accuracy and the narrowest IQR, indicating superior predictive power alongside reduced variability across repeated splits.

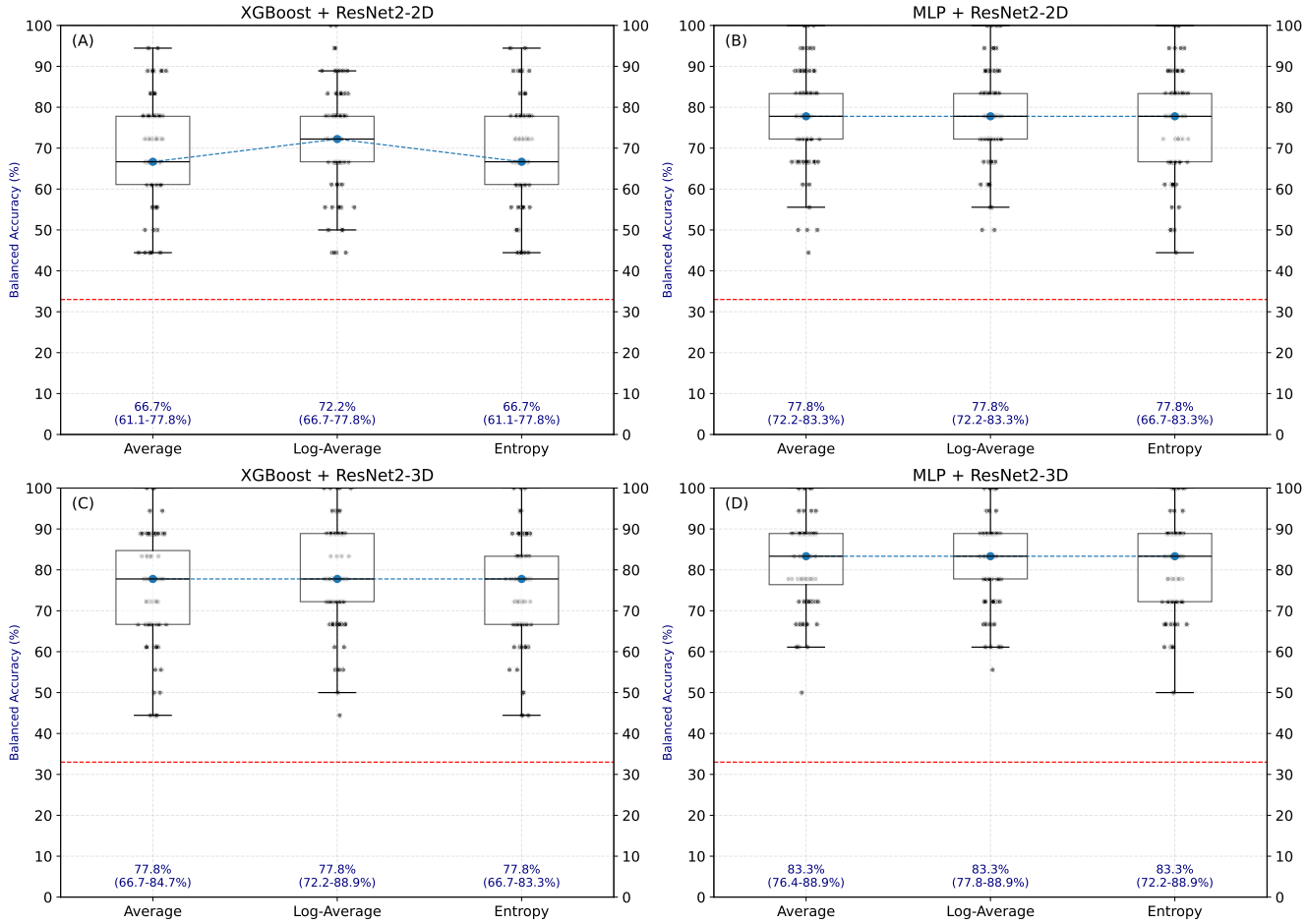

Figure 3.3: Late Separate Average Performance Results
Panels (A) and (B) report results obtained with 2D CNN predictions combined with XGBoost and MLP classifiers, respectively. Panels (C) and (D) show the analogous results using 3D CNN predictions . For each separate average fusion strategy (Average, Log-Average, Entropy), the black dots represent the balanced accuracies obtained across 100 train/validation/test splits under a stratified repeated hold-out strategy; the boxplots summarize their distribution (median and interquartile range, with whiskers), the blue dashed lines connect the medians across methods, and the red dashed line indicates the 33% chance level for a 3-class problem. The values printed below each box report the median (top) and the IQR (bottom). Overall, MLP-based fusion yields higher median balanced accuracy than XGBoost-based fusion, and 3D features outperform 2D, with all multimodal configurations clearly above chance. Among models with comparable performance, the best is **MLP + ResNet2-3D with Log-Average fusion**, which achieves the narrowest IQR.

3.4.2 Late Meta Fusion

A second Late fusion strategy explored in this study involved the use of meta-classifiers to integrate the probability outputs of the unimodal models. Three different approaches were considered: XGBoost (machine learning–based mechanism), Gated Fusion (deep learning–based mechanism) and Attention Gate (deep learning–based mechanism).

The boxplots in Figure 3.4 summarize the distributions of balanced accuracy across models, highlighting medians and interquartile ranges (IQRs).

The results reveal a clear trend: deep learning–based meta-classifiers (Gated Fusion and Attention Gate) consistently outperform the XGBoost meta-classifier. In addition to achieving higher balanced accuracy, they also demonstrate greater stability, as indicated by narrower interquartile ranges across splits. Notably, when combined with ResNet2-3D features, the Attention Gate yields the best balance between performance and consistency, with a higher frequency

of near-perfect outcomes and an elevated minimum balanced accuracy across splits. Overall, these findings indicate that deep learning–based meta-classifiers are better suited than traditional machine learning methods to capture the complex dependencies between modalities.

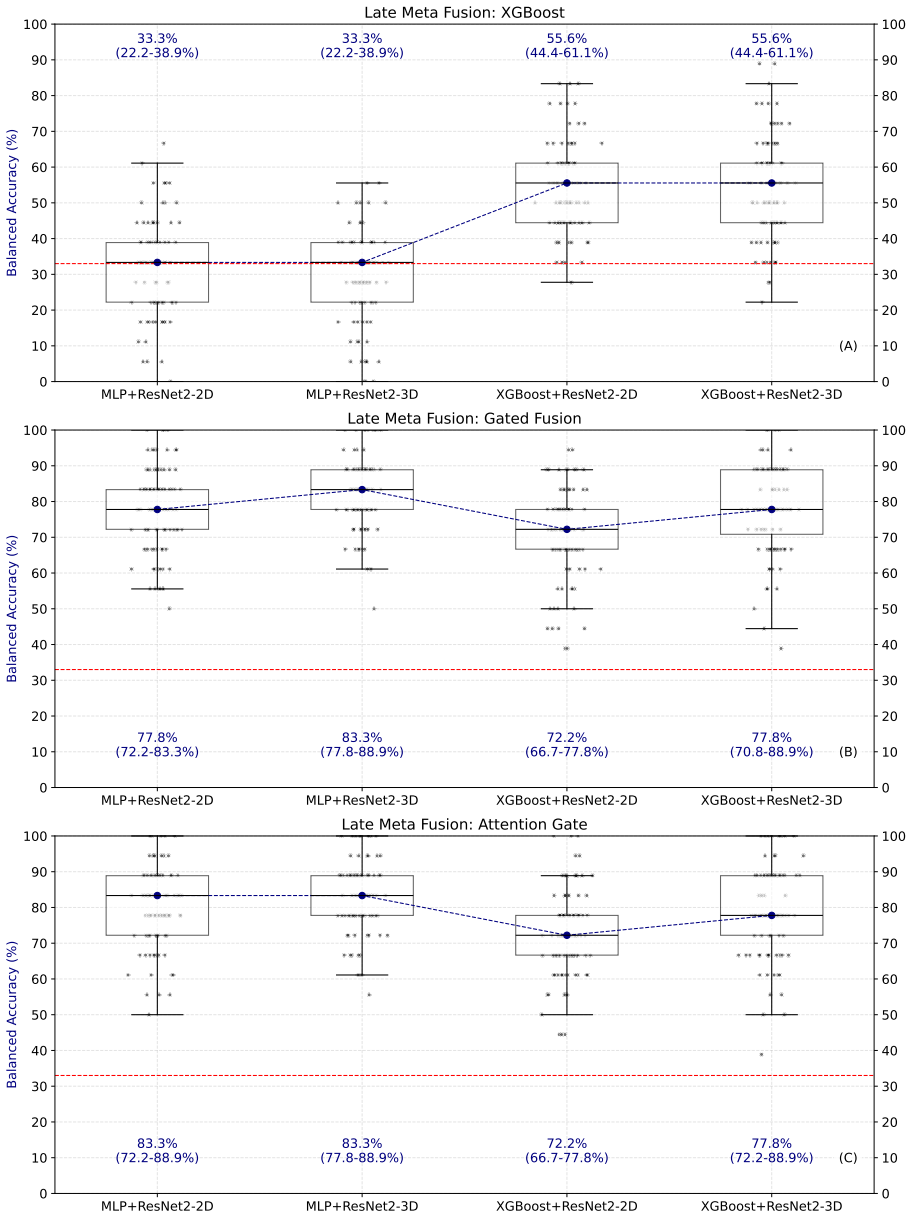

Figure 3.4: Late Meta Fusion Performance Results
Panels (A), (B), and (C) report the performance of three different late meta-fusion strategies: (A) XGBoost meta-classifier, (B) Gated Fusion, and (C) Attention Gate. For each configuration, the black dots represent the balanced accuracies obtained across 100 stratified train/validation/test splits; the boxplots summarize their distribution (median and interquartile range, with whiskers). The blue dashed lines connect the medians across methods, while the red dashed line indicates the 33% chance level for a 3-class classification problem. The values printed below each box correspond to the median (top) and interquartile range (IQR, bottom). Overall, Gated Fusion and Attention Gate yield more consistent performance than XGBoost. Moreover, the best-performing model at equal median and IQR is the **XGBoost + ResNet2-3D with Attention Gate** with a higher number of 100% occurrences and a higher minimum value.

3.5 Middle-Fusion Results

The third integration strategy explored was Middle Fusion. In this framework, different layers were selected to represent the modalities:

- MLP features:
  - X: clinical input features (input size(1,20))
  - FC1: first fully connected hidden layer of the MLP (64 units)
  - FC2: second fully connected hidden layer of the MLP (32 units)
  - LOGITS: final pre-softmax layer of the MLP (3 output nodes)
- CNN features:
  - GAP: global average pooling layer applied to convolutional feature maps (32 units)
  - LOGITS: final pre-softmax layer of the CNN classifier (3 output nodes)

Each fusion scheme corresponds to concatenating one MLP representation (X, FC1, FC2, or LOGITS) with one CNN representation (GAP or LOGITS), followed by a joint classifier head.

The boxplots in Figure 3.5 summarize the distributions of balanced accuracy across models, highlighting medians and interquartile ranges (IQRs).

Results indicate that all multimodal configurations outperform the chance level, confirming the benefit of middle-level feature integration. The most effective configuration overall is MLP FC1 features concatenated with 3D CNN GAP representations, which not only delivers the highest balanced accuracy but also exhibits the narrowest IQR, a higher number of near-perfect accuracies, and the strongest minimum performance across splits.

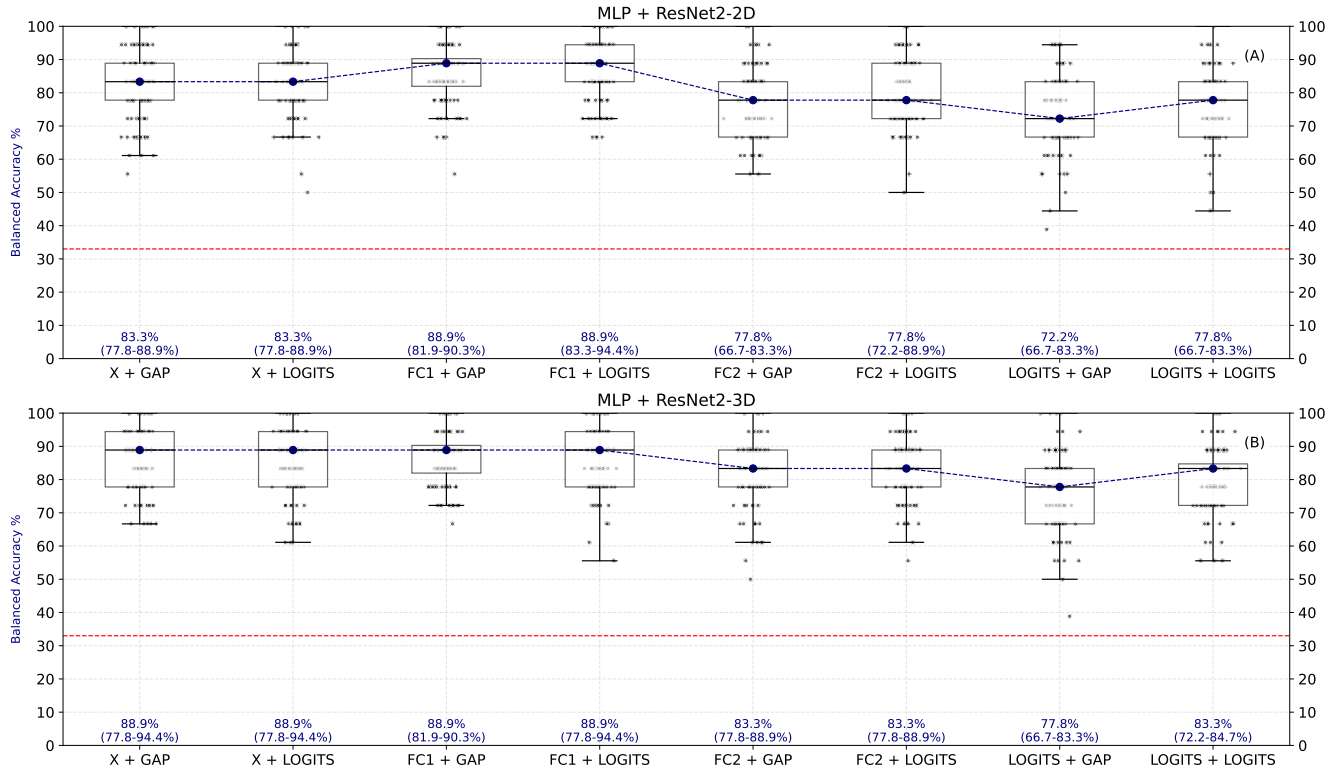

Figure 3.5: Middle Fusion Performance Results
Panel (A) reports results using 2D CNN; panel (B) shows the analogous results for 3D CNN. Each box refers to Middle fusion by concatenation of one MLP features (X, FC1, FC2, or LOGITS) with one CNN features (GAP or LOGITS), followed by the same prediction head. For each method, the black dots are the balanced accuracies obtained across 100 train/validation/test splits inside a stratified repeated hold-out strategy; the boxplots summarize their distribution (median and interquartile range, with whiskers); the blue dashed line connects the medians across methods; the red dashed line indicates the 33% chance level for a 3-class problem. The values printed below each box report the median (top) and the IQR (bottom). Overall, MLP FC1 combined with CNN Features tend to yield the best performance and IQR, with all multimodal configurations clearly above chance. Moreover, the best-performing model at equal median and IQR is the **FC1+GAP3D** with a higher number of 100% occurrences and a higher minimum value.

3.6 Results of Top Models Across Learning Approaches

To synthesize the evidence presented in the previous sections, we compare the performance of the four best-performing models identified across the Unimodal, Late fusion, and Middle fusion strategies. Specifically, the comparison includes:

- MLP (Unimodal CD model),
- ResNet2-3D (Unimodal EEG model),
- MLP + ResNet3D with Log-Average fusion(Late Separate Average Fusion),
- MLP + ResNet3D with Meta-Attention fusion (Late Meta Fusion),
- MLP FC1 + CNN3D GAP concatenation (Middle fusion).

The boxplots in Figure 3.6 summarize the distributions of balanced accuracy across models, highlighting medians and interquartile ranges (IQRs).

The results confirm the trends observed earlier. First, EEG-based unimodal models consistently outperformed those relying solely on clinical data, and the ResNet2-3D unimodal baseline already outperformed its 2D counterpart. Building upon this foundation, late fusion strategies with MLP integration improved performance over the unimodal baseline, with the Log-Average variant showing stronger consistency and robustness with no additional complexity. Finally, the middle fusion strategy, specifically the FC1 + GAP3D concatenation, achieved the highest overall performance. This configuration not only delivered the best median balanced accuracy but also demonstrated reduced variability, indicating greater stability across data partitions.

In conclusion, while each learning approach contributed valuable insights, the comparative analysis highlights the superiority of Middle-level multimodal integration. By combining MLP-derived features at the first fully connected layer with CNN3D-derived global average pooled representations, this model provided the most reliable and accurate predictions, underscoring the benefit of joint feature learning over both unimodal and late fusion strategies.

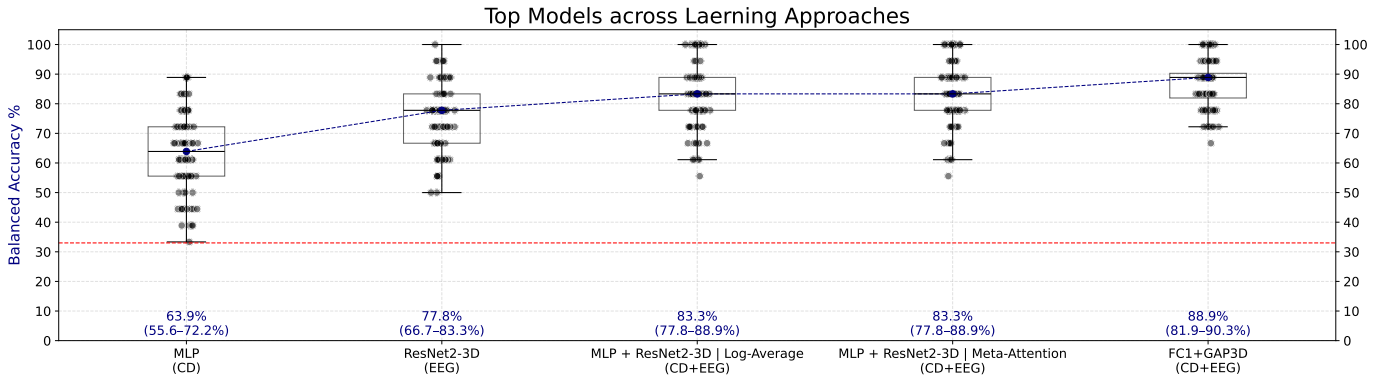

Figure 3.6: Top Models Performance Results

Balanced-accuracy distributions for five top-performing models accross Learning Approaches: MLP, ResNet2-3D baseline, MLP+ResNet3D (Log-Average fusion), MLP+ResNet3D (Meta-Attention fusion), and ResNet3D with FC1+GAP concatenation. Black dots are the balanced accuracies obtained across 100 train/validation/test splits inside a stratified repeated hold-out strategy; the boxplots summarize their distribution (median and interquartile range, with whiskers); the blue dashed line connects the medians across methods; the red dashed line indicates the 33% chance level for a 3-class problem. The values printed below each box report the median (top) and the IQR (bottom). Overall, MLP-based late fusion improves over the unimodal baseline, while **the Middle fusion with FC1+GAP3D** achieves the highest median balanced accuracy with reduced variability.

4 Discussion

4.1 Key Findings

This performance-oriented study yields several key insights into the relative contributions of different modalities and learning strategies for the classification of Alzheimer's disease (AD), frontotemporal dementia (FTD), and cognitively normal (CN) controls:

- Efficiency of modalities: Both EEG and clinical data are fast, cost-effective, and clinically accessible, making them particularly valuable for supporting diagnostic tasks.
- Modality hierarchy: EEG carries stronger discriminative power than clinical scores, confirming its central role in dementia classification.
- Volumetric advantage: ResNet2-3D outperformes its 2D baseline, showing the effectiveness of volumetric spatio-temporal representations.
- Middle fusion superiority: first fully connected hidden layer of the MLP concatenates with global average pooling layer features achieves the highest accuracy and lowest variability, proving the benefits of joint feature learning.
- Multimodal gains without added complexity: Late fusion, especially MLP+ResNet2-3D Log-Average, improves performance over unimodal baselines without additional computational complexity.

Taken together, these findings indicate that while EEG remains the most informative unimodal source, the practical accessibility of both EEG and clinical data further strengthens their clinical relevance. At the same time, Multimodal fusion strategies consistently surpass Unimodal models: Middle fusion achieves the most robust and accurate performance, whereas Late fusion provides a computationally efficient alternative for achieving meaningful gains without increasing model complexity. The following Subsections provide a more detailed discussion of each of these findings, their implications, and their relation to existing literature.

4.1.1 Efficiency of Modalities

Our experiments confirmed that electroencephalography (EEG) and clinical data (CD) provide complementary perspectives on brain function and cognitive performance, while remaining both cost-effective and time-efficient. In contrast, other commonly used modalities such as structural and functional MRI or PET (see Subsections 1.1.2, 1.1.3), though highly informative, are considerably more resource-intensive. Despite these low demands, EEG and CD together achieved three-class classification accuracies of up to 88.9% (81.9–90.3%) with the best performing model (see Figure 3.6), underscoring their diagnostic value and practical suitability for both routine clinical applications and large-scale research studies. This perspective is echoed by Cassani *et al.*(2018) [48]. Similarly, Costanzo *et al.*(2024) [49], in a systematic review of 55 studies, highlighted the promise of neurophysiological biomarkers, particularly EEG spectral power and event-related potentials, as reliable predictors of conversion to Alzheimer's disease.

4.1.2 Modality Hierarchy

Unimodal analyses revealed a clear hierarchy across modalities. EEG consistently outperformed clinical scores, reaching balanced accuracies up to 77.8%(66.7-83.3%) compared with 63.9% (55.6-72.2%) for clinical assessments alone (see Figure3.2). This gap highlights the central role of electrophysiological information in detecting disease-related alterations in brain activity and confirms EEG as the most informative unimodal source. These results align with previous work, specifically Dottori *et al.*(2017) [14], discussed in Subsection 1.1.3, which had already demonstrated the superior discriminative power of EEG over clinical measures.

4.1.3 Volumetric Advantage

Among the unimodal baselines, ResNet2-3D consistently outperformed its 2D counterpart, yielding an improvement of approximately 5.6% percentage points in balanced accuracy (see Figure 3.2). This highlights the added value of volumetric spatio-temporal representations, which more effectively capture the dynamics and structure of EEG spectrograms, thereby strengthening discriminative performance. A similar advantage was also observed in the multimodal configurations, where volumetric features continued to provide measurable improvements. These findings align with Bird *et al.*(2021) [50], who reported that 3D CNNs surpassed 2D CNNs in EEG-based classification tasks, particularly in emotional state and eye state recognition, achieving accuracies of 98.4% and 97.96%, respectively.

4.1.4 Middle Fusion Superiority

The Middle fusion strategy, specifically the FC1 + GAP3D concatenation, achieved the highest overall performance, with a median balanced accuracy of 88.9% and a reduced variability (81.9-90.3%) across 100 random stratified splits (see Figure 3.5). This stability indicates the advantage of joint feature learning, where multimodal representations leverage complementary information more effectively than unimodal approaches. However, external evidence is partially concordant: Vortmann *et al.*(2022) [51] reported that intermediate fusion of EEG and eye-tracking features often outperformed unimodal and early fusion strategies but was not consistently superior to late fusion.

4.1.5 Multimodal Gains Without Added Complexity

Finally, Late fusion strategies further improved results over unimodal baselines, even when using simple probability-level integration. In particular, MLP-based late fusion with Log-Average achieved balanced accuracies of 83.3%(77.8-88.9%), outperforming both unimodal EEG and clinical inputs while avoiding additional architectural complexity (see Figures3.3, 3.6). These findings demonstrate that even lightweight multimodal schemes can provide substantial benefits, offering an efficient trade-off between predictive performance and computational cost. This interpretation parallels the conclusions of Kline *et al.*(2022) [52], who discussed late fusion strategies, such as Late Elastic Average Fusion, in the context of precision health.

4.2 Strengths

Beyond the empirical findings, this study presents notable strengths at both the data and model levels:

- Data-level strength : Careful preparation and validation of the input data ensured reliable and representative samples, providing a solid foundation for fair and reproducible evaluation. In particular, custom EEG feature extraction was carried out through ROI-based spectrogram construction, as described in Subsection 2.2.3. This strategy was motivated by the rationale that key differences between diagnostic groups stem from disease-specific alterations in distinct brain regions. By capturing spatially localized spectral dynamics, this approach enhanced both the biological plausibility and the discriminative power of the input features.
- Model-level strength : The benchmarking framework was applied consistently across architectures and modalities, encompassing a total of 46 models evaluated under the same experimental conditions. By covering a broad range of configurations, the framework provided a comprehensive perspective on relative strengths and weaknesses, enabling robust and fair comparisons and offering a solid reference point for future studies in multimodal dementia classification.

4.3 Limitation

This study has one primary limitation that should be acknowledged: the analysis was conducted on a relatively small cohort of subjects. This inevitably restricts the statistical power of the results and may limit their generalizability. To mitigate this issue, the architectures were deliberately designed with a reduced number of trainable parameters.

Looking ahead, we hope that future releases of the BrainLat dataset or other resourches will include larger and more diverse cohorts of subjects. Expanding the dataset would enable more reliable validation of the proposed models, facilitate subgroup analyses across different dementia variants, and ultimately strengthen the clinical applicability of multimodal learning strategies. While the present work demonstrated promising results, the extension to larger samples remains essential for confirming their robustness and translational value

4.4 Future Directions

Building upon the key findings, the strengths, and limitation of this thesis, several avenues for future research can be identified. These perspectives can be articulated at two levels: the data, and the model.

4.4.1 Data Level

On one hand, among the different modalities considered, EEG emerged as the most discriminative unimodal source, confirming its central role in dementia classification. On the other hand, one of the strengths of this work was the careful preparation and validation of the input data, including the use of ROI-based EEG spectrograms. Consequently, further research should prioritize EEG-based approaches, with a particular emphasis on improving feature extraction strategies tailored to the pathophysiological characteristics of different dementia subtypes. A promising direction involves refining spectrogram construction through a region-of-interest (ROI) based approach guided by disease-specific neuroanatomical alterations. For Alzheimer's disease (AD), spectrograms could be generated specifically from ROIs corresponding to the temporal, parietal, occipital, and posterior cingulate cortices, regions typically affected by AD pathology. For frontotemporal dementia (FTD), spectrograms could instead focus on frontal and anterior temporal regions, which are preferentially impaired in this condition. By aligning spectrogram generation with the known topography of disease-related alterations, the resulting features would be more representative of the underlying pathology. This strategy would provide a dual advantage: (i) enhancing the discriminative power of EEG-based models by emphasizing spatially localized spectral patterns most relevant to diagnostic differentiation, and (ii) reducing the computational burden by focusing the feature space on a smaller set of clinically meaningful regions while still preserving essential spatial information.

4.4.2 Model Level

At the model level, future research could expand the benchmarking framework by testing new middle-fusion configurations, which in this study delivered the most robust results. A promising idea would be to integrate gating mechanisms into the fusion process, allowing the model to dynamically weigh the contribution of each modality. Furthermore, dimensionality management could be based on projection layers that align and transform modality-specific feature spaces before joint learning. Such extensions would combine the robustness of middle fusion with the flexibility of adaptive weighting, potentially leading to more powerful and generalizable multimodal architectures.

5 Conclusion

This thesis systematically evaluated Unimodal and Multimodal learning strategies for the three-class classification of Alzheimer's disease (AD), frontotemporal dementia (FTD), and cognitively normal (CN) controls using EEG and clinical data.

The results highlighted the practical value of these modalities, the strong discriminative power of EEG, the added benefit of volumetric spatio-temporal representations, and the consistent improvements provided by multimodal fusion, particularly through middle fusion, but also with lightweight approaches that did not increase architectural complexity.

The main strengths of this work lie in the careful preparation and validation of the dataset, the use of ROI-based spectrogram feature extraction to enhance the biological plausibility of EEG inputs, and the extensive benchmarking across diverse strategies. At the same time, the relatively small cohort represents a limitation, potentially affecting the generalizability of the findings.

This work highlights the potential of multimodal learning in the biomedical domain. Despite multimodal learning was applied to several tasks, its advantages are still to be fully explored. Multimodal learning can be beneficial for biomedical sciences as it allows to create models benefitting of information included in multiple modalities. In the case presented in this paper, EEG can provide knowledge related to the physiological and pathophysiological activity of the brain in the patients, while clinical data can capture the cognitive and behavioral consequences of disease. In the future, we expect multimodal learning to strongly improve medical sciences, allowing to infer relationships between features of different modalities and to improve medical knowledge as well as clinical practice.

Future research should build on these findings by further prioritizing EEG-based approaches with refined ROI-based spectrogram extraction tailored to disease-specific patterns, exploring novel Middle-fusion strategies such as gating with projection mechanisms to enhance multimodal integration, and validating the proposed models on larger and more diverse cohorts to improve robustness, generalizability, and clinical applicability.

References

[1] S. R. Stahlschmidt, B. Ulfenborg, and J. Synnergren, "Multimodal deep learning for biomedical data fusion: a review," *Briefings in Bioinformatics*, vol. 23, no. 2, p. bbab569, 2022.

[2] J. Duan, J. Xiong, Y. Li, and W. Ding, "Deep learning based multimodal biomedical data fusion: An overview and comparative review," *Information Fusion*, vol. 112, p. 102536, 2024.

[3] M. Salvi, H. W. Loh, S. Seoni, P. D. Barua, S. García, F. Molinari, and U. R. Acharya, "Multi-modality approaches for medical support systems: A systematic review of the last decade," *Information Fusion*, 2023.

[4] F. Mohsen, H. Ali, N. El Hajj, and Z. Shah, "Artificial intelligence-based methods for fusion of electronic health records and imaging data," *Scientific Reports*, vol. 12, no. 1, p. 17981, 2022.

[5] S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren, "Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines," *NPJ Digital Medicine*, vol. 3, p. 136, 2020.

[6] D. Zhang and D. Shen, "Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in alzheimer's disease," *NeuroImage*, vol. 59, no. 2, pp. 895–907, 2012.

[7] N. Bhagwat, J. D. Viviano, A. N. Voineskos, and M. M. Chakravarty, "Modeling and prediction of clinical symptom trajectories in alzheimer's disease using longitudinal data," *PLoS Computational Biology*, vol. 14, no. 9, p. e1006376, 2018.

[8] Y. Wang *et al.*, "Diagnosis and prognosis of alzheimer's disease using brain morphometry and white matter connectomes," *NeuroImage: Clinical*, vol. 23, p. 101859, 2019.

[9] O. Pelka *et al.*, "Sociodemographic data and apoe-ε4 augmentation for mri-based detection of amnestic mild cognitive impairment using deep learning systems," *PLOS ONE*, vol. 15, no. 9, p. e0236868, 2020.

[10] Y. Wang, X. Liu, and C. Yu, "Assisted diagnosis of alzheimer's disease based on deep learning and multimodal feature fusion," *Complexity*, vol. 2021, p. 6626728, 2021.

[11] M. Golovanevsky, C. Eickhoff, and R. Singh, "Multimodal attention-based deep learning for alzheimer's disease diagnosis," *Journal of the American Medical Informatics Association*, vol. 29, no. 12, pp. 2014–2022, 09 2022.

[12] M. N. K. P. and T. P., "Alzheimer's classification using dynamic ensemble of classifiers selection algorithms: A performance analysis," *Biomedical Signal Processing and Control*, vol. 68, p. 102729, 2021.

[13] F. R. Farina, D. D. Emek-Savaş, L. Rueda-Delgado, R. Boyle, H. Kiiski, G. Yener, and R. Whelan, "A comparison of resting state eeg and structural mri for classifying alzheimer's disease and mild cognitive impairment," *NeuroImage*, vol. 215, p. 116795, 2020.

[14] M. Dottori *et al.*, "Towards affordable biomarkers of frontotemporal dementia: A classification study via network's information sharing," *Scientific Reports*, vol. 7, no. 1, p. 3822, Jun 2017.

[15] L. Ilias and D. Askounis, "Multimodal deep learning models for detecting dementia from speech and transcripts," *Frontiers in Aging Neuroscience*, vol. 14, p. 830943, 2022.

[16] S. Liu *et al.*, "Multimodal neuroimaging feature learning for multiclass diagnosis of alzheimer's disease," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 4, pp. 1132–1140, 2015.

[17] J. Shi, X. Zheng, Y. Li, Q. Zhang, and S. Ying, "Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of alzheimer's disease," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 1, pp. 173–183, 2018.

[18] X. Gao, F. Shi, D. Shen, and M. Liu, "Task-induced pyramid and attention gan for multimodal brain image imputation and classification in alzheimer's disease," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 1, pp. 36–43, 2022.

[19] S. El-Sappagh, F. Ali, T. Abuhmed, J. Singh, and J. M. Alonso, "Automatic detection of alzheimer's disease progression: An efficient information fusion approach with heterogeneous ensemble classifiers," *Neurocomputing*, vol. 512, pp. 203–224, 2022.

[20] S. El-Sappagh, J. M. Alonso, S. M. R. Islam, A. M. Sultan, and K. S. Kwak, "A multilayer multimodal detection and prediction model based on explainable artificial intelligence for alzheimer's disease," *Scientific Reports*, vol. 11, no. 1, p. 2660, 2021.

[21] K. D. Miltiadous, N. Tzimourta, M. G. Giannakeas, T. Tsipouras, P. Afrantou, P. Ioannidis, and A. T. Tzallas, "Alzheimer's disease and frontotemporal dementia: A robust classification method of eeg signals and a comparison of validation methods," *Diagnostics*, vol. 11, no. 8, p. 1437, 2021.

[22] P. Scheltens *et al.*, "Alzheimer's disease," *The Lancet*, vol. 397, no. 10284, pp. 1577–1590, 2021.

[23] T. D. Bird, "Alzheimer disease overview," in *GeneReviews®*, updated 2018 dec 20 ed., M. P. Adam, J. Feldman, G. M. Mirzaa *et al.*, Eds. Seattle (WA): University of Washington, Seattle, 1998, 1993–2025.

[24] H. Braak, U. Rüb, C. Schultz, and K. D. Tredici, "Vulnerability of cortical neurons to Alzheimer's and Parkinson's diseases," *Journal of Alzheimer's Disease*, vol. 9, no. Suppl, pp. 35–44, 2006.

[25] J. D. Warren, J. D. Rohrer, and M. N. Rossor, "Frontotemporal dementia," *BMJ*, vol. 347, p. f4827, 2013.

[26] D. Kim and K. Kim, "Detection of early stage alzheimer's disease using EEG relative power with deep neural network," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2018, pp. 352–355.

[27] M. Amini, M. Pedram, A. Moradi, and M. Ouchani, "Diagnosis of alzheimer's disease by time-dependent power spectrum descriptors and convolutional neural network using eeg signal," *Computational and Mathematical Methods in Medicine*, vol. 2021, p. 5511922, 2021, published 2021 Apr 23. [Online]. Available: https://doi.org/10.1155/2021/5511922

[28] F. Lazzarotto, "Analisi non lineare del segnale eeg: l'esponente di hurst, l'entropia approssimativa e la dimensione frattale," Università degli Studi di Padova, 2024, relatore: Prof.ssa Alessandra Bertoldo, Matricola 2032639.

[29] P. M. Rossini, S. Rossi, C. Babiloni, and J. Polich, "Clinical neurophysiology of aging brain: From normal aging to neurodegeneration," *Progress in Neurobiology*, vol. 83, no. 6, pp. 375–400, 2007.

[30] C. Huang, L. Wahlund, T. Dierks, P. Julin, B. Winblad, and V. Jelic, "Discrimination of alzheimer's disease and mild cognitive impairment by equivalent eeg sources: a cross-sectional and longitudinal study," *Clinical Neurophysiology*, vol. 111, no. 11, pp. 1961–1967, 2000.

[31] R. Nardone *et al.*, "Usefulness of EEG techniques in distinguishing frontotemporal dementia from alzheimer's disease and other dementias," *Disease Markers*, vol. 2018, p. 6581490, 2018.

[32] L. Bonanni, A. Thomas, P. Tiraboschi, B. Perfetti, S. Varanese, and M. Onofrj, "Eeg comparisons in early alzheimer's disease, dementia with lewy bodies and parkinson's disease with dementia patients with a 2-year follow-up," *Brain*, vol. 131, no. 3, pp. 690–705, 2008, epub 2008 Jan 17.

[33] M. F. Folstein, S. E. Folstein, and P. R. McHugh, ""mini-mental state". a practical method for grading the cognitive state of patients for the clinician," *Journal of Psychiatric Research*, vol. 12, no. 3, pp. 189–198, 1975.

[34] Z. S. Nasreddine *et al.*, "The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment," *Journal of the American Geriatrics Society*, vol. 53, no. 4, pp. 695–699, 2005.

[35] N. Ciesielska, R. Sokołowski, E. Mazur, M. Podhorecka, A. Polak-Szabela, and K. Kędziora-Kornatowska, "Is the montreal cognitive assessment (moca) test better suited than the mini-mental state examination (mmse) in mild cognitive impairment (mci) detection among people aged over 60? meta-analysis," *Psychiatria Polska*, vol. 50, no. 5, pp. 1039–1052, 2016.

[36] T. Torralva, M. Roca, E. Gleichgerrcht, P. López, and F. Manes, "Ineco frontal screening (ifs): a brief, sensitive, and specific tool to assess executive functions in dementia," *Journal of the International Neuropsychological Society*, vol. 15, no. 5, pp. 777–786, 2009.

[37] B. Dubois, A. Slachevsky, I. Litvan, and B. Pillon, "The fab: a frontal assessment battery at bedside," *Neurology*, vol. 55, no. 11, pp. 1621–1626, 2000.

[38] G. S. A. Siqueira, P. M. S. Hagemann, D. S. Coelho, F. H. D. Santos, and P. H. F. Bertolucci, "Can moca and mmse be interchangeable cognitive screening tools? a systematic review," *The Gerontologist*, vol. 59, no. 6, pp. e743–e763, 2019.

[39] N. Ciesielska, R. Sokołowski, E. Mazur, M. Podhorecka, A. Polak-Szabela, and K. Kędziora-Kornatowska, "Is the montreal cognitive assessment (moca) test better suited than the mini-mental state examination (mmse) in mild cognitive impairment (mci) detection among people aged over 60? meta-analysis," *Psychiatria Polska*, vol. 50, no. 5, pp. 1039–1052, 2016.

[40] M. Delussu, "Diagnosi cognitiva nella malattia di parkinson: uno studio multicentrico sui subitem di mmse e moca," Laurea Magistrale Thesis, Università degli studi di Padova, 2023. [Online]. Available: https://hdl.handle.net/20.500.12608/79609

[41] S. Osowski, K. Siwek, and T. Markiewicz, "Mlp and svm networks-a comparative study," in *Proceedings of the 6th Nordic Signal Processing Symposium, 2004. NORSIG 2004.* IEEE, 2004, pp. 37–40.

[42] W. Xu, Y.-L. Fu, and D. Zhu, "Resnet and its application to medical image processing: Research progress and challenges," *Computer Methods and Programs in Biomedicine*, vol. 240, p. 107660, 2023.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[44] P. Prado *et al.*, "The brainlat project, a multimodal neuroimaging dataset of neurodegeneration from underrepresented backgrounds," *Scientific Data*, vol. 10, no. 1, p. 889, Dec 2023.

[45] scikit-learn developers, "scikit-learn: Parametergrid," https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.ParameterGrid.html, 2025, accessed: YYYY-MM-DD.

[46] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 785–794.

[47] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *2010 20th International Conference on Pattern Recognition*, 2010, pp. 3121–3124.

[48] R. Cassani, M. Estarellas, R. San-Martin, F. J. Fraga, and T. H. Falk, "Systematic review on resting-state EEG for Alzheimer's Disease diagnosis and progression assessment," *Disease Markers*, vol. 2018, p. 5174815, Oct. 2018.

[49] E. Costanzo, S. L. Sensi, and A. Trovò, "Exploring easily accessible neurophysiological biomarkers for alzheimer's disease," *Alzheimer's Research & Therapy*, vol. 16, no. 1, p. 91, 2024. [Online]. Available: https://alzres.biomedcentral.com/articles/10.1186/s13195-024-01607-4

[50] J. J. Bird, D. R. Faria, L. J. Manso, P. P. S. Ayrosa, and A. Ekárt, "A study on cnn image classification of eeg signals represented in 2d and 3d," *Journal of Neural Engineering*, vol. 18, no. 2, p. 026005, Feb. 2021.

[51] L. Vortmann, I. Behnke, K. Kühnlenz, and D. Ulrich, "Multimodal EEG and eye tracking feature fusion," *Frontiers in Computer Science*, vol. 4, p. 780580, 2022.

[52] A. Kline, A. Borthakur, H. Suresh, S. Ghosh, and F. Doshi-Velez, "Multimodal machine learning in precision health: A scoping review," *Journal of the American Medical Informatics Association*, vol. 29, no. 10, pp. 1749–1759, 2022. [Online]. Available: https://doi.org/10.1093/jamia/ocac102

A Appendix

A.1 Repository and File Reference

The core materials supporting this study can be accessed at the following repository: `https://github.com/MedMaxLab/tesirachele`.
Within the repository:

- The interactive 3D Topomap can be explored by downloading the file *InteractiveTopomap3D.html*, , available in the `files` folder.
- The Excel file of channel locations, kindly provided by Dr. Antonio, can be accessed as *ChanLocsAntonio.xlsx*, also available in the `files` folder.

A.2 Class Deep Learning Models

Listing A.1: Unimodal, Late- and Middle-Fusion models (PyTorch)

```
# ---------------------------------
# CD UNIMODAL Class
# ---------------------------------
class MLP(nn.Module):
    def __init__(self, input_size, num_classes=3, hidden_size1=64, activ1=nn.GELU(),
                 dropout1=0.0, hidden_size2=32, activ2=nn.GELU(), dropout2=0.3):
        super().__init__()
        self.fc1 = nn.Linear(input_size, hidden_size1)
        self.act1 = activ1
        self.drop1 = nn.Dropout(dropout1)
        self.fc2 = nn.Linear(hidden_size1, hidden_size2)
        self.act2 = activ2
        self.drop2 = nn.Dropout(dropout2)
        self.classifier = nn.Linear(hidden_size2, num_classes)

    def forward(self, x, return_features=False):
```

```
        h1 = self.drop1(self.act1(self.fc1(x)))
        h2 = self.drop2(self.act2(self.fc2(h1)))
        logits = self.classifier(h2)
        if return_features:
            return {"x": x, "fc1": h1, "fc2": h2, "logits": logits}
        return logits


# ----------------------------------------
#EEG UMIMODAL Class
# ----------------------------------------
class BasicBlock_2D(nn.Module):
    def __init__(self, in_channels, out_channels, stride=1):
        super().__init__()
        self.conv1 = nn.Conv2d(in_channels, out_channels, 3, stride=stride, padding=1, bias
            =False)
        self.bn1 = nn.BatchNorm2d(out_channels)
        self.act = nn.ReLU(inplace=True)
        self.conv2 = nn.Conv2d(out_channels, out_channels, 3, padding=1, bias=False)
        self.bn2 = nn.BatchNorm2d(out_channels)
        self.down = nn.Identity()
        if in_channels != out_channels or stride != 1:
            self.down = nn.Conv2d(in_channels, out_channels, 1, stride=stride, bias=False)

    def forward(self, x):
        identity = self.down(x)
        out = self.act(self.bn1(self.conv1(x)))
        out = self.bn2(self.conv2(out))
        out += identity
        return self.act(out)

class BasicBlock_3D(nn.Module):
    def __init__(self, in_channels, out_channels, stride=1):
        super().__init__()
        self.conv1 = nn.Conv3d(in_channels, out_channels, 3, stride=stride, padding=1, bias
            =False)
        self.bn1   = nn.BatchNorm3d(out_channels)
        self.act   = nn.ReLU(inplace=True)
        self.conv2 = nn.Conv3d(out_channels, out_channels, 3, padding=1, bias=False)
        self.bn2   = nn.BatchNorm3d(out_channels)
        self.down  = nn.Identity()
        if in_channels != out_channels or stride != 1:
            self.down = nn.Conv3d(in_channels, out_channels, 1, stride=stride, bias=False)

    def forward(self, x):
        identity = self.down(x)
        out = self.act(self.bn1(self.conv1(x)))
        out = self.bn2(self.conv2(out))
        out += identity
        return self.act(out)
```

```
65
66  class ResNet2-2D(nn.Module):
67      def __init__(self, in_channels=11, num_classes=3):
68          super().__init__()
69          self.block1 = BasicBlock_2D(in_channels, 16, stride=1)
70          self.pool1  = nn.MaxPool2d(2)
71          self.block2 = BasicBlock_2D(16, 32, stride=1)
72          self.gap    = nn.AdaptiveAvgPool2d(1)
73          self.fc     = nn.Linear(32, num_classes)
74
75      def forward(self, x, return_features=False):
76          b1 = self.block1(x)
77          p1 = self.pool1(b1)
78          b2 = self.block2(p1)
79          g  = self.gap(b2).flatten(1)
80          logits = self.fc(g)
81          if return_features:
82              return {"gap": g, "logits": logits}
83          return logits
84
85  class ResNet2-3D(nn.Module):
86      def __init__(self, in_channels=1, num_classes=3):
87          super().__init__()
88          self.block1 = BasicBlock_3D(in_channels, 16, stride=1)
89          self.pool1  = nn.MaxPool3d(2)
90          self.block2 = BasicBlock_3D(16, 32, stride=1)
91          self.gap    = nn.AdaptiveAvgPool3d((1, 1, 1))
92          self.fc     = nn.Linear(32, num_classes)
93
94      def forward(self, x, return_features=False):
95          x = x.unsqueeze(1)
96          b1 = self.block1(x)
97          p1 = self.pool1(b1)
98          b2 = self.block2(p1)
99          g  = self.gap(b2).flatten(1)
100         logits = self.fc(g)
101         if return_features:
102             return {"gap": g, "logits": logits}
103         return logits
104
105 class ResNet2-2D_reference(nn.Module):
106     def __init__(self, in_channels=11, num_classes=3):
107         super().__init__()
108         reset_everything(seed=SEED)
109         self.conv1 = nn.Conv2d(in_channels, 16, kernel_size=3, stride=1, padding=1, bias=
                False)
110         self.bn1 = nn.BatchNorm2d(16)
111         self.relu = nn.ReLU(inplace=True)
112         self.maxpool = nn.MaxPool2d(kernel_size=3, stride=2, padding=1)
113         self.layer1 = BasicBlock_2D(16, 32, stride=2)
```

```
        self.layer2 = BasicBlock_2D(32, 32, stride=1)
        self.global_pool = nn.AdaptiveAvgPool2d((1, 1))
        reset_everything(seed=SEED)
        self.fc = nn.Linear(32, num_classes)

    def forward(self, x):
        x = self.relu(self.bn1(self.conv1(x)))
        x = self.maxpool(x)
        x = self.layer1(x)
        x = self.layer2(x)
        x = self.global_pool(x)
        x = x.view(x.size(0), -1)
        return self.fc(x)

class ResNet2-3D_reference(nn.Module):
    def __init__(self, in_channels=1, num_classes=3):
        super().__init__()
        reset_everything(seed=SEED)
        self.conv1 = nn.Conv3d(in_channels, 16, kernel_size=3, stride=1, padding=1, bias=
            False)
        self.bn1 = nn.BatchNorm3d(16)
        self.relu = nn.ReLU(inplace=True)
        self.maxpool = nn.MaxPool3d(kernel_size=3, stride=2, padding=1)
        self.layer1 = BasicBlock_3D(16, 32, stride=2)
        self.layer2 = BasicBlock_3D(32, 32, stride=1)
        self.global_pool = nn.AdaptiveAvgPool3d((1, 1, 1))
        reset_everything(seed=SEED)
        self.fc = nn.Linear(32, num_classes)

    def forward(self, x):
        x = x.unsqueeze(1)
        x = self.relu(self.bn1(self.conv1(x)))
        x = self.maxpool(x)
        x = self.layer1(x)
        x = self.layer2(x)
        x = self.global_pool(x)
        x = x.view(x.size(0), -1)
        return self.fc(x)


# ----------------------------------------
# MULTIMODAL - Late Meta Fusion Class
# ----------------------------------------
class GatedFusion(nn.Module):
    def __init__(self, num_classes=3, hidden_size=32, activation=nn.GELU(), dropout=0.0,
        SEED=123):
        super(GatedFusion, self).__init__()
        reset_everything(seed=SEED)
        self.gate = nn.Sequential(
            nn.Linear(num_classes*2, hidden_size),
```

```
            nn.Dropout(dropout),
            activation,
            nn.Linear(hidden_size, 1),
            nn.Sigmoid()
        )

    def forward(self, mlp_proba, cnn_proba):
        x = torch.cat([mlp_proba, cnn_proba], dim=1)
        alpha = self.gate(x)        # shape: (batch, 1)
        fusion = alpha * mlp_proba + (1 - alpha) * cnn_proba
        return fusion

class AttentionFusion(nn.Module):
    def __init__(self, num_classes=3, hidden_size=32, dropout=0.0, activation=nn.GELU(),
        SEED=123):
        super(AttentionFusion, self).__init__()
        reset_everything(seed=SEED)
        self.attn_fc = nn.Sequential(
            nn.Linear(num_classes, hidden_size),
            nn.Dropout(dropout),
            activation,
            nn.Linear(hidden_size, 1)
        )
        self.num_classes = num_classes

    def forward(self, clin_proba, eeg_proba):
        x = torch.stack([clin_proba, eeg_proba], dim=1)
        attn_scores = self.attn_fc(x)
        attn_weights = F.softmax(attn_scores, dim=1)
        fused = torch.sum(attn_weights * x, dim=1)
        return fused


# --------------------------------
# MULTIMODAL - Middle Fusion Class
# --------------------------------
class MiddleFusionModel(nn.Module):
    def __init__(self, mlp_model: nn.Module, cnn_model: nn.Module,
                 mlp_layer: str = "fc2", cnn_layer: str = "gap",
                 num_classes: int = 3, head_hidden: int = 64, head_dropout: float = 0.4):
        super().__init__()
        assert mlp_layer in {"x", "fc1", "fc2", "logits"}
        assert cnn_layer in {"gap", "logits"}

        self.mlp = mlp_model
        self.cnn = cnn_model
        self.mlp_layer = mlp_layer
        self.cnn_layer = cnn_layer

        # ---- MLP feature dimensions ----
```

```
        mlp_dim_map = {
            "x":      mlp_model.fc1.in_features,
            "fc1":    mlp_model.fc1.out_features,
            "fc2":    mlp_model.fc2.out_features,
            "logits": mlp_model.classifier.out_features,
        }
        mlp_dim = mlp_dim_map[mlp_layer]

        # ---- CNN feature dimensions ----
        cnn_dim_map = {
            "gap":    cnn_model.fc.in_features,
            "logits": cnn_model.fc.out_features,
        }
        cnn_dim = cnn_dim_map[cnn_layer]

        fuse_in_dim = mlp_dim + cnn_dim

        # ---- classification head ----
            self.head = nn.Sequential(
            nn.Linear(fuse_in_dim, head_hidden),
            nn.GELU(),
            nn.BatchNorm1d(head_hidden),
            nn.Dropout(head_dropout),
            nn.Linear(head_hidden, num_classes),
        )

    def _pick_mlp_feat(self, feats_dict):
        return feats_dict[self.mlp_layer]

    def _pick_cnn_feat(self, feats_dict):
        return feats_dict[self.cnn_layer]

    def forward(self, x_clin, x_eeg, return_aux=False):
        mlp_feats = self.mlp(x_clin, return_features=True)
        cnn_feats = self.cnn(x_eeg,  return_features=True)
        a = self._pick_mlp_feat(mlp_feats)   # [B, D_mlp]
        b = self._pick_cnn_feat(cnn_feats)   # [B, D_cnn]
        fused = torch.cat([a, b], dim=1)     # [B, D_mlp + D_cnn]
        logits = self.head(fused)
        return (logits, {}) if return_aux else logits
```

A.3 Parameter and Hyperparameter Grids

Listing A.2: Parameter and Hyperparameter grid

```

# ---------------------------------
# CD UNIMODAL Grid
# ---------------------------------
param_grid_MLP = {
```

```
    'hidden_size1': [32, 64],
    'hidden_size2': [16, 32],
    'dropout1': [0.2, 0.8],
    'dropout2': [0.2, 0.8],
    'activ1': ['gelu'],
    'activ2': ['gelu']
}
param_grid_XG = {
    'max_depth': [2, 5],
    'min_child_weight': [1, 5],
    'subsample': [0.5, 1.0],
    'colsample_bytree': [0.2, 0.5],
    'learning_rate': [0.05, 0.1, 0.2],
    'gamma': [0],
    'reg_alpha': [0],
    'reg_lambda': [1],
    'n_estimators': [100],
    'objective': ['multi:softprob'],
    'num_class': [3],
    'random_state': [SEED],
    'eval_metric': ['mlogloss'],
    'tree_method': ['hist']
}
hyperparam_grid = {
    'test_size': [0.2],
    'normX': ['minmax'],
    'weight': ['exponential', 'inverse'],
    'SEED': [SEED],
    'splits': [100],
    'inner_val_size': [0.2]
}


# ----------------------------------------
#EEG UMIMODAL Grid
# ----------------------------------------
param_grid_CNN = {'param1': [0]}

hyperparam_grid = {
    'test_size': [0.2],
    'weight': ['exponential', 'inverse'],
    'SEED': [SEED],
    'splits': [100],
    'inner_val_size': [0.2]
}


# ----------------------------------------
# MULTIMODAL - Late Sepatate Average Grid
# ----------------------------------------
```

```
param_grid = {}
hyperparam_grid = {}


# ---------------------------------------
# MULTIMODAL - Late Meta Fusion Grid
# ---------------------------------------
param_grid_XG = {
    'max_depth': [2, 5],
    'min_child_weight': [1, 5],
    'subsample': [0.5, 1.0],
    'colsample_bytree': [0.2, 0.5],
    'learning_rate': [0.05, 0.1, 0.2],
    'gamma': [0],
    'reg_alpha': [0],
    'reg_lambda': [1],
    'n_estimators': [100],
    'objective': ['multi:softprob'],
    'num_class': [3],
    'random_state': [SEED],
    'eval_metric': ['mlogloss'],
    'tree_method': ['hist']
}
param_grid_Gated_Attention = {
    'hidden_size': [16, 32],
    'activation': [nn.GELU()],
    'dropout': [0.2, 0.8]
}
hyperparam_grid = {
    'test_size': [0.2],
    'weight': ['exponential', 'inverse'],
    'SEED': [SEED],
    'inner_val_size': [0.2],
    'mode': ['log'],
    'splits': [100]
}


# --------------------------------
# MULTIMODAL - Middle Fusion Grid
# --------------------------------
param_grid = {}
hyperparam_grid = {
    'test_size': [0.2],
    'normX': ['minmax'],
    'weight': ['exponential', 'inverse'],
    'SEED': [SEED],
    'inner_val_size': [0.2],
    'splits': [100]
}
```

A.4 Summary of Model Performance by Strategy

Strategy	Model	Balanced Accuracy (%)
Unimodal	XGBoost (CD)	61.1 (50.0–66.7)
	MLP (CD)	63.9 (55.6–72.2)
	ResNet2-2D (EEG)	72.2 (61.1–77.8)
	ResNet2-3D (EEG)	**77.8 (66.7–83.3)**
	ResNet2-2D_reference (EEG)	66.7 (55.6–77.8)
	ResNet2-3D_reference (EEG)	72.2 (66.7–77.8)
Late Separate Average Fusion	XGBoost + ResNet2-2D (Average)	66.7 (61.1–77.8)
	XGBoost + ResNet2-2D (Log-Average)	72.2 (66.7–77.8)
	XGBoost + ResNet2-2D (Entropy)	66.7 (61.1–77.8)
	XGBoost + ResNet2-3D (Average)	77.8 (66.7–84.7)
	XGBoost + ResNet2-3D (Log-Average)	77.8 (72.2–88.9)
	XGBoost + ResNet2-3D (Entropy)	77.8 (66.7–83.3)
	MLP + ResNet2-2D (Average)	77.8 (72.2–83.3)
	MLP + ResNet2-2D (Log-Average)	77.8 (72.2–83.3)
	MLP + ResNet2-2D (Entropy)	77.8 (66.7–83.3)
	MLP + ResNet2-3D (Average)	83.3 (77.8–88.9)
	MLP + ResNet2-3D (Log-Average)	**83.3 (77.8–88.9)**
	MLP + ResNet2-3D (Entropy)	83.3 (77.8–88.9)
Late Meta Fusion	XGBoost+ResNet2-2D (Meta-XGBoost)	55.6 (44.4–61.1)
	XGBoost+ResNet2-2D (Meta-Gated)	72.2 (66.7–77.8)
	XGBoost+ResNet2-2D (Meta-Attention)	72.2 (66.7–77.8)
	XGBoost+ResNet2-3D (Meta-XGBoost)	55.6 (44.4–61.1)
	XGBoost+ResNet2-3D (Meta-Gated)	77.8 (70.8–88.39)
	XGBoost+ResNet2-3D (Meta-Attention)	77.8 (72.2–88.9)
	MLP+ResNet2-2D (Meta-XGBoost)	33.3 (22.2–38.9)
	MLP+ResNet2-2D (Meta-Gated)	77.8 (72.2–83.3)
	MLP+ResNet2-2D (Meta-Attention)	83.3 (72.2–88.9)
	MLP+ResNet2-3D (Meta-XGBoost)	33.3 (22.2–38.9)
	MLP+ResNet2-3D (Meta-Gated)	83.3 (77.8–88.9)
	MLP+ResNet2-3D (Meta-Attention)	**83.3 (77.8–88.9)**
Middle Fusion	MLP + ResNet2-2D (X+GAP)	83.3 (77.8–88.9)
	MLP + ResNet2-2D (X+Logits)	83.3 (77.8–88.9)
	MLP + ResNet2-2D (FC1+GAP)	88.9 (81.9–90.3)
	MLP + ResNet2-2D (FC1+Logits)	88.9 (83.3–94.4)
	MLP + ResNet2-2D (FC2+GAP)	77.8 (66.7-83.3))
	MLP + ResNet2-2D (FC2+Logits)	77.8 (72.2-88.9)
	MLP + ResNet2-2D (Logits+GAP)	72.2 (66.7-83.3)
	MLP + ResNet2-2D (Logits+Logits)	77.8 (66.7-83.3)
	MLP + ResNet2-3D (X+GAP)	88.9 (77.8–94.4)
	MLP + ResNet2-3D (X+Logits)	88.9 (77.8–94.4)
	MLP + ResNet2-3D (FC1+GAP)	88.9 (81.9–90.3)
	MLP + ResNet2-3D (FC1+Logits)	88.9 (77.8–94.4)
	MLP + ResNet2-3D (FC2+GAP)	83.3 (77.8–88.9)
	MLP + ResNet2-3D (FC2+Logits)	83.3 (77.8–88.9)
	MLP + ResNet2-3D (Logits+GAP)	77.8 (66.7-83.3)
	MLP + ResNet2-3D (Logits+Logits)	83.3 (72.2-84.7)

Table A.1: Summary of Model Performance by Strategy
Balanced accuracy (%) of Unimodal and Multimodal models across different fusion strategies. Values indicate median performance with ranges in parentheses.