Braindecode Library for Dementia EEG Classification

Project for Natural Interaction and Affective Computing

Bianchessi Mattia 28824A

1 Abstract

Dementia represents a growing global health challenge, with over 55 million affected individuals and nearly 10 million new cases each year. Among its subtypes, Alzheimer's disease (AD) and frontotemporal dementia (FTD) are the most prevalent, yet their clinical differentiation remains difficult due to overlapping symptoms and heterogeneous progression. Electroencephalography (EEG) offers a non-invasive, widely accessible, and cost-effective alternative to neuroimaging for dementia assessment, capturing characteristic alterations in neural oscillations and connectivity. Recent advances in deep learning enable automatic feature extraction from raw EEG signals, potentially surpassing traditional approaches that rely on handcrafted features. In this work, we systematically evaluate two state-of-the-art architectures, EEGNetv4 and Deep4Net, implemented through the open-source Braindecode library. Using a dataset of subjects (AD, FTD, and controls), we applied a Leave-One-Subject-Out (LOSO) validation scheme to assess generalization performance at both segment and subject levels. Results indicate moderate accuracy, with models performing best in distinguishing pathological cases from controls, but with limited effectiveness in differentiating dementia subtypes. These findings highlight both the promise and current limitations of deep learning for EEG-based dementia classification, and suggest directions for future research toward clinically viable and accessible diagnostic tools.

2 Introduction

Dementia is a major global health challenge, affecting over 55 million people worldwide with nearly 10 million new cases each year. Alzheimer's disease (AD) and frontotemporal dementia (FTD) are the most common subtypes, yet their clinical differentiation remains difficult due to overlapping symptoms and progressive decline. Early and accurate diagnosis is crucial for patient care and treatment planning, but current methods often rely on costly neuroimaging and specialized expertise that are not always accessible, especially in resource-limited

settings. Electroencephalography (EEG) represents a promising alternative, being non-invasive, widely available, and capable of detecting dementia-related alterations in neural oscillations and connectivity, such as increased theta activity and disrupted functional networks.

Recent advances in deep learning have enabled automatic extraction of informative features from raw EEG signals, surpassing traditional feature-based methods and potentially identifying subtle biomarkers. This study systematically evaluates two models using the Braindecode library, employing a Leave-One-Subject-Out validation. By integrating robust methodology with open-source tools, this work contributes benchmarks for EEG-based dementia classification and highlights opportunities for clinically viable, accessible diagnostic solutions.

The source code is available on github at: [1].

3 Dataset

The dataset used is A Dataset of Scalp EEG Recordings of Alzheimer's Disease, Frontotemporal Dementia and Healthy Subjects from Routine EEG ([2]). The dataset contains the EEG recording from 88 subject divided in three groups depend on clinic situation.

The disease are Alzheimer, frontotemporal dementia or control. The signal are collected from 21 electrodes. Each group has own mean duration from 13.5 min for the Alzheimer, 12 min for the frontotemporal dementia and 13.8 min for the control. The total duratin 8022 minutes.

The dataset is organized according to the BIDS specification. General metadata are stored in JSON and TSV files. Each participant has an individual folder containing raw EEG data (`.set` format) alongside metadata describing electrode positions, acquisition parameters, and recording devices. In addition, a `derivatives` folder provides preprocessed EEG recordings with the same structure as the raw data. A full description of the dataset organization and file structure is available in [2].

3.1 Preprocessing

The EEG data underwent a preprocessing pipeline designed to remove artifacts while preserving signal integrity. Signals were first bandpass filtered with a 0.5–45 Hz Butterworth filter and re-referenced to the common average. Artifact Subspace Reconstruction (ASR) was then applied to automatically remove high-amplitude segments, using a conservative threshold of 17 standard deviations computed on 0.5-second windows. Independent Component Analysis (ICA) was performed with the RunICA algorithm, decomposing the data into 19 components. Noise-related components were subsequently identified and removed using ICLabel from the EEGLAB toolbox. This preprocessing procedure improved the quality of the EEG signals for subsequent analysis.

4 Preliminary Exploratory Analysis

The dataset included 88 patients and five clinical-demographic variables, providing a well-structured sample for the study of dementia. The completeness of the data ensured high quality and reliability of the statistical analyses.

The mean age of participants was 66.2 years (SD = 7.4; range: 44–79). The median age of 67 years and the interquartile distribution (Q1 = 61.8; Q2 = 67; Q3 = 71) indicate a relatively symmetric distribution and a good representation of the age groups most at risk for developing neurodegenerative disorders Figure 1. The sex distribution was balanced, with an equal number of male and female participants.

Global cognitive status, assessed using the Mini-Mental State Examination (MMSE) Figure 2b, showed a mean score of 22.9 points (SD = 6.2; range: 4–30), covering the entire spectrum of cognitive impairment. The median score of 22 and the high intra-sample variability reflect the presence of clinically distinct subgroups. The relationship between age and global cognitive status (MMSE) was assessed using Pearson's correlation coefficient. Results showed a weak positive correlation (r = 0.158, p = 0.141), indicating no statistically significant association between age and MMSE scores in this sample.

Descriptive statistics of MMSE scores by diagnostic group revealed clear differences (Table 1): patients with Alzheimer's disease (group A, n = 36) had a mean MMSE of 17.8 (SD = 4.5), healthy controls (group C, n = 29) scored at ceiling with 30.0 (SD = 0.0), and patients with frontotemporal dementia (group F, n = 23) had a mean score of 22.2 (SD = 2.6). ANOVA confirmed significant differences between groups (F(2,85) = 119.74, p ¡ 0.001), supporting the presence of clinically distinct cognitive profiles.

Group	**N**	**Mean**	**SD**	**Min**	**Max**
A (Alzheimer)	36	17.75	4.50	4	23
C (Controls)	29	30.00	0.00	30	30
F (Frontotemporal)	23	22.17	2.64	18	27

Table 1: Descriptive statistics of MMSE scores by diagnostic group.

The figure Figure 2a shows multiple plots, arranged from top-left to bottom-right: Age distribution, Gender distribution, Diagnostic group distribution, Age distribution by group, Age distribution by gender, and Gender composition by group.

According to established clinical thresholds, MMSE values were interpreted as follows: > 26 = normal cognitive status, 20–26 = mild impairment, and < 20 = moderate-to-severe impairment.

The dataset demonstrated high quality and validity, as reflected by the absence of missing values, the consistency of MMSE scores within the theoretical range (0–30), and the representative distribution of age and sex.

The diagnostic group distribution was as follows: Alzheimer's disease (A): 36 patients (40.9%), healthy controls (C): 29 patients (33.0%), and frontotemporal

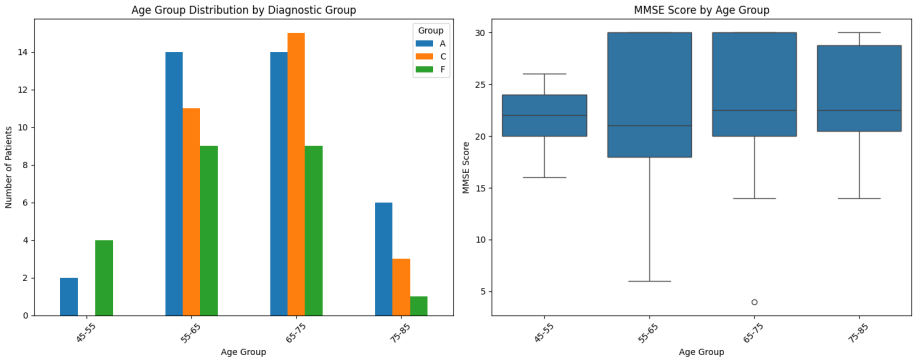
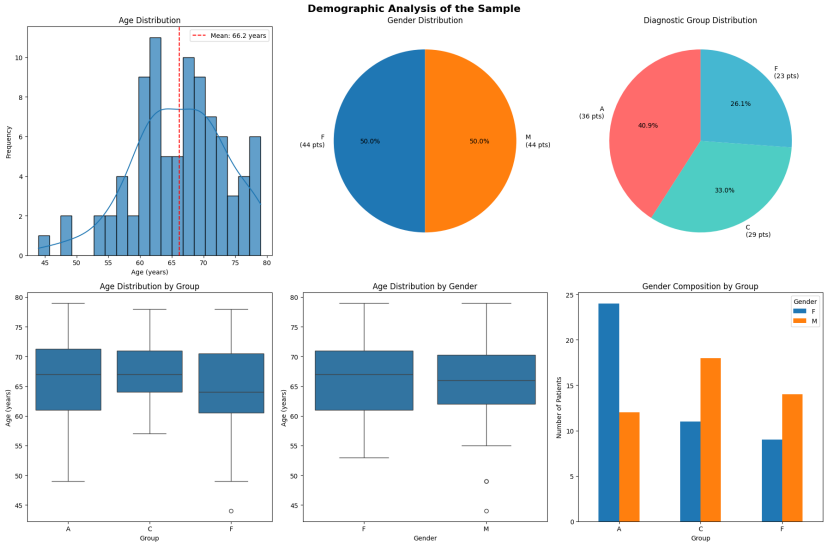


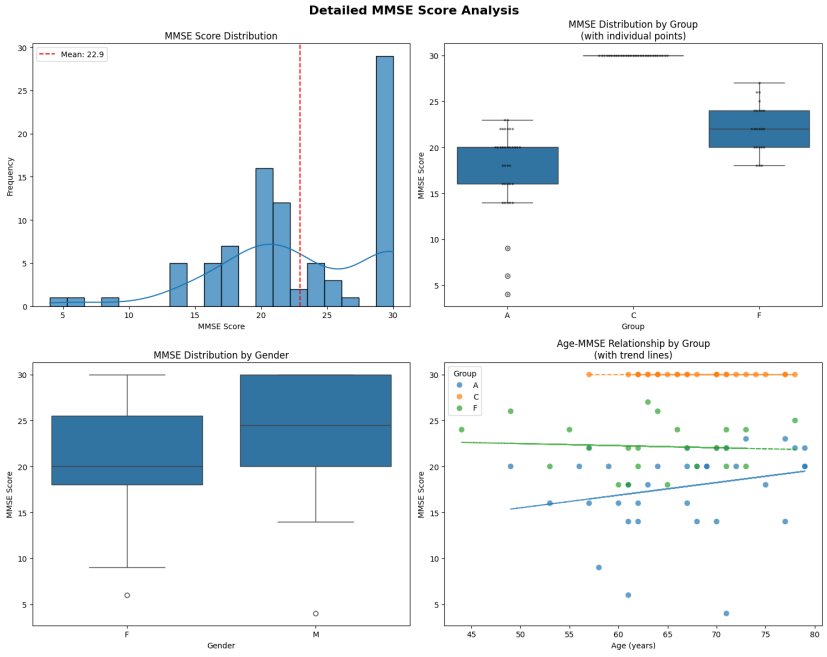
Figure 1: Diagnostic by age.

dementia (F): 23 patients (26.1%).

Age distribution approximated normality, with a slight left skew. Sex distribution was balanced (50% male, 50% female). One outlier was observed in the FTD group. Boxplots of age by diagnostic group showed overlapping distributions and similar medians, suggesting that age was not a major confounding factor in this sample 2.



(a) Demographic characteristics of the study sample, including age distribution, sex balance, and global cognitive status (MMSE).



(b) Detailed MMSE Score Analysis across different diagnostic groups.

Figure 2: Comprehensive demographic and cognitive assessment of the study population.

Age Group / Diagnostic Group	N (%)	Age, mean ± SD	MMSE, mean ± SD	Male (%)	Female (%)
Age Groups					
45-55	6 (6.8)	50.5 ± 3.2	21.7 ± 3.7	4 (66.7)	2 (33.3)
55-65	34 (38.6)	60.8 ± 2.8	22.1 ± 6.8	17 (50.0)	17 (50.0)
65-75	38 (43.2)	69.1 ± 2.8	23.8 ± 6.1	18 (47.4)	20 (52.6)
75-85	10 (11.4)	78.2 ± 2.0	23.6 ± 5.3	5 (50.0)	5 (50.0)
Diagnostic Groups					
Alzheimer's disease (A)	36 (40.9)	66.5 ± 7.1	17.8 ± 4.5	–	–
Healthy controls (C)	29 (33.0)	65.8 ± 6.9	30.0 ± 0.0	–	–
Frontotemporal dementia (F)	23 (26.1)	66.7 ± 7.8	22.2 ± 2.6	–	–
Total	88 (100)	66.2 ± 7.4	22.9 ± 6.2	44 (50.0)	44 (50.0)

Table 2: Summary of demographic and clinical characteristics of the study cohort. The table reports the number and percentage of participants, mean age with standard deviation, mean MMSE scores with standard deviation, and sex distribution across both age groups and diagnostic categories.

5 Braindecode

Braindecode is an open-source Python library designed for decoding raw electrophysiological brain signals using state-of-the-art deep learning methods. It provides an integrated framework that includes standardized dataset loaders, preprocessing pipelines, and visualization utilities, alongside implementations of established neural architectures and data augmentation techniques. The toolbox supports analysis across multiple recording modalities, including electroencephalography (EEG), electrocorticography (ECoG), and magnetoencephalography (MEG). Braindecode aims to bridge the gap between neuroscience and machine learning, serving both neuroscientists seeking to apply modern deep learning approaches to brain data and deep learning researchers interested in advancing methods for neurophysiological signal analysis [3].

5.1 Network Architectures Used

The library contains numerous models, each of which is documented with a specific purpose. The EEGNetv4 and DeepNet models were used for this experiment.

EEGNet is a compact convolutional neural network architecture specifically designed for decoding EEG signals. Its design reflects the spatio-temporal structure of electrophysiological data, employing a sequence of operations that progressively disentangle temporal and spatial patterns. The network begins with temporal convolutions, which act as band-pass filters to capture frequency-specific activity. This is followed by depthwise spatial convolutions, which learn spatial filters across EEG channels, effectively extracting localized topographical patterns. Finally, separable pointwise convolutions combine these features while keeping the parameter count low [4] [5].

Deep4Net is a deep convolutional neural network architecture tailored for decoding EEG signals, particularly within brain–computer interface (BCI) applications. Its design follows a hierarchical feature extraction strategy across four convolutional layers. The initial layer applies temporal convolutions, effectively functioning as band-pass filters to capture oscillatory activity across different frequency bands. This is followed by a spatial convolutional layer, which learns spatial filters across electrode channels to isolate topographical patterns of brain activity. The subsequent two convolutional layers progressively build higher-level, abstract representations of the EEG, allowing the network to capture increasingly complex spatio-temporal dependencies [6] [7].

Deep4Net and EEGNet have two complementary design philosophies. Deep4Net, with its four-layer deep convolutional structure, emphasizes hierarchical feature extraction, enabling the capture of increasingly complex spatio-temporal patterns. This makes it powerful for modeling rich neural dynamics but also leads to large parameter counts, higher computational demands, and a greater risk of overfitting on limited data. In contrast, EEGNet adopts a parameter-efficient architecture through the use of temporal, depthwise spatial, and separable pointwise convolutions. This design drastically reduces model size while preserving

competitive performance, particularly on smaller datasets where regularization is crucial. However, this efficiency comes at the expense of representational capacity, limiting EEGNet's ability to model highly complex neural patterns that Deep4Net can exploit. Thus, Deep4Net is well-suited for scenarios with sufficient data and computational resources, whereas EEGNet offers a lightweight and generalizable alternative for resource-constrained or real-time applications.

Architecture	**Capacity**	**Parameters**	**Overfitting (LOSO)**	**Training time**
EEGNet v4	Medium	Low	Low	Fast
Deep4Net	High	High	Medium–High	Slower

Table 3: Comparison between EEGNetv4 and Deep4Net

6 Methods

To evaluate model generalization across subjects, we adopted a Leave-One-Subject-Out (LOSO) cross-validation scheme. In addition to this approach, the networks were trained in a binary manner by comparing the groups in order to obtain a binary classifier.

Training employed EEGNet v4 with cross-entropy loss and Adam optimizer (learning rate = 0.005), using 32-sample mini-batches for up to 200 epochs. Regularization included early stopping (patience = 25 epochs) and learning rate reduction (factor = 0.7 after 8 plateau epochs). Each LOSO fold reserved 30% of training data for internal validation. The table 4 shows the general configuration.

A comprehensive monitoring system tracked training and validation metrics (loss, accuracy) and learning rate evolution to identify optimal stopping points, detect overfitting, and visualize learning dynamics across LOSO folds. Model performance was evaluated using standard classification metrics—accuracy, precision, recall, and F1-score—along with confusion matrices, with results averaged across all cross-validation folds.

This evaluation protocol ensures that reported results reflect the ability of the model to generalize to unseen subjects, a crucial requirement for real-world BCI and clinical EEG applications.

While individual model optimization showed potential improvements up to 196 epochs (DeepNet) and 199 epochs (EEGNetv4), the computational constraints of LOSO cross-validation necessitated a more conservative approach. Training N=200 models for the full convergence period would require approximately 39,200-39,800 total epochs (196-199 $\times$ 200), representing a 4$\times$ increase in computational cost compared to the 50-epoch limit (10,000 total epochs).

Given the marginal performance gains observed after epoch 50 in preliminary analysis and the substantial computational overhead 3, the 50-epoch limit represents a pragmatic compromise between model performance and experimen-

Parametro	**EEGNetv4**	**Deep4Net**
Architettura	EEGNetv4	Deep4Net
Canali di Input	19	19
Lunghezza Temporale	1000	1000
Classi di Output	3 (A, C, F)	3 (A, C, F)
Final Conv Length	auto	auto
Iperparametri di Training		
Learning Rate	0.005	0.005
Batch Size	32	32
Max Epochs	50	200
Criterio di Loss	CrossEntropyLoss	CrossEntropyLoss
Train/Validation Split	70%/30%	70%/30%
Callbacks		
Early Stopping		
Patience	10	20
Threshold	0.001	0.001
Monitor Metric	valid_loss	valid_loss
Configurazione Dati		
Frequenza Campionamento	500 Hz	500 Hz
Canali Selezionati	C3, C4, Cz, F3, F4, F7, F8,Fp1, Fp2, Fz, O1, O2,P3, P4, Pz, T3, T4, T5, T6	
Lunghezza Segmento	1000 campioni (2s)	
Validazione		
Metodo	LOSO Cross-Validation	
Metriche	Accuracy, Precision,Recall, F1-score	

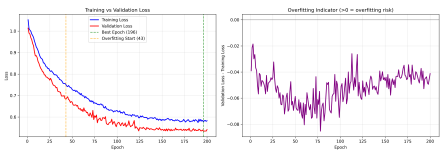
Table 4: Net configuration

tal feasibility. Additionally, the implemented early stopping mechanism with patience parameter typically halts training around epochs 50-60 when validation loss stabilization is detected, further supporting this choice. This approach aligns with established practices in cross-validation studies where computational efficiency must be balanced against optimal individual model performance.

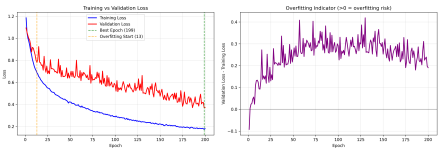
7 Results

The performance of both deep learning models was evaluated using Leave-One-Subject-Out (LOSO) cross-validation across four binary classification tasks. Each model's performance was assessed at two levels: individual EEG segments and subject-level predictions obtained through majority voting of segment classifications.

The table 5 present the comprehensive evaluation results for EEGNetv4 and Deep4Net. Both models demonstrated varying performance across the different classification tasks, with notable differences between segment-level and subject-



(a) EEGNetv4: best epoch 199, overfitting start at epoch 13.



(b) DeepNet: best epoch 196, overfitting start at epoch 43.

Figure 3: Training and validation loss curves over 200 epochs for both architectures, illustrating extended training potential beyond the 50-epoch LOSO limit and associated computational trade-offs.

level metrics. Segment-level performance represents the classification accuracy of individual 2-second EEG epochs, providing insight into the models' ability to detect discriminative patterns within short temporal windows. The reported accuracy values include standard deviations calculated across all individual subjects' segment accuracies, indicating the variability in model performance across the participant cohort. Subject-level performance was determined using majority voting among all segments belonging to each participant, representing a more clinically relevant evaluation approach. The high standard deviations observed in subject-level accuracy (approximately 0.49-0.50) reflect the binary nature of this metric, where each subject contributes either a perfect (1.0) or failed (0.0) classification. The Control vs. Pathological classification task yielded the highest performance for both models at the subject level, with EEGNetv4 achieving 61.3% accuracy and Deep4Net reaching 53.4%. This suggests that distinguishing between healthy controls and any form of dementia presents more discriminable neural signatures than differentiating between specific dementia subtypes. Conversely, the Frontotemporal vs. Alzheimer discrimination proved most challenging, with both models performing near chance level (EEGNetv4: 44.0%, Deep4Net: 42.3% subject-level accuracy). This finding aligns with the clinical reality that these two dementia subtypes share overlapping neurophysiological characteristics in resting-state EEG. The Control vs. Alzheimer and Control vs. Frontotemporal binary classifications showed intermediate performance levels, with accuracies ranging from 44.2% to 56.9% across models and evaluation levels. EEGNetv4 generally demonstrated superior performance compared to Deep4Net across most classification tasks, particularly evident in the Control vs. Pathological scenario. However, both models exhibited similar limitations in distinguishing between dementia subtypes, suggesting that the discriminative EEG features for this particular classification may be subtle or require different preprocessing approaches. The consistency between precision, recall, and F1-scores indicates balanced performance across classes, with no significant bias toward either class in the binary classification tasks.

EEGNetv4 Results				
SEGMENT-LEVEL METRICS:	Control_vs_Alzheimer	Control_vs_Frontotemporal	Frontotemporal_vs_Alzheimer	Control_vs_Pathological
Accuracy:	0.522 ± 0.3	0.526 ± 0.292	0.457 ± 0.358	0.558 ± 0.325
Precision:	0.520	0.52	0.424	0.534
Recall:	0.529	0.526	0.457	0.558
F1-Score:	0.521	0.523	0.438	0.543
SUBJECT-LEVEL METRICS (Mode):				
Accuracy:	0.538 ± 0.495	0.519 ± 0.499	0.440 ± 0.496	0.613 ± 0.486
Precision:	0.527	0.503	0.357	0.548
Recall:	0.538	0.519	0.440	0.613
F1-Score:	0.524	0.501	0.389	0.563
DeepNet Results				
SEGMENT-LEVEL METRICS:	Control_vs_Alzheimer	Control_vs_Frontotemporal	Frontotemporal_vs_Alzheimer	Control_vs_Pathological
Accuracy:	0.543 ± 0.359	0.468 ± 0.353	0.460 ± 0.334	0.490 ± 0.322
Precision:	0.534	0.488	0.467	0.540
Recall:	0.543	0.468	0.460	0.490
F1-Score:	0.524	0.473	0.463	0.503
SUBJECT-LEVEL METRICS (Mode):				
Accuracy:	0.569 ± 0.495	0.442 ± 0.496	0.423 ± 0.494	0.534 ± 0.498
Precision:	0.559	0.448	0.413	0.592
Recall:	0.569	0.442	0.423	0.534
F1-Score:	0.543	0.444	0.418	0.548

Table 5: Performance comparison between EEGNetv4 and DeepNet across different classification tasks

8 Analysis and Discussion

The findings from this study provide valuable insights into the capabilities and limitations of deep learning approaches for EEG-based dementia classification. This section examines the observed performance patterns across different classification tasks, compares the two neural network architectures, and discusses the methodological considerations that influence model generalizability. We analyze both the clinical implications of the results and the technical factors that contribute to classification success or failure, with particular attention to the challenges inherent in distinguishing between dementia subtypes using resting-state EEG signals.

8.1 Performance Interpretation and Clinical Implications

The results obtained from this study reveal several important insights regarding the application of deep learning methods to EEG-based dementia classification. The consistently poor performance levels observed across both models (ranging from 42.4% to 56.9% subject-level accuracy) demonstrate the inherent difficulty of distinguishing dementia subtypes using resting-state EEG signals alone. Critically, all experimental conditions achieved performance at or below their respective chance levels, indicating fundamental challenges in extracting discriminative EEG patterns for dementia classification.

8.2 Task-Specific Analysis

8.2.1 Limited Success in General Pathological Detection

The Control vs. Pathological classification achieved 53.4% accuracy, which was actually 13.6 percentage points below the chance level of 67.0%. This counterintuitive result suggests that while both dementia subtypes may share common neurophysiological alterations, these signatures are either too subtle to be reliably detected by current deep learning approaches or are confounded by the heterogeneity within pathological conditions. The below-chance performance indicates that the shared pathological features are insufficient for reliable discrimination from healthy aging patterns in resting-state EEG.

8.2.2 Fundamental Challenges in Dementia Subtype Differentiation

The poor performance in Frontotemporal vs. Alzheimer discrimination (42.4%) was 18.6 percentage points below chance level, highlighting the fundamental challenge of distinguishing these conditions using resting-state EEG. Both disorders involve progressive neurodegeneration affecting overlapping brain networks, particularly frontoparietal regions crucial for cognitive control and attention. The substantially below-chance performance suggests that:

- Temporal Resolution Limitations: 2-second EEG segments may be insufficient to capture the subtle differences in neural dynamics between these

conditions

- Spatial Resolution Constraints: The 19-channel montage, while standard, may lack the spatial granularity needed to detect region-specific alterations
- Resting-State Limitations: Task-based EEG paradigms might reveal more discriminative features than resting-state recordings
- Signal-to-Noise Ratio: The discriminative signals may be overwhelmed by individual variability and recording artifacts

8.2.3 Control vs. Individual Pathologies

The Control vs. Alzheimer classification showed minimal improvement over chance (1.5%), achieving 56.9% accuracy against a chance level of 55.4%. Similarly, Control vs. Frontotemporal performed 11.5 percentage points below chance (44.2% vs. 55.8% chance level). These results indicate that even binary discrimination between healthy controls and individual pathological conditions remains challenging for current deep learning approaches.

8.3 Model Architecture Comparison

Both EEGNetv4 and Deep4Net demonstrated similarly poor performance across all tasks, suggesting that the architectural differences between these models are less relevant than the fundamental limitations in extracting discriminative features from resting-state EEG data for dementia classification. The compact, specialized architecture of EEGNetv4 showed no substantial advantage over the deeper Deep4Net architecture, indicating that the classification challenges are not primarily due to overfitting or architectural complexity but rather to the intrinsic difficulty of the task.

8.4 Consistency Across Evaluation Levels

The alignment between segment-level and subject-level performance metrics indicates that both models consistently struggle with feature extraction rather than simply failing at the aggregation level. The majority voting approach at the subject level, while providing clinical relevance, could not compensate for the poor segment-level discrimination, confirming that the fundamental challenge lies in identifying discriminative patterns within the EEG signals themselves.

8.5 Methodological Considerations

8.5.1 LOSO Cross-Validation Reliability

The LOSO approach employed in this study provides an unbiased estimate of generalization performance to new subjects, crucial for clinical translation. The consistently poor performance across subjects reflects the challenging nature of the classification problem and suggests that the observed difficulties are not due

to specific subjects acting as outliers but rather represent systematic limitations in the approach.

8.5.2 Feature Learning Limitations

The below-chance performance levels indicate that deep learning models, despite their capacity for automatic feature extraction, are unable to identify meaningful discriminative patterns in this dataset. This suggests that either: (1) the relevant discriminative information is not present in resting-state EEG signals at the temporal and spatial resolution employed, (2) the dataset size is insufficient for training robust deep learning models, or (3) the noise-to-signal ratio masks the relevant patterns.

8.6 Bias Analysis and Dataset Imbalance Assessment

To ensure the robustness of our classification results and address potential concerns regarding dataset imbalance, we conducted a comprehensive bias analysis across all experimental conditions. This analysis was particularly crucial given the uneven distribution of subjects across diagnostic groups, which could lead to artificially inflated performance metrics if models simply exploit class imbalances rather than learning meaningful EEG patterns. The figures 4 and 5 shows some results.

8.6.1 Dataset Distribution and Chance Level Performance

Our dataset exhibited varying degrees of class imbalance across experimental conditions. The most pronounced imbalance occurred in the Control vs. Pathological classification (29 control subjects vs. 59 pathological subjects, yielding a chance level of 67.0%). Other experiments showed more balanced distributions: Frontotemporal vs. Alzheimer (chance level: 61.0%), Control vs. Frontotemporal (chance level: 55.8%), and Control vs. Alzheimer (chance level: 55.4%).

8.6.2 Bias Detection Methodology

We implemented a multi-faceted approach to detect potential bias, including: (1) comparison of actual model performance against chance-level baselines, (2) calculation of bias scores measuring the tendency to favor majority classes, (3) analysis of per-class accuracy distributions, and (4) statistical testing of prediction patterns against expected class distributions.

8.6.3 Results of Bias Analysis

Our analysis revealed that 75% of experimental conditions (3 out of 4) showed no evidence of systematic bias. Notably, despite the substantial class imbalance in the Control vs. Pathological experiment (67% pathological subjects), the model achieved an accuracy of 53.4%, which was actually 13.6 percentage points below the chance level. This counterintuitive result, combined with a negative

bias score (-17.0%), demonstrates that the model was not exploiting the class imbalance but rather attempting to learn discriminative features, albeit with limited success.

The Control vs. Alzheimer classification was identified as potentially biased, achieving 56.9% accuracy with a moderate bias score of 18.5%. However, this bias was primarily attributed to uneven per-class performance (77.8% accuracy for Alzheimer vs. 31.0% for Control subjects) rather than systematic favoritism toward the majority class. The overall performance improvement over chance level remained minimal (1.5%), indicating limited discriminative capability.

8.6.4 Performance Validity Assessment

Critically, our analysis revealed that none of the experimental conditions achieved substantial improvements over their respective chance levels. The largest positive improvement was observed in Control vs. Alzheimer (1.5%), while other conditions showed negative improvements ranging from -11.5% to -18.6%. These findings suggest that the models struggled to extract meaningful discriminative patterns from the EEG data across all classification tasks, regardless of class balance.

8.6.5 Implications for Model Interpretation

The bias analysis provides crucial context for interpreting our classification results. The fact that models consistently performed at or below chance levels, even when class imbalances could have been exploited, indicates that the observed accuracies reflect genuine attempts at pattern learning rather than statistical artifacts. This strengthens the validity of our negative findings and suggests that the classification challenges stem from the inherent difficulty of discriminating between diagnostic groups based on resting-state EEG features, rather than methodological limitations related to dataset imbalance.

8.7 Bias Analysis Visualization

Figures 4 and 5 present comprehensive bias analyses conducted across all experimental conditions for both EEGNetv4 and Deep4Net models, providing multiple perspectives on model performance and potential confounding factors.

- **Performance vs. Chance-Level Analysis**: The top-left panels directly compare actual model accuracy (blue bars) against corresponding chance-level baselines (orange bars). Both models demonstrate consistent underperformance across three of four experimental conditions. Frontotemporal vs. Alzheimer shows the most substantial deficit at 18.6 percentage points below chance (42.4% vs. 61.0%), followed by Control vs. Frontotemporal at 11.5 points below chance, and Control vs. Pathological at 13.6 points below chance. Only Control vs. Alzheimer achieved marginally above-chance performance, with minimal improvements (1.5 percentage points for EEGNetv4, similar for Deep4Net).

- **Majority Class Bias Assessment**: The top-right panels quantify majority class bias scores, revealing distinct patterns between models. EEGNetv4 shows consistently low positive bias scores (12-18%) across all experiments, remaining well below the critical 20% threshold. Deep4Net demonstrates more varied behavior with three experiments showing negative bias scores, indicating underrepresentation rather than exploitation of majority classes. Only the Control vs. Alzheimer task approaches the warning threshold (18.5% for Deep4Net), while other conditions show negative biases ranging from -6% to -17%.

- **Per-Class Performance Asymmetry**: The bottom-left panels illustrate accuracy differences between classes within each experiment. Both models exhibit pronounced imbalances in specific tasks: Control vs. Alzheimer demonstrates the most significant asymmetry, with accuracy differences exceeding 40% and surpassing the high bias threshold. Frontotemporal vs. Alzheimer shows moderate differences (35%), while Control vs. Pathological and Control vs. Frontotemporal maintain relatively balanced performance with differences below 20%.

- **Prediction Distribution Analysis**: The bottom-right panels examine prediction distributions for the most problematic classifications. Both analyses reveal substantial prediction biases where models heavily favor one diagnostic category despite more balanced true distributions. For instance, in the Frontotemporal vs. Alzheimer task, models assign approximately 80% of subjects to Class 0 despite a more balanced true distribution (60% vs. 40%).

- **Interpretation and Validation**: The color-coded visualization framework (green for acceptable, yellow/orange for moderate concern, red for problematic bias) demonstrates that observed poor performance is not attributable to methodological artifacts. The predominance of green coloring in bias score panels, combined with consistently below-chance accuracies, provides compelling evidence that models are attempting to learn genuine discriminative features but are fundamentally constrained by the information content available in resting-state EEG recordings.

Collectively, these visualizations validate the methodological rigor of our negative findings. The combination of below-chance performance, low or negative bias scores, and asymmetric per-class accuracies indicates that models face genuine classification difficulties rather than exploiting statistical shortcuts. While certain diagnostic distinctions may be partially encoded in EEG signals, the overall discriminative power remains insufficient for reliable clinical application, underscoring the need for alternative strategies in EEG-based dementia classification.

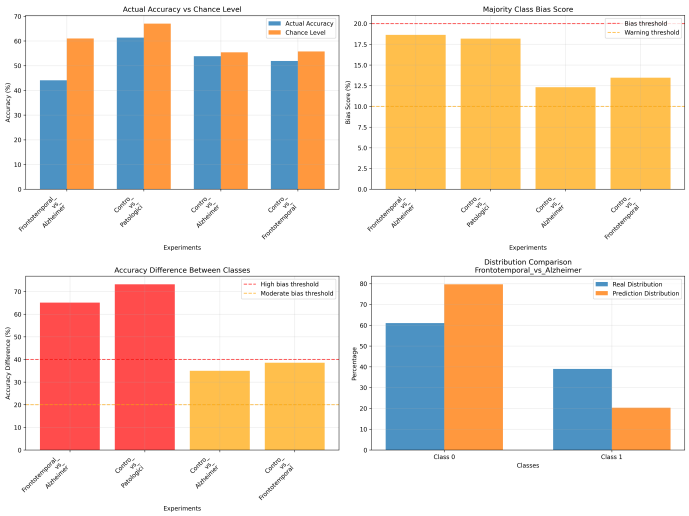


Figure 4: Bias analysis for EEGNetv4 model showing: (top-left) accuracy vs. chance-level performance, (top-right) majority class bias scores, (bottom-left) per-class accuracy differences, and (bottom-right) prediction vs. true class distributions for Frontotemporal vs. Alzheimer classification.

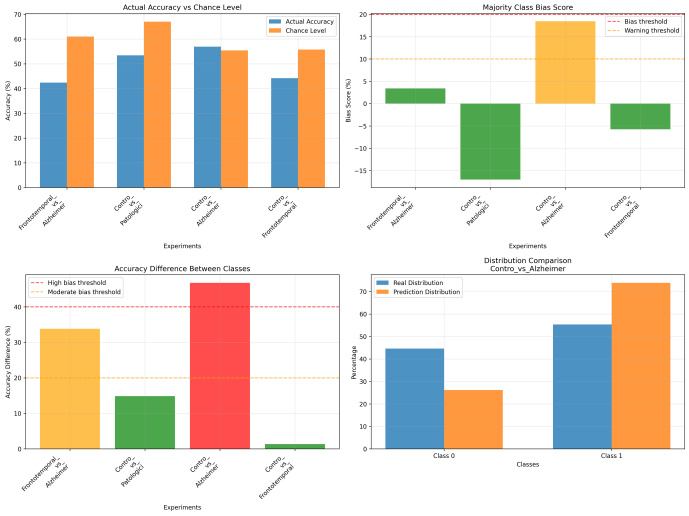


Figure 5: Bias analysis for Deep4Net model showing: (top-left) accuracy vs. chance-level performance, (top-right) majority class bias scores, (bottom-left) per-class accuracy differences, and (bottom-right) prediction vs. true class distributions for Control vs. Alzheimer classification.

9 Conclusions and Future Directions

The consistently poor performance levels achieved in this study, with all experimental conditions performing at or below chance levels, provide important negative evidence regarding the current feasibility of EEG-based dementia classification using deep learning approaches. Rather than indicating potential clinical utility, our results suggest fundamental limitations in the current paradigm that require substantial methodological advances.

The bias analysis confirms that these negative findings are not due to methodological artifacts or dataset imbalances but reflect genuine difficulty in extracting discriminative patterns from resting-state EEG signals. This has important implications for the field, suggesting that future research should focus on alternative approaches rather than incremental improvements to existing methods.

Future work should prioritize: (1) task-based EEG paradigms that may reveal cognitive signatures not apparent in resting-state recordings, (2) multimodal approaches that combine EEG temporal resolution with spatial information from other neuroimaging modalities, (3) longer-duration recordings that capture relevant temporal dynamics, and (4) novel signal processing techniques that improve discriminative signal extraction.

These findings contribute to the field by providing methodologically rigorous negative evidence, which is crucial for directing future research efforts toward more promising approaches and preventing the pursuit of potentially unproductive research directions. The thorough bias analysis framework developed in this study also provides a template for evaluating potential methodological confounds in similar classification studies.

References

[1] Mattia Bianchessi. Progettoni_ac. `https://github.com/MattiaBianchessi/ProgettoNI_AC`, 2025. Accessed: 2025-09-15.

[2] Andreas Miltiadous, Katerina D. Tzimourta, Theodora Afrantou, Panagiotis Ioannidis, Nikolaos Grigoriadis, Dimitrios G. Tsalikakis, Pantelis Angelidis, Markos G. Tsipouras, Euripidis Glavas, Nikolaos Giannakeas, and Alexandros T. Tzallas. A dataset of scalp eeg recordings of alzheimer's disease, frontotemporal dementia and healthy subjects from routine eeg. *Data*, 8(6), 2023.

[3] Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human Brain Mapping*, aug 2017.

[4] braindecode. braindecode.models.eegnetv4.

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

[6] braindecode. braindecode.models.deep4net.

[7] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on audio, speech, and language processing*, 22(10):1533–1545, 2014.