NeuroXAI: Explainable Deep Learning for EEG-Based Detection of Alzheimer's and Parkinson's Diseases

1st Chayut Bunterngchit
Division of Industrial and Logistics Engineering Technology, Faculty of Engineering and Technology, King Mongkut's University of Technology North Bangkok, Rayong Campus
Rayong, Thailand
chayut.b@eat.kmutnb.ac.th

2nd Laith H. Baniata
Department of Autonomous Systems, Faculty of Artificial Intelligence, Al-Balqa Applied University
Al-Salt, Jordan
laith.baniata@bau.edu.jo

3rd Abdur Rasool
Department of Information and Computer Sciences, University of Hawaii at Manoa
Honolulu, HI, USA
abdur@hawaii.edu

2025 IEEE 25th International Conference on Bioinformatics and Bioengineering (BIBE) | 979-8-3315-5899-4/25/$31.00 ©2025 IEEE | DOI: 10.1109/BIBE66822.2025.00107

***Abstract*—The accurate and interpretable detection of neurodegenerative disorders remains a critical challenge in clinical neuroscience. NeuroXAI is introduced as a high-performance and explainable artificial intelligence (XAI)-driven deep learning framework for the electroencephalography (EEG)-based classification of Alzheimer's disease (AD) and Parkinson's disease (PD). The framework integrates a hybrid temporal convolutional network with disease-specific attention mechanisms to capture multi-granular spatial–temporal representations, achieving 98.79% accuracy for PD and 90.64% for AD on benchmark datasets, with robust generalization under leave-one-subject-out cross-validation. Clinical interpretability is supported through a comprehensive XAI pipeline that combines gradient-based saliency, integrated gradients, and perturbation-based occlusion analyses, which expose channel- and time-specific activation patterns consistent with known motor and cognitive biomarkers. Functional connectivity and nonlinear signal complexity measures, including fractal dimension and entropy, further validate disease-specific neural disruptions captured by the model. By uniting high predictive performance with clinically meaningful explanations, NeuroXAI offers a robust and deployment-ready solution for EEG-based neurological disorder screening, advancing the adoption of XAI in real-world diagnostics.**

***Index Terms*—Electroencephalography (EEG), Alzheimer's disease, Parkinson's disease, Explainable artificial intelligence (XAI), Deep learning, Attention mechanism**

I. Introduction

Neurodegenerative disorders, including Alzheimer's disease (AD) and Parkinson's disease (PD), pose a significant and growing global health burden [1], [2]. The number of people living with dementia is projected to triple by 2050 [3], [4]. AD remains the most prevalent form of dementia, predominantly affecting older populations, while PD, the second most common neurodegenerative disorder, is expected to rise to 25 million cases by 2050 [5]–[7]. The aging global population is a primary driver of this increase, underscoring the urgent need for accurate and early diagnostic methods to enable timely intervention and improved long-term outcomes.

In response to this, deep learning (DL) frameworks have emerged as promising tools for the diagnosis of neurodegenerative diseases [8], [9]. Applied to electroencephalography (EEG) data, these methods automatically extract complex, high-dimensional neural features that are often difficult to capture with conventional analysis. Despite their high diagnostic accuracy, DL models remain inherently opaque, functioning as black boxes [10]–[12], which hinders clinical adoption where transparency and interpretability are essential. To address this limitation, explainable artificial intelligence (XAI) has gained prominence [13], [14], aiming to develop models that produce human-interpretable explanations, thereby enhancing clinician trust and facilitating clinical validation.

Several studies have leveraged XAI techniques for the diagnosis of AD and PD, aiming to combine the predictive strength of the DL models with clinically interpretable decision-making. Prior work has explored EEG-centric models, attention-based mechanisms, and post hoc explainability methods to uncover neurophysiological insights. Within this context, multimodal frameworks have gained increasing traction by integrating magnetic resonance imaging (MRI), EEG, genetic, and clinical data with interpretable layers. For instance, an interpretable transformer-based model [15] was developed using MRI and tabular data, while vision transformers combined with a gated recurrent unit [16] employed Shapley additive explanations (SHAP) and local interpretable model-agnostic explanation (LIME) to highlight region-specific brain contributions. Genomic studies have also incorporated XAI [17], leveraging LIME and permutation importance for early AD identification. Similarly, neuroimaging and gene expression data have been fused [18] to localize biomarkers through LIME, a transformer-driven multimodal approach integrating MRI, demographic, and cognitive features were characterized [19]. For comparative analysis, multiple convolutional neural network (CNN) variants evaluated with LIME [20] further reinforced region-level interpretability. Collectively, these approaches signify a shift toward comprehensive, biologically informed XAI applications.

Within EEG-focused studies, both lightweight architectures

2471-7819/25/$31.00 ©2025 IEEE
DOI 10.1109/BIBE66822.2025.00107

TABLE I
POSITIONING NEUROXAI RELATIVE TO REPRESENTATIVE XAI-DRIVEN MODELS

Model	Modality	XAI type	Granularity	Strengths / Limitations
[15]	MRI & tabular	Intrinsic and post hoc	Region-level	Interpretable attention; non-EEG.
[16]	MRI	Post hoc	Region-level	Region importances; no electrophysiology.
[21]	EEG	Post hoc	Channel/time	Efficient; no uncertainty or biomarker link.
[22]	EEG	Intrinsic	Channel/time	Disentangled attention; limited clinical validation.
[23]	EEG (handcrafted)	Post hoc	Feature-level	Requires feature engineering; limited temporal insight.
NeuroXAI	EEG	**Intrinsic and post hoc**	**Channel/time (multi-granular)**	**Biomarker alignment and functional connectivity validation**

and attention-enhanced networks have been explored, with growing emphasis on real-time applicability and clinical interpretability. For instance, a compact CNN architecture [21] was developed for efficient AD detection, employing SHAP and gradient-weighted class activation mapping (Grad-CAM) to localize predictive features. The combination of CNN and recurrent neural network layers [24] was proposed for mental health diagnosis, using layer-wise relevance propagation and SHAP visualizations tailored for wearable EEG systems. Tree-based models [23] were introduced to EEG-derived features, leveraging SHAP and LIME for feature relevance attribution. More advanced attention mechanisms were introduced [22], where orthogonal attention provided disentangled temporal–spatial contributions for AD classification. Additionally, visualization-driven studies [25], [26], demonstrated comparative saliency mapping across LIME, Grad-CAM, providing a cross-method perspective on feature attribution.

These studies demonstrate the growing adoption of XAI techniques in neurodiagnostic pipelines and the progress in model interpretability, yet several critical limitations remain. Most prior approaches rely on post hoc explanations that are decoupled from model training objectives, resulting in visualizations that lack strong clinical correlation or neurobiological validation of EEG biomarkers. Additionally, nonlinear signal dynamics, such as fractal dimension and sample entropy, are often underutilized, despite their potential to differentiate pathological EEG patterns. Model calibration and uncertainty quantification are also rarely addressed, even though these components are essential for reliable clinical deployment.

To overcome these gaps, this study introduces NeuroXAI, a novel framework that integrates a temporal convolutional network backbone with disease-specific, channel-conditioned attention mechanisms and a comprehensive suite of XAI tools. The architecture employs a three-block 1D CNN backbone for feature extraction, enhanced by channel-wise attention layers that emphasize disease-relevant neural patterns. EEG signals are homogenized and segmented into overlapping frames, combined with data augmentation strategies such as noise injection and temporal shifts to improve generalization. Multi-granular explainability is achieved using saliency maps, integrated gradients, Grad-CAM, LIME, SHAP-like gradient attributions, and temporal occlusion analysis, while Monte Carlo dropout provides model-level uncertainty estimation for subject-specific interpretability reports. Finally, nonlinear dynamics in the form of fractal dimension and sample entropy serve as post hoc validation measures to bridge model activations with clinically recognized EEG biomarkers.

The existing work has prioritized the interpretability for the neurodiagnostics. Nevertheless, their direct comparisons are confounded by differing modalities, datasets, and evaluation protocols. To contextualize NeuroXAI, a contrast representation has been qualitatively been presented in XAI-driven models along four axes as presented in Table I. This leads to making a material impact clinical utility: (i) modality, (ii) XAI type, (iii) attribution granularity, and (iv) biomarker alignment and auxiliary validation. Thus, NeuroXAI differs by combining intrinsic, disease-specific, channel-conditioned attention with multi-granular post hoc analyses on EEG.

Overall, the work contributes as follows:

- Proposes an interpretable CNN model with embedded channel-conditioned attention layers to differentiate the dynamics of AD and PD.
- Introduces multi-granular explainability by integrating visual, statistical, and attribution-based XAI methods for EEG interpretation.
- Embeds fractal dimension and entropy measures to connect DL predictions with clinical EEG markers.
- Incorporates Monte Carlo dropout for uncertainty estimation and evaluates model calibration for reliable deployment.
- Provides comprehensive visual analytics, including t-SNE embeddings, channel attention summaries, and case-level diagnostic reports.

II. METHODOLOGY

This section presents a complete pipeline that has been adopted in the development of the model.

A. Preprocessing and Augmentation Strategy

Let $\mathcal{X} = \{\mathbf{X}^{(i)} \in \mathbb{R}^{C\times T_i}\}_{i=1}^{N}$ denote raw EEG recordings, where C is the number of electrodes and T_i the time-series length for subject i. All trials are temporally truncated to $\tau = \min_i T_i$ to ensure homogeneity.

Each sequence is subdivided into M overlapping windows $\{\mathbf{X}_j^{(i)} \in \mathbb{R}^{C\times \ell}\}_{j=1}^{M}$ by using a sliding window of length ℓ and overlap rate $\rho \in (0, 1)$:

$$\mathbf{X}_j^{(i)} = \left[\mathbf{X}^{(i)}\right]_{[:,\, j\delta:j\delta+\ell]}, \quad \delta = \lfloor \ell(1-\rho) \rfloor.$$

Each frame has then been normalized across time as follows:

$$\hat{\mathbf{X}}_j^{(i)} = \frac{\mathbf{X}_j^{(i)} - \mu_j^{(i)}}{\sigma_j^{(i)} + \epsilon}, \quad \mu_j^{(i)} = \mathbb{E}[\mathbf{X}_j^{(i)}], \quad \sigma_j^{(i)} = \sqrt{\mathbb{V}[\mathbf{X}_j^{(i)}]}.$$

To regularize representations under small samples, stochastic transformations are applied as follows:

$$\tilde{\mathbf{X}} = \mathbf{X} + \epsilon + \mathcal{P}_r(\mathbf{X}) + \mathcal{F}_p(\mathbf{X}),$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is Gaussian noise, $\mathcal{P}_r$ denotes a circular temporal shift with offset r, and $\mathcal{F}_p$ is a Bernoulli channel-flip operator.

B. Deep Neural Architecture and Training

The input tensor $\mathbf{X} \in \mathbb{R}^{C \times \ell}$ has been subjected to several transformations through convolutional blocks. Define three filter banks $\{\mathbf{W}_k\}_{k=1}^3$ with increasing channel depth and receptive field widths. The transformation hierarchy is defined recursively as follows:

$$\mathbf{H}^{(1)} = \sigma \circ \phi \left(\mathbf{W}_1 * \mathbf{X} \right), \tag{1}$$

$$\mathbf{H}^{(2)} = \sigma \circ \phi \left(\mathbf{W}_2 * \mathbf{H}^{(1)} \right), \tag{2}$$

$$\mathbf{H}^{(3)} = \sigma \circ \phi \left(\mathbf{W}_3 * \mathbf{H}^{(2)} \right), \tag{3}$$

where $*$ denotes 1D convolution, ϕ is a normalization operator (e.g., batch normalization), and σ is a nonlinearity (e.g., ReLU or $\tanh$).

1) Attention modulation: A disease-specific attention mechanism refines the third-layer activations:

$$\mathbf{A}^{(d)} = \text{Softmax} \left(\mathbf{V}_d * \mathbf{H}^{(3)} \right), \quad d \in \{\text{PD}, \text{AD}\},$$

where $\mathbf{V}_d$ is a convolutional attention kernel tailored to disease type d. The attended representation is:

$$\widetilde{\mathbf{H}} = \mathbf{H}^{(3)} \odot \mathbf{A}^{(d)}.$$

2) Dense output projection: Feature aggregation and classification proceed via:

$$\mathbf{z} = \zeta \left(\mathbf{W}_4 \cdot \text{GAP}(\widetilde{\mathbf{H}}) + \mathbf{b}_4 \right), \tag{4}$$

$$\hat{y} = \sigma \left(\mathbf{w}_5^\top \mathbf{z} + b_5 \right), \tag{5}$$

where GAP is global average pooling, ζ is a dropout-regularized nonlinearity, and $\hat{y}$ is the predicted probability of disease presence.

3) Loss and optimization: Training has been conducted using binary cross-entropy loss:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i) \right].$$

Parameter updates follow the Adam optimizer:

$$\theta \leftarrow \theta - \eta \cdot \frac{\widehat{m}_t}{\sqrt{\widehat{v}_t} + \epsilon},$$

where $\widehat{m}_t$ and $\widehat{v}_t$ are from exponential moment estimates.

The forward and backward passes to the model, along with the established pipeline, have been depicted in Algorithm 1.

Algorithm 1 End-to-end attention-modulated forward-backward training

Require: Dataset $\mathcal{D} = \{(\mathbf{X}_i, y_i)\}_{i=1}^N$, epochs E, batch size B, learning rate η

for $e = 1$ to E **do**
  Shuffle $\mathcal{D}$
  for each mini-batch $\mathcal{B} = \{(\mathbf{X}_i, y_i)\}_{i=1}^B$ **do**
    for $i = 1$ to B **do**
      $\mathbf{H}_1 \leftarrow \sigma(\phi(\mathbf{W}_1 * \mathbf{X}_i))$
      $\mathbf{H}_2 \leftarrow \sigma(\phi(\mathbf{W}_2 * \mathbf{H}_1))$
      $\mathbf{H}_3 \leftarrow \sigma(\phi(\mathbf{W}_3 * \mathbf{H}_2))$
      $\mathbf{A} \leftarrow \text{Softmax}(\mathbf{V}_d * \mathbf{H}_3)$
      $\widetilde{\mathbf{H}} \leftarrow \mathbf{H}_3 \odot \mathbf{A}$
      $\mathbf{z} \leftarrow \zeta(\mathbf{W}_4 \cdot \text{GAP}(\widetilde{\mathbf{H}}) + \mathbf{b}_4)$
      $\hat{y}_i \leftarrow \sigma(\mathbf{w}_5^\top \mathbf{z} + b_5)$
    end for
    Compute $\mathcal{L}_{\text{BCE}}$ on batch
    Backpropagate gradients via the chain rule
    Update $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}$
  end for
end for

C. Explainability Framework

Several gradient-based and model-agnostic interpretability techniques are employed as follows:

Saliency Maps:

$$\mathbf{S}_i = \left| \frac{\partial \hat{y}_i}{\partial \mathbf{X}_i} \right|, \quad \text{visualized as heatmaps over } C \times \ell.$$

Integrated gradients:

$$\text{IG}_i = (\mathbf{X}_i - \mathbf{X}_0) \cdot \frac{1}{K} \sum_{k=1}^{K} \frac{\partial \hat{y}}{\partial \mathbf{X}} \bigg|_{\mathbf{X}_0 + \frac{k}{K}(\mathbf{X}_i - \mathbf{X}_0)}.$$

Temporal occlusion: Time-indexed occlusion masks $\mathbf{M}_t$ suppress $\mathbf{X}_{[:,t:t+w]}$ and record $\Delta \hat{y}_t$.

LIME: A sparse linear model $\hat{y} \approx \sum_j \beta_j x_j$ is fit on flattened, perturbed input segments $\mathbf{x}_j$.

Grad-CAM:

$$\text{CAM}_t = \sum_c \alpha_c \mathbf{H}_{c,t}, \quad \alpha_c = \frac{1}{\ell} \sum_{t'} \frac{\partial \hat{y}}{\partial \mathbf{H}_{c,t'}}.$$

The saliency aggregation by class has been carried out by using Algorithm 2.

Algorithm 2 Class-wise saliency map aggregation

Require: Labeled dataset $\mathcal{D} = \{(\mathbf{X}_i, y_i)\}_{i=1}^{N}$, class d
Ensure: Average saliency map $\overline{\mathbf{S}}^{(d)}$

```
Initialize S̄^(d) ← 0
n_d ← 0
for i = 1 to N do
    if y_i = d then
        S_i ← |∇_{X_i} F_θ(X_i)|
        S̄^(d) ← S̄^(d) + S_i
        n_d ← n_d + 1
    end if
end for
return S̄^(d) / n_d
```

D. Nonlinear Biomarker Analysis

Each EEG channel $\mathbf{x}_c \in \mathbb{R}^{\ell}$ is projected into a nonlinear feature space:

$$\mathbf{f}_c = [D_c, E_c], \quad D_c = \text{Higuchi}(\mathbf{x}_c), \quad E_c = \text{SampEn}(\mathbf{x}_c),$$

with class-level distributions plotted via kernel density and violin plots.

E. Latent Space Structure and Epistemic Uncertainty

Embedding vectors $\mathbf{z}_i$ from the penultimate layer are projected via t-SNE for clustering analysis. Epistemic uncertainty is estimated via Monte Carlo Dropout:

$$\hat{y}_i^{(t)} = \mathcal{F}_\theta(\mathbf{X}_i; \text{Dropout}), \quad \mu_i = \mathbb{E}_t[\hat{y}_i^{(t)}], \quad \sigma_i^2 = \text{Var}_t[\hat{y}_i^{(t)}].$$

Reliability diagrams and histogram plots quantify confidence calibration.

III. Results

The diagnostic framework presented in the study has been evaluated using two neurodegenerative disorders (PD and AD). The first dataset is the UC San Diego Resting-State EEG dataset that include records from the individuals with PD and cognitively normal (CN) [27]. The second dataset includes the resting-state EEG records from individuals with AD, frontotemporal dementia (FTD), and CN [28]. These data were collected from AHEPA University Hospital, Greece. The key attributes of both datasets has been presented in Table II.

EEG open source datasets from open-neuro have been employed, and the model has been evaluated using several quantitative metrics (accuracy, precision, recall, F1-score, sensitivity, and specificity), along with the XAI outcomes.

A. Classification Performance Summary

To ensure subject-independent evaluation and avoid data leakage, the study employed both standard train-test splits and leave-one-subject-out (LOSO) cross-validation to evaluate the generalizability of the proposed model. In LOSO, for each iteration, data from one subject was held out as the test set while the model was trained on data from all other subjects. Performance metrics were averaged across all subjects to report mean values along with standard deviation.

The proposed DL model has been found to achieve high accuracy across both AD and PD. For PD, the model attained an accuracy of **98.79%**, precision of **99.32%**, recall of **98.83%**, F1-score of **99.07%**, sensitivity of **98.83%**, and specificity of **98.73%**. In contrast, AD classification retuned an accuracy of **90.64%**, precision of **90.97%**, recall of **87.72%**, and F1-score of **89.32%**. These findings have been summarized in Table III.

Under the more rigorous LOSO evaluation, the model performance decreased slightly due to subject variability. The PD classification accuracy was **96.41% ± 1.87**, precision **96.87% ± 1.54**, recall **96.12% ± 2.03**, F1-score **96.49% ± 1.62**, sensitivity **96.12% ± 2.03**, and specificity **96.68% ± 1.45**. For AD, the LOSO accuracy was **88.03% ± 2.74**, precision **88.96% ± 2.31**, recall **85.20% ± 3.02**, and F1-score **86.83% ± 2.48**. These results demonstrate strong generalization performance across unseen subjects.

The training and validation for both diseases have rendered a stable convergence and generalization. The training and loss curves for PD have depicted that minimal overfitting has been recorded (Figures 1a, 1b). Similarly, AD has shown accepted levels of convergence despite exhibiting a greater class overlap and variability in the signals (Figures 1a, 1b).

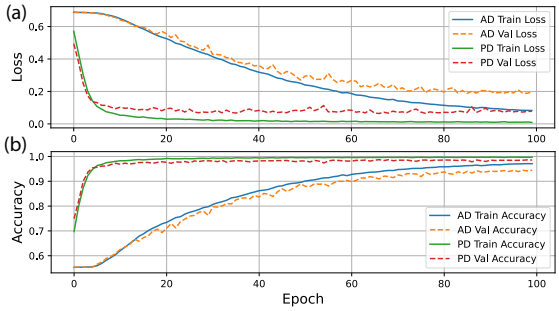

Fig. 1. Training performance metrics for AD and PD classification: (a) loss curves and (b) accuracy curves.

This performance gap between AD and PD can be attributed to the distinct neurophysiological substrates of the disorder. PD-related EEG alterations can manifest motor-relevant frequency bands and rendering improved temporal-spatial patterns. This helps in enhancing the separability of the representations that have been learned by the model. On the other side, the AD signatures are more diffused and subtle and thus require fine-grained abstractions. This can benefit them for further architectural or multimodal enhancements. Nevertheless, the overall performance is found to be scalable and serves as a foundation for automated neurological disorder screening using EEG signals.

B. Explainability Analysis and Clinical Interpretability

To further ensure that the model's decision-making process meets the clinical interpretability standards, several XAI methods have been employed as part of the trained architecture. The focus has been kept on interpretability and explainability. The formet refers to the model's inherent transparency including attention weights that are directly learned and visualized during

TABLE II
OVERVIEW OF THE EEG DATASETS USED

Attribute	PD dataset (UC San Diego)	AD/FTD dataset (AHEPA University Hospital)
Participants	31 (15 PD and 16 CN)	88 (36 AD, 23 FTD, and 29 CN)
Age (mean ± SD)	PD: 63.2 ± 8.2, CN: 63.5 ± 9.6	63.6 to 67.9 years across groups
Modality	Resting-state EEG	Resting-state EEG (eyes closed)
Channels	40 channels	19 scalp + 2 reference (10–20 system)
Sampling rate	500 Hz	500 Hz
Recording duration	5–10 min per session	13.5 min (AD), 12 min (FTD), 13.8 min (CN)
Diagnosis classes	PD, CN	AD, FTD, CN

TABLE III
PERFORMANCE METRICS FOR PD AND AD DETECTION

Condition	Accuracy	Precision	Recall	F1-Score	Sensitivity	Specificity
Standard evaluation						
PD	98.79%	99.32%	98.83%	99.07%	98.83%	98.73%
AD	90.64%	90.97%	87.72%	89.32%	92.7%	92.63%
LOSO cross-validation (mean ± SD)						
PD	96.41% ± 1.87	96.87% ± 1.54	96.12% ± 2.03	96.49% ± 1.62	96.12% ± 2.03	96.68% ± 1.45
AD	88.03% ± 2.74	88.96% ± 2.31	85.20% ± 3.02	86.83% ± 2.48	–	–

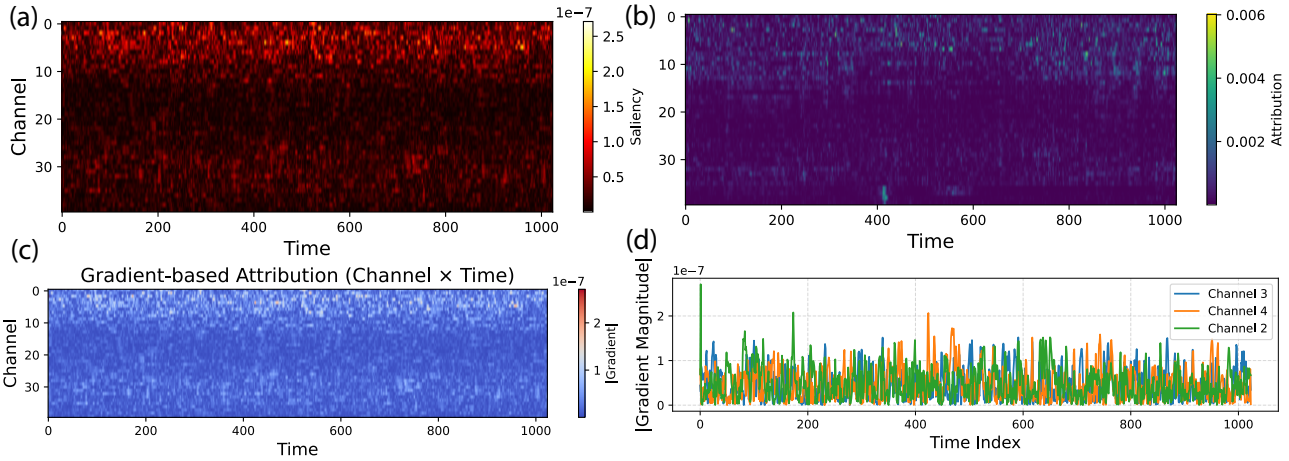

Fig. 2. Gradient-based attribution and interpretability results: (a) Saliency map showing channel-wise and temporal contributions to classification; (b) Integrated gradients indicating cumulative attributions across EEG time series; (c) Gradient-based heatmap illustrating localized activations in the Channel × Time domain; (d) Temporal saliency of top contributing channels in PD detection.

training. The latter denotes post hoc analysis methods applied after model training, such as SHAP, LIME, and saliency maps, which provide external approximations of feature importance. The employed methods are grouped into four functional categories that offer distinct yet complementary insights:

- **Gradient-Based Attribution:** Includes saliency maps, integrated gradients, and gradient × input methods. These quantify the influence of input features by measuring sensitivity of the output with respect to small perturbations.
- **Perturbation-Based Analysis:** Includes temporal occlusion and LIME. These explicitly alter input regions to assess their causal impact on prediction, offering intuitive insights into local feature importance.
- **Model-Intrinsic Attention:** Exploits internal attention weights from trained modules, shedding light on which channels/time segments the model inherently focuses on during prediction.
- **Global Interpretability and Complexity Profiling:** Includes SHAP-like summary statistics, disease-specific saliency averaging, functional connectivity (FC), t-SNE visualization, and calibration analysis. These methods interpret group-level patterns and model confidence to aid clinical relevance.

This combination avoids redundancy by spanning both local and global interpretability, while targeting temporal, spatial, and structural insights across the EEG signals.

1) Gradient-based attribution: Saliency maps (Figure 2a) have been computed using first-order gradients in contrast to the revealed concentration activity levels. These are recorded along the specific EEG channels and time segments. The regions have exhibited high-activation scores and project a high influence on the predicted neurological class. These focused activation patterns reveal that localized neurophysiological

TABLE IV
TOP SALIENT EEG CHANNELS OF PD AND CN

Channel	PD mean	CN mean	$\lvert\Delta\rvert$
31	6.1×10^{-5}	2.0×10^{-6}	5.9×10^{-5}
32	5.5×10^{-5}	3.0×10^{-6}	5.2×10^{-5}
29	5.4×10^{-5}	2.0×10^{-6}	5.2×10^{-5}
30	5.4×10^{-5}	3.0×10^{-6}	5.2×10^{-5}
33	5.4×10^{-5}	3.0×10^{-6}	5.1×10^{-5}

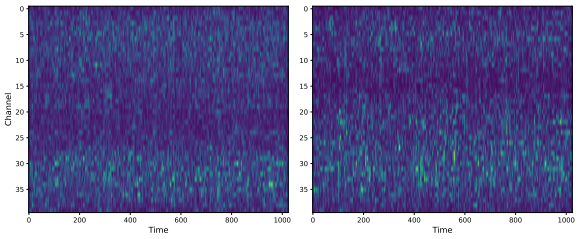

Fig. 3. Average saliency maps for CN (left) and PD (right) cohorts.

signatures have been employed and are deemed consistent with the domain knowledge. This applies to the region-specific abnormalities in PD and AD.

The integrated gradients have further been presented to infer the cumulative importance of the EEG signals. This integration of gradients has been carried out along the path from the baseline to the actual input. The findings are presented in Figure 2b, which depicts that a heightened attribution has been recorded across mid-latency temporal bands. These patterns are aligned with the known disruptions in the mid-range neural oscillations, which are associated with cognitive impairments and motor degradations.

The gradient-based attribution heatmaps further offer a temporal view of sensitivity across the input dimension as depicted in (Figure 2c). These maps highlight the localized signal bursts in inter-channel transitions, which modulate the output of the model. Notably, sporadic regions of high attribution intensity reveal the time-channel segments where the model focuses most during decision-making, suggesting the importance of transient neural activations.

To expand upon the XAI strategies, the saliency maps were further computed by using backpropagated gradients. These maps led to a quantification of the local sensitivity of the predictions and perturbations across each temporal channel location. Channels 31–35 exhibited distinctly elevated saliency in PD compared to CN (Table IV). These findings highlight disease-specific cortical involvement.

The SHAP representations derived from the gradients have been analyzed in the context of global importance as depicted in (Figure 2d). These affirm that the saliency peak is around channels 28-33. The disease-specific saliency maps have been averaged across 20 subjects for the PD and CN groups and presented in Figure 3.

The consistency of the attributions reported across the saliency and integrated gradient methods suggests that the model is not reliant on the spurious correlations. Instead, the model captures the salient neural dynamics and reflects the disease-specific alterations. This is especially important for the real-world clinical settings where transparency can lead to improved hypothesis generation and diagnosis of disease with more confidence.

2) Perturbation-based analysis: The temporal occlusions have been carried out for the identification of the segments where removal can lead to alterations in the prediction probabilities. The results have been depicted in Figure 4a and depict that distinctive temporal fragments which are localized across 200–600 ms across most individuals with PD.

LIME explanations (Figure 4b) have been presented in complement to this by using the flattened input representations. These were exposed to the top 10 channel time feature combinations that contributed to the class predictions.

3) Model-intrinsic attention: From the learned attention mechanisms of the mode, mean attention weights were extracted for EEG channels for visualization (Figure 4c). The channels with high attention weights refer to the top saliency regions as depicted in (Figure 4d). These findings indicate that the model's focus aligned with gradient-based relevance.

4) Functional connectivity and nonlinear complexity: To further probe into the inter-channel relationships, the EEG functional connectivity matrices across trials have been extracted. Figure 5 illustrates the resulting matrix and suggests that the localized synchronic disruptions exist in PD.

5) Latent space structure and model confidence: The XAI feature embedding further involves a visualization using t-SNE as depicted in (Figure 6). These class clusters have shown a clear separability that confirms the learned representations and the discriminative power of the model. In addition, the model has exhibited reliability using calibration curves in (Figure 7). These findings have shown that a high agreement exists between the predicted confidence and empirical accuracy.

IV. DISCUSSION AND FUTURE DIRECTIONS

The results of this study highlight the potential of NeuroXAI to provide both high diagnostic accuracy and interpretable insights for EEG-based detection of AD and PD. Compared to conventional EEG classification methods and prior XAI-driven approaches, the framework offers a multi-granular understanding of model behavior, bridging the gap between DL predictions and clinically meaningful biomarkers. The channel- and time-specific relevance patterns revealed by saliency and integrated gradients align with known neurophysiological pathways involved in motor and cognitive impairment, while LIME and temporal occlusion analyses highlight mid-latency periods that are consistent with disrupted neural timing in AD and PD. Furthermore, functional connectivity and nonlinear complexity measures support the biological plausibility of the learned representations. Despite these strengths, the approach is limited to EEG data from controlled datasets, and the interpretability pipeline still requires clinical validation to ensure that the highlighted regions and temporal windows correspond to actionable biomarkers in diverse patient populations.

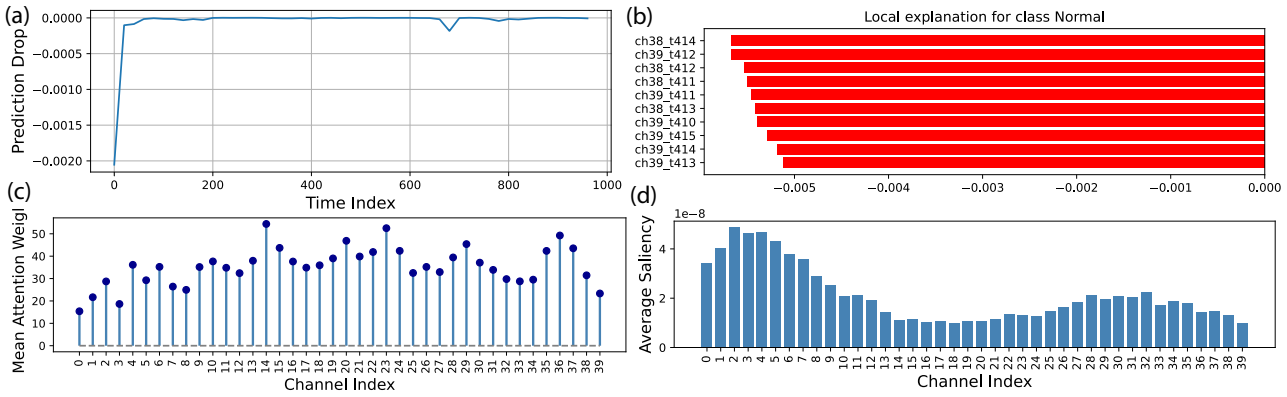

Fig. 4. Perturbation-based and model-intrinsic interpretability analyses for EEG-based PD detection: (a) Temporal occlusion sensitivity showing prediction drops when critical segments are removed; (b) LIME explanation highlighting the top 10 influential time–channel feature combinations; (c) Mean attention weights per EEG channel derived from the trained model; (d) Average gradient-based saliency scores per channel.

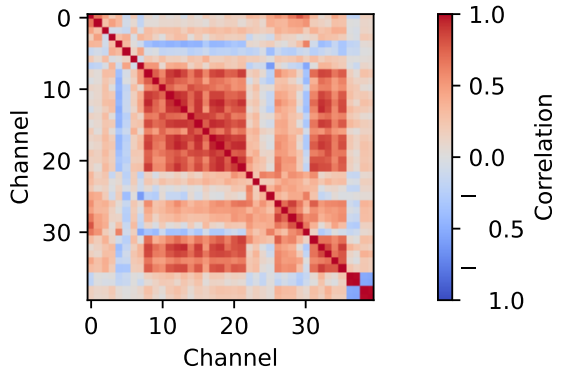

Fig. 5. Functional connectivity matrix for a PD trial.

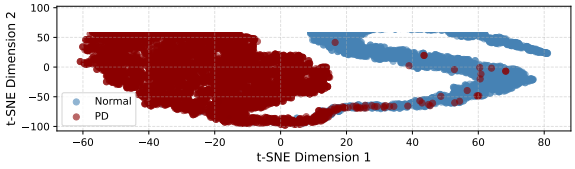

Fig. 6. t-SNE projection of latent space embeddings for PD and Normal test samples.

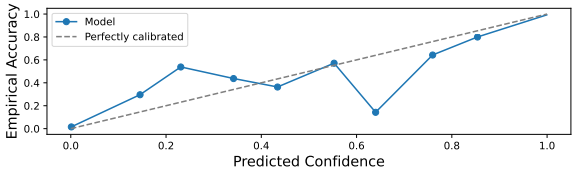

Fig. 7. Reliability diagram indicating calibration quality of the model's probabilistic predictions.

Although the datasets employed in this study were collected in controlled laboratory or clinical settings, real-world EEG acquired from wearable, long-term, or ambulatory devices is frequently affected by artifacts (e.g., ocular and muscular activity, motion-induced drift, electrode displacement, and environmental noise). Such artifacts can obscure disease-relevant signatures and reduce model reliability.

The NeuroXAI framework incorporates several design choices that enhance robustness against these challenges.The disease-specific attention mechanism naturally reduces the load on noisy or unstable channels by focusing the relevance on more consistent signal sources. Then, the inclusion of uncertainty estimation and calibration provides a safeguard: predictions accompanied by high uncertainty may indicate artifact-contaminated or unreliable inputs, prompting re-evaluation or data quality checks. Nonetheless, the validation on long-term, wearable EEG remains a critical next step. Prospective deployment studies should integrate online artifact-detection modules, adaptive filtering, and active learning loops that allow NeuroXAI to dynamically recalibrate in the presence of evolving noise profiles. Such extensions will be essential for clinical translation into real-world, non-laboratory environments.

Future research should focus on broadening the applicability and reliability of NeuroXAI. Integrating multimodal data sources such as MRI, functional near-infrared spectroscopy, or genetic information could provide richer context for early-stage detection and improve cross-disease discrimination. In addition, incorporating advanced uncertainty estimation, model calibration, and domain adaptation would enhance clinical trustworthiness and robustness in real-world settings. Exploring longitudinal modeling, cross-condition transfer learning, and population-level validation could facilitate early diagnosis, disease monitoring, and generalization to heterogeneous clinical cohorts. These directions will help bridge the gap between research-grade EEG models and clinically deployable neurodiagnostic tools.

V. Conclusion

In conclusion, this study introduces NeuroXAI, an EEG-based DL framework that simultaneously achieves high classification performance and clinical interpretability for AD and PD detection. The framework integrates disease-specific attention mechanisms with a multi-granular explainability pipeline that includes saliency, integrated gradients, LIME-based occlusion, and SHAP-like analyses, providing transparent, biomarker-aligned insights that highlight disease-relevant channels and temporal windows. Complementary functional connectivity and nonlinear complexity analyses validate the neurophysiological relevance of the learned representations, and calibration assessment confirms the model's reliability for real-world use. Collectively, these contributions establish NeuroXAI as a robust, interpretable, and deployment-ready solution for EEG-based neurological disorder screening, with future work focusing on multimodal integration, enhanced uncertainty quantification, and cross-condition transfer learning.

References

[1] W. A. Rocca, "The burden of Parkinson's disease: a worldwide perspective," *The Lancet Neurology*, vol. 17, no. 11, pp. 928–929, 2018.

[2] S. F. Javaid, C. Giebel, M. A. Khan, and M. J. Hashim, "Epidemiology of Alzheimer's disease and other dementias: Rising global burden and forecasted trends," *F1000Research*, vol. 10, p. 425, 2021.

[3] G. Livingston, J. Huntley, K. Y. Liu, S. G. Costafreda, G. Selbæk, S. Alladi, D. Ames, S. Banerjee, A. Burns, C. Brayne *et al.*, "Dementia prevention, intervention, and care: 2024 report of the Lancet standing commission," *The Lancet*, vol. 404, no. 10452, pp. 572–628, 2024.

[4] A. Lastuka, E. Bliss, M. R. Breshock, V. C. Iannucci, W. Sogge, K. V. Taylor, P. Pedroza, and J. L. Dieleman, "Societal costs of dementia: 204 countries, 2000–2019," *Journal of Alzheimer's Disease*, vol. 101, no. 1, pp. 277–292, 2024.

[5] C. Bunterngchit, L. H. Baniata, H. Albayati, M. H. Baniata, K. Alharbi, F. H. Alshammari, and S. Kang, "A hybrid convolutional–transformer approach for accurate electroencephalography (EEG)-based Parkinson's disease detection," *Bioengineering*, vol. 12, no. 6, 2025.

[6] D. Martinez-Ramirez, A. Ramirez-Zamora, M. Rodríguez-Violante, and S.-M. Fereshtehnejad, *Managing Parkinson's Disease With a Multidisciplinary Perspective*. Frontiers Media SA, 2022.

[7] E. Kelechi Wisdom, T. Soyemi, S. Mayowa, N. S. Ede, E. Ubalaeze Solomon, C. A. Iloanusi, C. E. Agbo, O. Suzan Idogen, C. Augustine Ikechukwu, O. Clinton Ifeanyi *et al.*, "Building healthcare capacity for neurodegenerative disease management in Nigeria: Challenges and opportunities," *Journal of Public Health Research*, vol. 14, no. 2, p. 22799036251350957, 2025.

[8] M. B. T. Noor, N. Z. Zenia, M. S. Kaiser, S. A. Mamun, and M. Mahmud, "Application of deep learning in detecting neurological disorders from magnetic resonance images: a survey on the detection of Alzheimer's disease, Parkinson's disease and schizophrenia," *Brain informatics*, vol. 7, pp. 1–21, 2020.

[9] C. Bunterngchit, J. Wang, J. Su, Y. Wang, S. Liu, and Z.-G. Hou, "Temporal attention fusion network with custom loss function for EEG-fNIRS classification," *Journal of Neural Engineering*, vol. 21, no. 6, p. 066016, Nov. 2024.

[10] Q. Teng, Z. Liu, Y. Song, K. Han, and Y. Lu, "A survey on the interpretability of deep learning in medical diagnosis," *Multimedia Systems*, vol. 28, no. 6, pp. 2335–2355, 2022.

[11] F. Valente, S. Paredes, J. Henriques, T. Rocha, P. de Carvalho, and J. Morais, "Interpretability, personalization and reliability of a machine learning based clinical decision support system," *Data Mining and Knowledge Discovery*, vol. 36, no. 3, pp. 1140–1173, 2022.

[12] C. Bunterngchit, J. Wang, J. Su, Y. Wang, S. Liu, and Z.-G. Hou, "Enhanced cross-subject classification of hybrid EEG-fNIRS data using the simplified multimodal transformer network," in *Neural Information Processing*, M. Mahmud, M. Doborjeh, K. Wong, A. C. S. Leung, Z. Doborjeh, and M. Tanveer, Eds. Singapore: Springer Nature Singapore, 2025, pp. 300–313.

[13] S. Bouazizi and H. Ltifi, "Enhancing accuracy and interpretability in EEG-based medical decision making using an explainable ensemble learning framework application for stroke prediction," *Decision Support Systems*, vol. 178, p. 114126, 2024.

[14] M. C. Maurer, J. M. Metsch, P. Hempel, T. Bender, N. Spicher, and A.-C. Hauschild, "Explainable artificial intelligence on biosignals for clinical decision support," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 6597–6604.

[15] M. S. Kamal and S. F. Nimmy, "Interpretable transformers for Alzheimer disease diagnosis on multi-modal data," in *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024, pp. 1–8.

[16] S. Mahim, M. S. Ali, M. O. Hasan, A. A. N. Nafi, A. Sadat, S. Al Hasan, B. Shareef, M. M. Ahsan, M. K. Islam, M. S. Miah *et al.*, "Unlocking the potential of XAI for improved Alzheimer's disease detection and classification using a ViT-GRU model," *IEEE Access*, vol. 12, pp. 8390–8412, 2024.

[17] R. Alzoubi, A. Turky, A. Hussain, and S. Foufou, "Interpretable deep learning for Alzheimer's disease through genetic data and explainable artificial intelligence," in *2024 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT)*. IEEE, 2024, pp. 50–59.

[18] M. S. Kamal, A. Northcote, L. Chowdhury, N. Dey, R. G. Crespo, and E. Herrera-Viedma, "Alzheimer's patient analysis using image and gene expression data and explainable-AI to present associated genes," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–7, 2021.

[19] V. Viswan, N. Shaffi, M. Mahmud, K. Subramanian, and F. Hajamohideen, "Explainable artificial intelligence in Alzheimer's disease classification: A systematic review," *Cognitive Computation*, vol. 16, no. 1, pp. 1–44, 2024.

[20] H. A. Shad, Q. A. Rahman, N. B. Asad, A. Z. Bakshi, S. F. Mursalin, M. T. Reza, and M. Z. Parvez, "Exploring Alzheimer's disease prediction with XAI in various neural network models," in *TENCON 2021-2021 IEEE Region 10 Conference (TENCON)*. IEEE, 2021, pp. 720–725.

[21] Z. M. Leion, O. Jyoti, M. I. H. Abir, M. M. R. Kontho, I. Arafat, and H. I. Peyal, "An explainable AI approach using lightweight-CNN model for Alzheimer's disease detection," in *2024 27th International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2024, pp. 1588–1593.

[22] A. Mahdizadeh, P. A. Moghadam, S. Mirabbasi, and P. Nasiopoulos, "Explainable orthogonal attention networks for EEG-based analysis: Leveraging disentangled representations to enhance diagnosis," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[23] V. Vimbi, N. Shaffi, and M. Mahmud, "Interpreting artificial intelligence models: a systematic review on the application of LIME and SHAP in Alzheimer's disease detection," *Brain Informatics*, vol. 11, no. 1, p. 10, 2024.

[24] S. T. Siddiqui, Z. Syed, S. G. Hashmi, J. Ahmad, A. K. Singha, and K. A. Qidwai, "A wearable EEG-based hybrid CNN-RNN framework for psychiatric disorder diagnosis: Leveraging XAI for enhanced clinical insights," in *2025 Devices for Integrated Circuit (DevIC)*. IEEE, 2025, pp. 743–748.

[25] S. Khanapur, C. B. Bharadwaj, R. Bhardwaj, and J. S. Nayak, "An approach for XAI visualizations for explainability of Alzheimer's detection," in *2024 2nd International Conference on Networking, Embedded and Wireless Systems (ICNEWS)*. IEEE, 2024, pp. 1–6.

[26] D. Mansouri, A. Echtioui, R. Khemakhem, and A. B. Hamida, "Explainable AI framework for Alzheimer's diagnosis using convolutional neural networks," in *2024 IEEE 7th International Conference on Advanced Technologies, Signal and Image Processing (ATSIP)*, vol. 1. IEEE, 2024, pp. 93–98.

[27] N. Jackson, S. R. Cole, B. Voytek, and N. C. Swann, "Characteristics of waveform shape in Parkinson's disease detected with scalp electroencephalography," *eneuro*, vol. 6, no. 3, 2019.

[28] A. Miltiadous, K. D. Tzimourta, T. Afrantou, P. Ioannidis, N. Grigoriadis, D. G. Tsalikakis, P. Angelidis, M. G. Tsipouras, E. Glavas, N. Giannakeas *et al.*, "A dataset of scalp EEG recordings of Alzheimer's disease, frontotemporal dementia and healthy subjects from routine EEG," *Data*, vol. 8, no. 6, p. 95, 2023.