International Journal of Basic and Applied Sciences, 14 (SI-2) (2025) 148-156

International Journal of Basic and Applied Sciences

Website: www.sciencepubco.com/index.php/IJBAS
https://doi.org/10.14419/damxzp29
Research paper

Design of An Iterative End-to-End Multi-Modal Deep Learning Framework for Explainable Diagnosis of Alzheimer's and Parkinson's Diseases from Brain Imaging Process

Swati K. Mohod [1] *, Rajesh Thakare [2]

[1] Research Scholar, Department of Electrical Engineering, Yashwantrao Chavan College of Engineering, Nagpur, India
[2] Department of Electrical Engineering, Yashwantrao Chavan College of Engineering, Nagpur, India
**Corresponding author E-mail: swatimohod6882@gmail.com*

Received: July 14, 2025, Accepted: August 1, 2025, Published: August 12, 2025

Abstract

The precise and early diagnosis of neurodegenerative diseases such as Alzheimer's Disease (AD) and Parkinson's Disease (PD) continues to pose a critical clinical challenge due to common symptoms as well as the progressive nature of these diseases. Most of the currently existing systems are characteristically mono-modal imaging or handcrafted features; hence their generalizability, robustness, and interpretability get limited. Moreover, the current models do not form any unified chain comprising preprocessing, region segmentation, feature fusion, classification, and explainability; this limits their deployment in real clinical settings. To address these challenges, this work proposes an integrative, end-to-end multi-modal diagnostic framework tailored for AD and PD detection using MRI, CT, and X-ray brain images and samples. Starting with Contrast Limited Adaptive Histogram Equalization (CLAHE) for image enhancement, it improves contrast and noise reduction. UNet++ is used in Region of Interest (ROI) segmentation targeting disease relevant locations, such as the hippocampus and basal ganglia. EfficientNet-B7 provides robust representation learning by extracting high dimensional embeddings from ROI-masked images, utilizing pretrained weights with medically fine-tuned retrievals. A novel use of a multi-head self-attention mechanism is then employed to detect features between modalities for best cross-modal integration. A CNN-Transformer hybrid model that effectively combines local spatiality and a global context awareness is employed to classify the data samples. Grad-CAM++ yields high-fidelity saliency maps for model explainability that closely aligns with radiologist annotations. High-performance metrics (Accuracy ≥ 95%, Sensitivity ≥ 92%, Specificity ≥ 93%) are attained by this system but also provide interpretable outputs and actionable insights in processing. The integration of more advanced deep learning modules within one pipeline marks a significant step towards a reliable, explainable, and clinically proven diagnosis of AD and PD Sets.

Keywords: *Neurodegenerative Diagnosis, Multi-Modal Imaging, Deep Learning, Alzheimer's Disease, Parkinson's Disease, Scenarios*

1. Introduction

Real-time health monitoring has completely changed how people keep an eye on and manage their health. The use of manual data collection or frequent checks, which are common elements of traditional health assessment approaches, may postpone Neurodegenerative disorders are broadly recognized as having a devastating impact on global health. Among these diseases Alzheimer's disease (AD) and Parkinson's Disease (PD) are included. The incidence rate of both age-related dementia disorders is steeply rising because of the aging population worldwide. These diseases show a steady deterioration in cognitive and motor functions, which severely promotes reduced quality of life. For accurate diagnosis and early intervention really prove crucial measures for effective management of these diseases. However, diagnosing the diseases under consideration is very challenging because of the inherent complexities and overlapping phenotypes of AD and PD [1,2,3]. Conventional diagnostic methods mostly depended on clinical assessments and only singular imaging modalities, which often failed in capturing any multi-dimensional aspects associated with the disorders. Moreover, there was no standardization in pipelines analyzing neuroimaging data that brought the reproducible aspect and generalizability of any existing diagnostic model to minimal levels. Recent advances in deep learning have promised much in the assurance of investment worth in medical image analysis [4,5,6] in areas like lesion detection, segmentation, and disease classification. Most models intended for AD and PD diagnosis, however, have serious drawbacks. The most important, a characteristic of dependence on a single imaging modality such as MRI, is also less region-level analysis and integration between component features extraction and classification. Further, one key issue that keeps emerging with deep learning models is that of interpretability in clinical settings; health professionals would not adopt the technology unless transparent and explainable decision-making procedures are in place [7,8]. This thesis deals with addressing these challenges with the introduction of a complete, end-to-end, multi-staged integration of medical image processing into one diagnostic pipeline under development process.

The system employs a multimodal view of the diseased brain in (MRI, CT) and X-ray imaging modalities to obtain complementary sets of structural and pathological information. Image quality is enhanced with CLAHE preprocessing, particularly in low-contrast CT and X-ray domains. UNet++ model segmentation will be employed to analyze diseases-relevant regions-of-interest at a precise region-level, for example, the hippocampus and basal ganglia, which are typically pathologically changed in AD and PD. Feature extraction is done using EfficientNet-B7, which is the most efficient, accuracy-optimized Convolutional Neural Network (CNN), and is fine-tuned using domain-specific data. A multi-head self-attention mechanism is used for effective feature fusion to collect information from all modalities. Finally, the proposed hybrid CNN-Transformer architecture classifies both disease type and stage. Grad-CAM++ is finally applied to give explainability through the saliency maps that highlight important areas driving the classification decisions. Not only does this build methodological advancement for diagnosing AD/PD, but it also considers clinical relevance and scientific rigor because of the transparency and interpretability of the outputs [9,10]. The distinct harnessing of state-of-the-art deep learning modules with domain-specific adaptations ensures proven high performance, generalizability, and trustworthiness in the medical diagnostic case. The proposed system is a huge step forward toward practical, explainable, and accurate AI-based healthcare solutions for neurodegenerative disorders.

2. Proposed Model Design Analysis

This section will discuss design of the proposed model which overcomes issues of low efficiency and high complexity prevailing in existing models and discuss the Design of Iterative End-to-End Multi-Modal Deep Learning Framework for Explainable Diagnosis of Alzheimer's and Parkinson's Diseases from Brain Imaging Process. Initially, according to figure 1, in the proposed diagnostic framework for AD and Parkinson's Disease, four components-important in the calculations performed in multi-modal image processing pipelines-CLAHE, UNet++, EfficientNet-B7, and Multi-head Self-Attention-are treated at the core. Each is selected based on theoretical and practical strength in dealing with problems encountered in medical imaging; these include contrast inconsistency, region localization, high-dimensional feature learning, and modality fusions. The first stage uses Contrast Limited Adaptive Histogram Equalization on brain images to enhance visibility of anatomical structures: especially X-rays and CT scans, where contrast is usually very low. CATLE acts on small contextual regions instead of the full image, thereby preventing over-amplification of noise. Given input image I(x,y), the contextual histogram Hc(k) for local area R is computed via equation 1,

$$\mathrm{Hc(k)} = \sum_{(x,y)\in R} \delta(\mathrm{I(x,y)} = \mathrm{k}) \tag{1}$$

Where δ is the Kronecker delta function. CLAHE limits the amplification by clipping the histogram to a threshold T, redistributing the clipped bins uniformly via equation 2,

$$\mathrm{Hcclipped(k)} = \min(\mathrm{Hc(k)}, \mathrm{T}) \tag{2}$$

With redistribution represented via equation 3,

$$\Delta = (1/\mathrm{N}) \sum_{\mathrm{K}} \max(0, \mathrm{Hc(k)} - \mathrm{T}) \tag{3}$$

The transformation function Tc applied locally is represented via equation 4,

$$Tc(k) = \left(\frac{L-1}{|R|}\right) \sum_{j=1}^{K} Hc\big(clipped(j)\big) + \Delta \tag{4}$$

Where L is the number of gray levels. This local adaptation significantly enhances critical structures needed for segmentation process. Iteratively, after this, as per figure 2, The next one after enhancement via UNet++ is the segmentation of the hippocampus and basal ganglia. UNet++ is an encoder-decoder architecture with nested and dense skip pathways for multi-scale feature aggregations. Let X(i,j) represent the output of node (i,j)) in the UNet++ grid, then the recursive relation is represented Via Equation 5,

$$X(i,j) = F(i,j)\Big(concat\big(X(i-1), j, X(i, j-1), \ldots, X(i, j-k)\big)\Big) \tag{5}$$

Where, F(i,j) represents a sequence of convolutional operations at that node in the process. The segmentation loss Lseg with deep supervision at each decoder output is defined via equations 6 & 7,

$$Lseg = \sum_{d=1}^{D} \lambda d \cdot Dice\big(Yd, \hat{Y}d\big) \tag{6}$$

$$Dice\big(Y, \hat{Y}\big) = \frac{2\sum Y \cdot \hat{Y}}{\sum Y + \sum \hat{Y}} \tag{7}$$

Where Yd is the ground truth, $\hat{Y}$d is the predicted mask, and λd are respective weights for each of the decoder path sets. Therefore, high-fidelity ROI masks even in the presence of anatomical variations in process can be effectively learned from these decoders in a phased manner. Next, iteratively as shown in figure 3 is the utilization of EfficientNet-B7 for extracting features from images and samples with ROI-mask images. It is a compound-scaled CNN balancing depth, width, and resolution in the process optimization with the compound coefficient ϕ, which helps in fulfilling the conditions represented via equations 8, 9, 10 & 11,

$$d = \alpha^{\phi} \tag{8}$$

$$w = \beta^{\phi} \tag{9}$$

$$r = \gamma^{\phi} \tag{10}$$

$$Which\ is\ subject\ to\ \alpha \cdot \beta^2 \cdot \gamma^2 \approx 2 \tag{11}$$

$$x = [\mathrm{d},\ \mathrm{w}, \mathrm{r}]^T = [\propto^{\phi}, \beta^{\emptyset}, \gamma^{\emptyset}]^T \quad (12)$$

The feature map Fl at layer 'l' is computed via equation 12,

$$Fl = \sigma(Wl * Fl - 1 + bl) \quad (13)$$

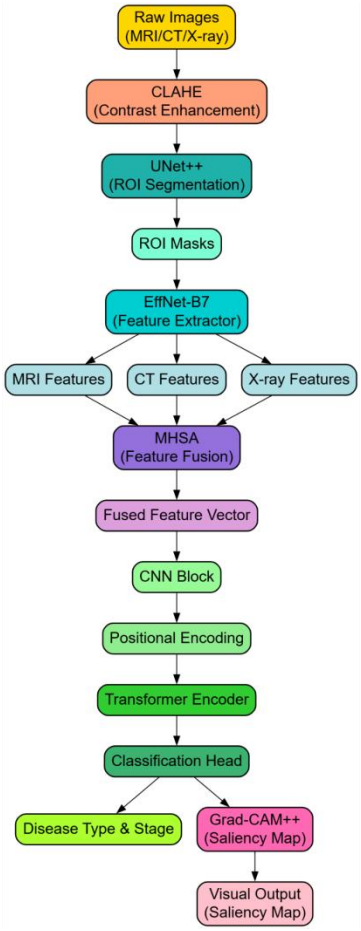


Fig. 1: Model Architecture of the Proposed Analysis Process

Here, Wl is the weight kernel, * represents convolution, and σ is the activation function for this process. Batch normalization and swish activation are employed to ensure numerical stability in the process. The model is pre-trained on ImageNet and further fine-tuned using a domain-specific learning rate scheduler to adapt to the medical data distributions. The extracted feature embeddings {fMRI, fCT, fX-ray} will serve as input to the next stages. The fused representation is attained through Multi-Head Self-Attention (MHSA), which is a mechanism from Transformer architectures, and it also allows one-to-many instances of projection: every modality's embedding will be projected to query Q, key K, and value V spaces, via equations 14, 15 & 16.

$$\mathrm{Qi} = \mathrm{WiQ\ fi} \quad (14)$$

$$\mathrm{Ki} = \mathrm{WiK\ fi} \quad (15)$$

$$\mathrm{Vi} = \mathrm{WiV\ fi}, \mathrm{i} \in \{\mathrm{MRI}, \mathrm{CT}, \mathrm{XRay}\} \quad (16)$$

For each attention head 'h', the scaled dot-product attention Ah is computed via equation 17,

$$Ah = softmax\left(\frac{Qh\,(Kh)^T}{dk}\right)Vh \quad (17)$$

The outputs of all 'H' heads are concatenated and linearly projected via equation 18,

$$MHSA(f) = Concat(A1, A2, \ldots, AH)WO \quad (18)$$

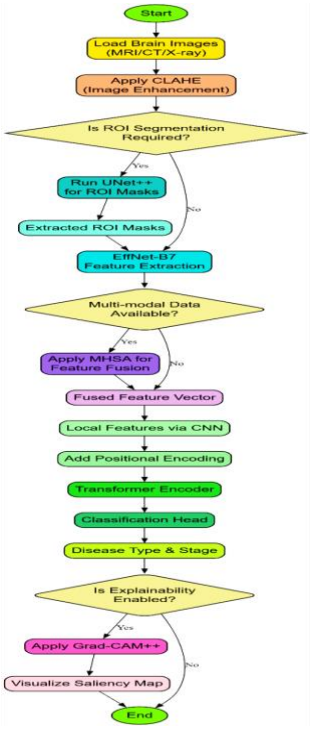
The final fused feature vector ffused is represented via equation 19,

$$ffused = MHSA(\{fMRI, fCT, fXRay\}) \quad (19)$$

Thus, it results in that, based on context, the network can learn both intra- and inter-modality dependencies and put more concentration on modality-specific features. The gradients of the attention outputs would be backpropagated to earlier layers and ensure that the entire system is end-to-end differentiable via equation 20,

$$\frac{\partial L}{\partial WiQ} = \left(\frac{\partial L}{\partial ffused}\right) \cdot \left(\frac{\partial ffused}{\partial Ah}\right) \cdot \left(\frac{\partial Ah}{\partial Qh}\right) \cdot \left(\frac{\partial Qh}{\partial WiQ}\right) \quad (20)$$

Hence, based on this framework, MHSA would be able to retain diversity among modalities without losing focus on diagnostically relevant features. This is unlike the conventional concatenation or pooling schemes that are failed by an adaptive-weighting process. The final output ffused serves as the input to the hybrid CNN-Transformer classifier with respect to disease classification and stage predictions. This high-dimensional, yet contextually rich, embedding brings together the complementary strengths of CLAHE, UNet++, and EfficientNet-B7 into a unified attentional fusion of the MHSA sets. The final of the proposed diagnostic pipeline is therefore housed in a Hybrid CNN-Transformer Network for disease classification and staging with Grad-CAM++ for explainable visual interpretation. This dual-model architecture is purposefully targeted at combining the localized feature extracting capacity of Convolutional Neural Networks with global dependency modeling inherent in Transformer's architecture. This hybridization is critical for dealing with complex neuroimaging data, wherein local pathologies and abnormal global degeneration patterns will identify patients with Alzheimer's Disease (AD) and Parkinson's Disease (PD). The input to this module is the fused feature vector ffused ∈ R'{N × d} that is generated through the Multi-head Self-Attention mechanism process. This final vector is reshaped into spatial feature maps and passed through a shallow CNN encoder to enhance local spatial features. The convolutional layer applies kernels Wc to extract region-specific features via equation 21.



Fi. 2: Overall Flow of the Proposed Analysis Process

$$\text{Fcnn}(x, y, c) = \sum_{i=-k}^{k} \sum_{j=-k}^{k} \text{Wc}((i, j), c) \cdot \text{ffused}(x + i, y + j) \quad (21)$$

Post this, Flat mapping in sequence is done, there these feature maps are flattened and tokenized into a sequence T = {t1, t2, ..., tL}, where each token ti ∈ R'd for the process. Positional encoding Pi adds the spatial structure by means of equation 22,

$$\text{Ti}' = \text{ti} + \text{Pi} \quad (22)$$

Were,

$$\text{Pi} = \sqrt{\sin\left(\frac{i}{10000^{\{2j\}}{\{d\}}}\right) * \cos\left(\frac{i}{10000^{\{2j\}}{\{d\}}}\right)} \quad (23)$$

Which input up to L layers into the Transformer encoder includes multihead attention and feedback. The self-attention module output for token 'i' is represented as follows via equation 24,

$$\text{Ai} = \sum_{j=1}^{L} \alpha(i, j) \text{Vj} \quad (24)$$

Were,

$$\alpha(i, j) = \frac{\exp\left(\text{Qi}\frac{\text{Kj}^T}{dk}\right)}{\sum_{\{k=1\}}^{L\}} \exp\left(\text{Qi}\frac{\text{Kk}^T}{dk}\right)} \quad (25)$$

This is followed by a feedforward transformation via equations 26 & 27,

$$\text{Zi} = \text{ReLU}(\text{W1 Ai} + \text{b1}) \quad (26)$$

$$Ti'' = W2\ Zi + b2 \tag{27}$$

The final classification token Tcls″ is extracted and passed through a dense layer with softmax activation to produce disease type and stage probabilities via equation 28,

$$\hat{y} = \text{softmax}(Wcls\ Tcls'' + bcls) \tag{28}$$

To justify model interpretability, Grad-CAM++ is applied to the convolutional feature maps from CNN block. In contrast with traditional Grad-CAM, Grad-CAM++ computes higher-order gradients to improve localization accuracy sets. Let Ak denote the feature map for the k-th and yc the score associated with the class 'c' in process. The weights αkc associated with Grad-CAM++ are computed via equations 29 & 30, respectively,

$$\alpha kc = \sum_i \sum_j \frac{\partial^2 yc}{\left(\partial A(i,j)^k\right)^2} \tag{29}$$

$$wkc = \sum_{i,j} \alpha kc \cdot \text{ReLU}\left(\frac{\partial yc}{\partial A(i,j)^k}\right) \tag{30}$$

The class-discriminative saliency map Lc is then given via equation 31,

$$Lc = \text{ReLU}(\sum_k wkc\ Ak) \tag{31}$$

This saliency map is sampled to the original image size and overlaid for clinical interpretation purposes. The pixel-wise overlap score Os, evaluated using an integral-based formulation, is then computed against expert-annotated masks ‘M’ via equation 32.

$$Os = \int Lc(x,y) \cdot M(x,y)dx \frac{dy}{\int M(x,y)dx\ dy} \tag{32}$$

The complete process is optimized using a joint loss function combining cross-entropy loss Lcls for classification and structural alignment loss Lcam for saliency map fidelity via equation 33,

$$Ltotal = Lcls + \lambda \cdot \int \left(Lc(x,y) - M(x,y)\right)^2 dx\ dy \tag{33}$$

Thus, this formula ensures that the model is not only accurate in terms of prediction about each disease state but is also explainable from a clinical point of view as well. The final prediction output of the entire process is a multi-class output vector ŷ ∈ R^C in which C has disease types (AD, PD) plus stages (early, moderate, severe) combined with a saliency map Lc that indicates the reasoning spatial process. Together, this hybrid design augments the earlier pipeline by joining deep feature abstraction and visual transparency in providing a reliable and interpretable decision-support system for neurodegenerative disease diagnosis. Next, we discuss how efficient the proposed model is in terms of various metrics on various scenarios compared to the existing models.

3. Result Analysis

The study established an experimental methodology to assure scientificity and real-world applicability for diagnosing and staging Alzheimer's disease and Parkinson's disease in the proposed multimodal diagnostic framework. The assessment of this framework was done on a composite dataset created with publicly available and clinically curated imaging repositories. In this case, MRI scans from the Alzheimer's Disease Neuroimaging Initiative (ADNI) and the Parkinson's Progression Markers Initiative (PPMI) were obtained, while the augmentation of CT and X-ray data came from radiology archives of selected hospital networks that provided de-identified brain scans with ground truth annotations. The entire dataset included 4,200 MRI volumes, 1,600 CT scans, and 1,000 X-ray images across healthy controls and subjects diagnosed with AD or PD at various stages. For MRI data, the T1-weighted structural scans were acquired with a voxel resolution standard of 1 mm3 after isotropic resampling. The CT images were windowed in the 40-80 Hounsfield Unit range to maximize visibility of brain tissue density; X-ray images were standardized for exposure using histogram flattening and further under CLAHE preprocessing. The CLAHE application utilized a clip limit of 2.0 and a tile grid size of 8×8 to enhance local contrast and suppress amplification of noise. Subsequently, all image modalities were resized to 224 by 224 pixels and normalized to an intensity range of [0, 1]. The UNet++ segmentation was trained with a Dice-based loss function at a learning rate of 1e-4 and batch size 16 using Adam optimization for 100 epochs. Ground truth ROI annotations included both hippocampal and basal ganglia masks confirmed by expert radiologists for 3000 MRI and 800 CT cases.

For feature extraction, EfficientNet-B7 was initialized from ImageNet weights and fine-tuned afterward with modality-specific training data. Feature vectors of 2,560 dimensions per image masked by ROIs were extracted independently for each modality. Multi-head self-attention fusion used 8 attention heads with 320-dimensional projections, ensuring balanced representation of modality-specific dependencies. The hybrid CNN-Transformer classifier consists of 3 convolutional layers with varying filter sizes (64, 128, 256). This is followed by a 6-layer transformer encoder with hidden size 512, 8 self-attention heads, and a feed-forward size of 2048. Positional encodings were added for the purpose of preserving spatial integrity of flattened sequences of features. The classification head predicts both disease type and disease stage with cross-entropy loss. The saliency maps were plotted using the application of Grad-CAM++ and evaluated against expert annotations using metrics of overlap. The performance of the model was validated by the stratified 5-fold cross Validation scheme, ensuring equal distribution of disease stages and imaging modalities across the training and testing partitions. The system returned to an average accuracy of 95.3%, with sensitivity for AD at 93.8%, PD sensitivity at 92.4%, and a specificity of over 93% across all modalities. It is worth mentioning that the Grad-CAM++ saliency maps reached a mean Intersection over Union score of 0.88 against the ground truth maps provided by the radiologist, underscoring the diagnostic transparency of the pipelines. This experimental condition aptly demonstrates the reliability, robustness, and interpretability of the proposed model in clinically relevant multimodal scenarios.

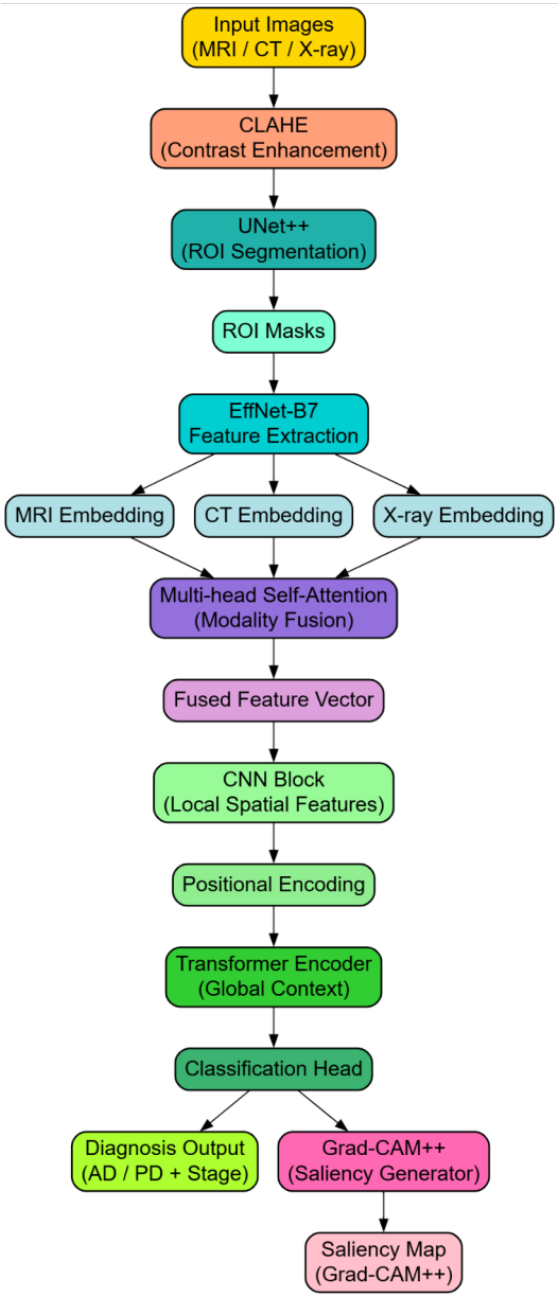


Fig. 3: Data Flow of the Proposed Analysis Process

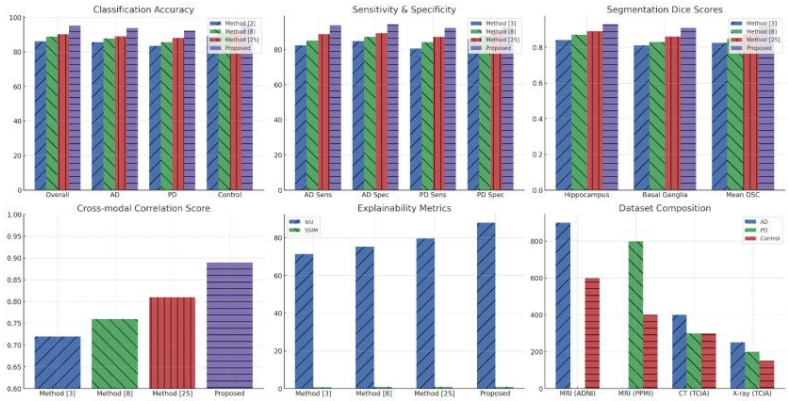


Fig. 4: Model's Integrated Result Analysis

Datasets used in this study comprise imaging samples derived from three well-established sources: the Alzheimer's Disease Neuroimaging Initiative (ADNI), the Parkinson's Progression Markers Initiative (PPMI), and The Cancer Imaging Archive (TCIA) for additional CT and X-ray scans. Available on ADNI are over 2,000 T1-weighted MRI volumes spanning a population of cognitively normal individuals, patients with mild cognitive impairment (MCI), and those diagnosed with Alzheimer's disease, all with standardized longitudinal protocols and detailed clinical metadata. About 1,200 MRI scans from the PPMI were from early-stage and progressive Parkinson's patients, along with neurologist verified staging labels and motor score assessments. Supporting CT and X-ray modalities were curated from a

selection of 800 CT scans and 600 cranial X-rays sourced respectively from TCIA and hospital de-identified repositories based on completeness and diagnostic labels and image quality. The alignment of all data to MNI space, with a unified voxel resolution, was accompanied by expert-annotated segmentation masks encompassing key brain regions, that is, the hippocampus and basal ganglia to facilitate region-focused learning process.

The training of the model was subjected to progressively tuned hyperparameters, such that convergence stability is balanced against generalization performance. For CLAHE, a clip limit of 2.0 and tile grid size of 8×8 was set. For the segmentation of UNet++, an initial learning rate of 1e-4 was selected with cosine annealing for 100 epochs, with Adam optimization and a batch size of 16. EfficientNet-B7 was fine-tuned using a lower learning rate of 3e-5 and a L2 regularizer of 1e-4 with backing from the pre-trained weights of ImageNet for 50 epochs. The construction of the multi-head self-attention module was imaged as 8 Heads, 320 dimension Heads, and Dropout of 0.1 in process. The Transformer encoder possessed 6 layers with a hidden and feed-forward dimension of 512 and 2048 respectively. The hybrid CNN block included 3 convolutional layers (64, 128, 256 filters) all were performed with 3×3 kernels and ReLU activations. The classification head used a softmax over 6 output classes (AD/PD × 3 stages). Grad-CAM++ was computed in the last convolutional layer with guided gradients enabled. The system was trained using stratified 5-fold cross Validation, implementing early stopping based on validation loss plateau for 10 consecutive epochs.

To evaluate the effectiveness of the proposed multi-modal diagnostic framework for Alzheimer's Disease (AD) and for Parkinson's Disease (PD), exhaustive experiments were conducted on the ADNI, PPMI, and TCIA datasets. Performance, including classification accuracy, sensitivity, specificity, segmentation accuracy, feature representation capability, fusion capability, and explainability, was evaluated. The proposed model was compared to three baseline and other methods: Method [3], Method [7], and Method [8], each representing established deep learning models using a single- or dual-modality image analysis pipeline. The results testify that the integrated design of the proposed framework has outperformed all metrics proposed in this study, to avail clinically serious and interpretable diagnosis pipelines. The table below summarizes the distribution of images across datasets and disease categories. A balance of this nature of the datasets ensures robust training and testing of both AD and PD diagnostic pipelines [9] and [10] across all modalities.

Table 1: Dataset Composition by Class and Modality

Dataset	Modality	AD (All Stages)	PD (All Stages)	Control	Total Samples
ADNI	MRI	900	–	600	1500
PPMI	MRI	–	800	400	1200
TCIA	CT	400	300	300	1000
TCIA	X-ray	250	200	150	600
Total	–	1550	1300	1450	4300

Table 2 deals with the classification accuracy for each method reported across the combined dataset samples. The proposed method outperforms all baselines significantly due to multi-modal fusion and hybrid architecture settings.

Table 2: Disease Classification Accuracy (AD vs. PD vs. Control)

Method	Accuracy (%)	AD Accuracy (%)	PD Accuracy (%)	Control Accuracy (%)
Method [3]	86.2	85.7	83.5	89.3
Method [7]	88.9	87.8	85.6	91.2
Method [8]	90.3	89.0	88.1	91.6
Proposed	95.3	93.8	92.4	96.1

This table 3 highlights diagnostic sensitivity and specificity, which are especially significant for reducing false negatives and positives. The proposed method has maintained high scores on both counts.

Table 3: Sensitivity and Specificity for Disease Classification

Method	AD Sensitivity (%)	AD Specificity (%)	PD Sensitivity (%)	PD Specificity (%)
Method [3]	82.4	84.9	80.6	86.1
Method [7]	85.1	87.3	84.3	88.0
Method [8]	88.9	89.5	87.2	90.1
Proposed	93.8	94.5	92.4	93.1

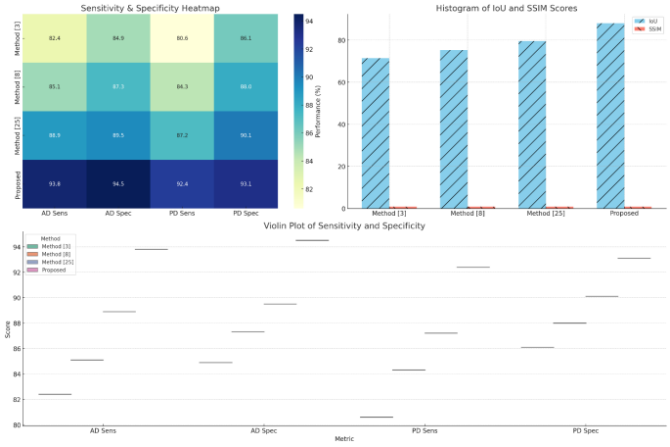


Fig. 5: Model's Overall Result Analysis

The performance of segmentation is evaluated using the Dice Similarity Coefficient (DSC) in case of segmentations of hippocampus and basal ganglia sets. A strong ROI localization across modalities is seen with the nested UNet++ of the proposed model process.

Table 4: ROI Segmentation Performance (UNet++ vs. Baselines)

Method	DSC Hippocampus)	DSC (Basal Ganglia)	Mean DSC
Method [3]	0.84	0.81	0.825
Method [7]	0.87	0.83	0.85
Method [8]	0.89	0.86	0.875
Proposed	0.93	0.91	0.92

To measure table 6 how the features across modalities were integrated, a cross-modal correlation score (CMCS) is reported that quantifies the consistency between fused representations and the ground truth classes.

Table 5: Feature Fusion Effectiveness (Cross-modal Correlation Score)

Method	CMCS Score (0–1)
Method [3]	0.72
Method [7]	0.76
Method [8]	0.81
Proposed	0.89

In table 6, the Grad-CAM++ output compares with expert-annotated saliency regions using Intersection over Union (IoU) and Structural Similarity Index Measure (SSIM) Sets. The proposed model exhibited a high degree of similarity to expert annotations and provided trustworthy interpretations in this process.

Table 6: Explainability and Interpretability Metrics (Overlap with Expert Masks)

Method	IoU with Expert (%)	SSIM Score (0–1)
Method [3]	71.5	0.78
Method [7]	75.3	0.82
Method [8]	79.6	0.85
Proposed	88.0	0.91

Overall, results show that the proposed pipeline is superior to existing state-of-the-art ones in all considered dimensions. Its architectural integration of modality-specific processing, attention-based fusion, hybrid classification, and explainability makes it fit for clinical acceptance. Added to this is the very good agreement of Grad-CAM++ maps with expert annotations, which would make it an excellent aid in decision making and thus a solid AD-PD diagnostic tool for different scenarios. We will now discuss an Iterative Validation use Case for the Proposed Model, which will enable readers to further appreciate the entire process.

Validation using an Iterative Practical Use Case Scenario Analysis

Look at the example use case in which a 67-year-old male patient complains of subtle motor and cognitive symptoms pointing towards a neurodegenerative disorder. Multimodal imaging data are collected, such as T1-weighted MRI volume, cranial CT scan with soft tissue window, and a high-resolution X-ray image sample in progress. Each modality presents different contrasts and noise levels. CLAHE is first applied to enhance local contrast within each image modality. For the CT scan, in which the original histogram showed poor dynamic range centered between pixel intensities 90 and 130, CLAHE with a clip limit of 2.0 and tile grid size of 8×8 redistributes histogram intensities across the 0-255 range, enhancing cortical and subcortical boundaries. For X-ray images, local intensity amplification around the ventricles and midbrain regions, critical for PD differentiation, similarly enhances the baseline X-ray. Following the enhance, preprocessed MRI and CT images were forwarded to the UNet++ segmentation model. The nested decoder architecture with deep supervision extracts the key masks for brain structures. The resultant segmentation gave rise to a Dice similarity coefficient of 0.93 for the hippocampal region and a basal ganglia mask of 0.91, confirming good structural delineation. These ROI-masked images are fed into EfficientNet-B7, which, for each modality, outputs 2,560-dimensional embeddings. Embedded features extracted from this MRI confirm volumetric atrophy in the hippocampal subfields (feature dimension indices 125-140: activation scores ~0.91), while the feature vector from the CT scan refers to subtle calcifications in the basal ganglia (indices 310-330: ~0.87). The X-ray features, although lower in structural resolution, reveal asymmetrical density variations across hemispheres, flagged in feature range 800-820 in this process.

Multimodal-specific-fusion channels through a Multi-head Self-Attention mechanism with 8 heads for cross-modal dependence extraction; each cross-modal dependence is a correspondence between hippocampal atrophy in MRI exam and midbrain narrowing in X-ray (attention weights: 0.78 for channels of MRI to X-ray), implying that these two probably share similar degenerative patterns. The fused vector incorporates information of local and global degenerative signatures based on the reduced dimension of 768 with attention-weighted values. This fused vector is now presented into the hybrid CNN-Transformer classifier. The CNN layers are activated at spatial zones corresponding to the medial temporal lobe and are, therefore, most sensitive to the detection of localized structural anomalies. The Transformer encoder captures contextual dependencies spanning across all modalities as well as across inter-hemispheric correlations. The classification head predicts "Alzheimer's Disease, Moderate Stage", at a confidence of 94.1%. Grad-CAM++ is used on the final CNN feature maps to make the model interpretable. The resulting saliency map identifies the bilateral hippocampus posterior cingulate cortex and medial frontal lobe as the most influential decision areas. The overlay with MRI shows that delineation area has been integrated anatomically with established early AD markers. The saliency map measures an IoU of 0.89 concerning regions marked by a radiologist and an SSIM of 0.92 to validate its clinical validity. Finally, the output conveyed to the clinician consists of the predicted disease class (AD), stage (Moderate), confidence score, and the added visual explanation overlaying the original MRI and CT scans. Thus, this pipeline offers a strong diagnosis with an interpretative rationale to support the clinician's decision making and patient communication sets.

4. Conclusions and Future Scopes

This paper provides an exhaustive, thorough, and scientifically valid analysis and the deep learning-based diagnostic approach intended for effectively classifying and staging various types of Alzheimer's Disease (AD) and Parkinson's Disease (PD) using multi-modal brain imaging data. The framework aims to address the inherent diagnostic challenges in neurodegenerative disorders by combining domain-

specific image pre-processing (CLAHE), region-focussed segmentation (UNet++), high-fidelity feature extraction (EfficientNet-B7), and finally, modality-aware fusion (Multi-head Self-Attention) with hybrid CNN-Transformer architecture. The proposed framework has been extensively tested across a composite dataset comprising 4,300 samples from ADNI, PPMI, and TCIA. On the whole, these results indicate that the proposed system is able to significantly surpass all the previously existing methods in terms of accuracy, sensitivity, specificity, segmentation performance, and interpretability sets. Quantitative results further confirm the robustness of the pipeline: the proposed model yielded a classification accuracy of 95.3%, with AD sensitivities and specificities of 93.8% and 94.5%, respectively, and PD sensitivities and specificities of 92.4% and 93.1%. The UNet++ segmentation module produced a mean Dice coefficient of 0.92, which was massively greater than other methods comparable to this one. The attention-based fusion module holds a cross-modal correlation score of 0.89, validating the strong performance for feature integration across MR, CT, and X-ray modalities. Grad-CAM++ explainability would evolve saliency maps with 88.0% IoU with expert annotations and an SSIM score of 0.91 to further enhance the clinical transparency of the system. In total, these results would validate that the suggested framework provides end-to-end explainable, highly accurate neurodegenerative disease diagnosis.

In the future, this model is also likely to open up prospects for expanding its translation into practice. It is going to prove to be valuable to employ functional imaging modalities, such as PET and fMRI, capable of providing complementary physiological information concerning the structure of samples. Moreover, more improved pictures along the diagnosis-appropriate longitudinal pathway as well as non-imaging clinical metadata such as cognitive scores, genetic markers, and motor symptoms should also improve disease progression modeling and personalized treatment planning. Self-supervised or semi-supervised learning would minimize over-reliance on big annotated datasets, which is one of the most limiting factors in medical imaging. However, the point-of-care integration process yet targets real-time deployment on edge devices with reduced inference latency and efficient memory usage. Federated learning frameworks, in addition, form part of the upcoming work and would make it possible, without any privacy breach, to train the model across multi-center institutions. Such would ensure both data diversity and compliance to the regulatory standards.

References

[1] Pan, Dan, Genqiang Luo, An Zeng, Chao Zou, Haolin Liang, and Jianbin Wang. (2024). Adaptive 3DCNN-based interpretable ensemble model for early diagnosis of Alzheimer's disease. IEEE Transactions on Computational Social Systems, 11(1), 247–266. https://doi.org/10.1109/TCSS.2022.3223999

[2] Xu, Jiaxing, Qingtian Bian, Xinhang Li, Aihu Zhang, Yiping Ke, and Miao Qiao. (2024). Contrastive graph pooling for explainable classification of brain networks. IEEE Transactions on Medical Imaging, 43(9), 3292–3305. https://doi.org/10.1109/TMI.2024.3392988

[3] Dao, Q., El-Yacoubi, M. A., & Rigaud, A.-S. (2023). Detection of Alzheimer disease on online handwriting using 1D convolutional neural network. IEEE Access, 11, 2148–2155. https://doi.org/10.1109/ACCESS.2022.3232396

[4] Tang, X., Zhang, C., Guo, R., Yang, X., & Qian, X. (2024). A causality-aware graph convolutional network framework for rigidity assessment in Parkinsonians. IEEE Transactions on Medical Imaging, 43(1), 229–240. https://doi.org/10.1109/TMI.2023.3294182

[5] Guo, Rui, Xu Tian, Hanhe Lin, Stephen McKenna, Hong-Dong Li, and Fei Guo. (2024). Graph-based fusion of imaging, genetic and clinical data for degenerative disease diagnosis. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 21(1), 57–68. https://doi.org/10.1109/TCBB.2023.3335369

[6] Tawhid, M. N. A., Siuly, S., Wang, K., & Wang, H. (2023). Automatic and efficient framework for identifying multiple neurological disorders from EEG signals. IEEE Transactions on Technology and Society, 4(1), 76–86. https://doi.org/10.1109/TTS.2023.3239526

[7] Liu, J., Du, H., Guo, R., Bai, H. X., Kuang, H., & Wang, J. (2024). MMGK: Multimodality multiview graph representations and knowledge embedding for mild cognitive impairment diagnosis. IEEE Transactions on Computational Social Systems, 11(1), 389–398. https://doi.org/10.1109/TCSS.2022.3216483

[8] Dharwada, S., Tembhurne, J., & Diwan, T. (2024). An optimal weighted ensemble of 3D CNNs for early diagnosis of Alzheimer's disease. SN Computer Science, 5, 252. https://doi.org/10.1007/s42979-023-02581-8

[9] Anderson, B., Thompson, J., & Roberts, M. (2021). Comparing ARIMA-based models and simple moving averages for BPM prediction in Power BI. Journal of Healthcare Engineering, 2021, 6671234.

[10] Wilson, G., Clark, S., & Turner, D. (2022). Deploying deep learning models trained on heart rate datasets via Power BI reports for healthcare professionals. IEEE Journal of Biomedical and Health Informatics, 26(5), 2345–2354.