SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Vision Transformer-Driven Multi-Modal Medical Imaging Framework for Early Diagnosis of Neurodegenerative Diseases

M. Padmapriya[1], Haritha Akkineni[2], M. Jeevana Sujitha[3], Lavanya Gottemukkala [4], Para Rajesh[5]

[1]Assistant Professor, School of Computing SRM Institute of Science and Technology Tiruchirappalli.
E-mail: padmapriya.m@ist.srmtrichy.edu.in, ORCID:0000-0002-3656-7477
[2]Associate Professor, Department of CSE (Data Science), PVP Siddhartha Institute of Technology, Kanuru, Vijayawada, Andhra Pradesh-520007. E-mail: aharitha@pvpsiddhartha.ac.in, ORCID: 0000-0002-5909-4078
[3]Assistant Professor, Computer Science and Engineering, SRKR Engineering college, Chinaamiram, Bhimavaram, Andhra Pradesh, INDIA, -534202. E-mail: jeevanasujitha@gmail.com, ORCID:0000-0002-9753-752X
[4]Assistant Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Bowrampet, Hyderabad – 500043, Telangana, India. E-mail: g.lavanya@klh.edu.in, ORCID: 0000-0002-1866-1582
[5]Assistant Professor, Department of Computer Science and Engineering, VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, Telangana-500090, India. E-mail: pprr21@gmail.com, ORCID: 0000-0001-9056-5918

Abstract
Accurate diagnosis of neurodegenerative diseases such as AD and PD in their early stages remains a critical challenge in modern medicine, due to the subtle and overlapping structural and functional changes in the brain during early stages. Conventional diagnostic methods, which rely on single modality imaging or subjective clinical evaluation, are not sensitive enough for early detection. We introduce a new multi-modal medical imaging framework based on Vision Transformers (ViTs) to improve early diagnosis of neurodegenerative diseases. The proposed framework is designed by integrating data from MRI and PET and exploiting the advantages of the strong self-attention of ViTs for modelling the overall contextual relationship and local structural anomalies. The multi-modal fusion strategy combines structural variations from MRI and metabolic activity from PET in a meaningful way, resulting in a comprehensive representation of brain pathology. The ViT-driven multi-modal framework outperforms the conventional Convolutional Neural Network (CNN)-based and single-modality methods in terms of classification accuracy, sensitivity and specificity, as demonstrated by experiments on publicly available datasets such as the Alzheimer's Disease Neuroimaging Initiative (ADNI) and the Parkinson's Progression Markers Initiative (PPMI). These results highlight the potential of Vision Transformers for multi-modal medical image analysis and provide a promising tool for early clinical intervention of neurodegenerative diseases.

Keywords: Vision Transformers, Multi-Modal Medical Imaging, Neurodegenerative Diseases, Alzheimer's Disease, Parkinson's Disease, Deep Learning, Medical Image Fusion.

Introduction

Neurodegenerative diseases (NDDs) including Alzheimer's Disease (AD) and Parkinson's Disease (PD), are one of the largest health crises in the world characterised by progressive degeneration of structure and function of the central nervous system or peripheral nervous system. World populations are ageing, increasing the incidence of these conditions and presenting major emotional, physical and economic challenges for patients, carers and health-care systems globally (Erkkinen et al., 2018). Current clinical interventions have been largely symptomatic and disease-modifying therapies are lacking or most efficacious when administered in the earliest, often asymptomatic stages of the disease. Therefore, early and accurate diagnosis of NDDs is needed for early intervention, better patient outcome and appropriate trial candidate selection.

Imaging is a useful non-invasive tool in the assessment of brain pathology. MRI gives a high-resolution structural view of the brain, showing brain atrophy, cortical thinning and structural abnormalities relating to disease progression (Frisoni et al., 2010).

However, Positron Emission Tomography (PET) offers information on functions and metabolism like decrease in glucose metabolism and depositions of amyloid or tau proteins that are often preceded by structural changes (Vlassenko et al., 2012). Diagnosis is often made with single-modality imaging, but this does not necessarily reflect the entire spectrum of pathologic change. Structural imaging is insensitive to early metabolic change, and functional imaging has limited anatomical localisation. Therefore, multi-modal imaging, which integrates complementary information from multiple modalities (e.g., MRI and PET), has been a promising approach to improve the diagnostic accuracy and robustness (Hao et al., 2020).

1. In recent years, deep learning (DL) with Convolutional Neural Networks (CNNs) has revolutionised medical image analysis by automatically learning hierarchical features from complex imaging data (Litjens et al., 2017). Some CNN based models have been proposed to classify NDDs from multi-modal data and they have shown

huge improvements over traditional machine learning methods (Jo et al., 2019; Wen et al., 2020). But CNNs have limitations. The local receptive fields of these models limit their ability to learn long-range dependencies and global contextual information, which are important for understanding the distributed network alterations that characterise many NDDs. In addition, the effective fusion of multi-modal features in CNN architectures is still a challenge, which generally adopts simple concatenation or addition operations that may not be optimal for the alignment and integration of complementary information.

To solve these problems, recently, Vision Transformers (ViTs) have been suggested for computer vision tasks by using the self-attention mechanisms of Natural Language Processing (Dosovitskiy et al., 2020). ViTs split an image into a set of patches and feed them into a transformer encoder. This enables the model to learn global dependencies and complex relations between the different image regions already at the lowest layers. Such a global receptive field makes ViTs particularly promising for medical image analysis where subtle pathological changes could be distributed thru the brain (Shamshad et al., 2023). Recently, ViTs have been explored for medical applications such as organ segmentation, tumour detection, and disease classification (Hatamizadeh et al., 2022).

However, ViTs potential in the multi-modal medical imaging for the neurodegenerative diseases setting is relatively under explored. So that we can effectively fuse features from different modalities (e.g. structural MRI, functional PET) via transformer architectures, we need some fusion strategies to explore cross-modal correlations and remove modality specific noises. To this end, we propose a new Vision Transformer Driven Multi-Modal Medical Imaging Framework for early diagnosis of NDDs. We propose a framework that extracts rich global contextualised feature representation of MRI and PET images using independent ViT encoders. Then a cross-attention based fusion module is utilised to effectively fuse the multi-modal features, which enables the model to pay attention to the salient and complemental information in both structural and functional domains. The main contributions of this paper can be summarised as follows:

1. We propose a novel multi-modal approach with Vision Transformers for global context modelling for early diagnosis of neurodegenerative diseases using MRI and PET data.
2. We propose a cross-attention fusion module for adaptively fusing the structural and functional features, which can effectively model the complex inter-modal relationships and improve the representation capability of the fused features.
3. Experiments on benchmark datasets (ADNI and PPMI) demonstrate that the proposed framework achieves the state-of-the-art performance for early-stage Alzheimer's and Parkinson's disease classification compared to traditional CNNs and baseline transformer models.

2. Related Work

2.1 Deep Learning in Neurodegenerative Disease Diagnosis

Deep learning has been applied to medical image analysis with exponential growth, especially for neurodegenerative disease diagnosis. Early approaches relied mostly on 2D CNNs processed slice by slice. For example, Sarraf and Tofighi (2016) used a modified LeNet architecture to classify the Alzheimer's disease with functional MRI data. But 2D models lose important context of the spatial information in the z axis. In later studies, 3D CNNs were used to incorporate 3D spatial information. Payan and Montana (2015) proposed a 3D CNN based on sparse autoencoders to predict the Alzheimer's disease status from structural MRI, showing important improvements over conventional methods. Similarly, Cho et al. (2021) demonstrated the efficacy of 3D CNNs with DaTscan imaging for Parkinson's disease discriminating PD patients and healthy controls. Although CNNs have been successful, they are limited by their local receptive fields, which may not be suitable for the dispersed neuroanatomical alterations in complex brain disorders (Wen et al., 2020).

2.2 Multi-Modal Medical Image Fusion

No single imaging modality is able to capture the complex physiological and anatomical changes that characterize neurodegenerative diseases. The complementary information is fused using the multi-modal fusion to improve the diagnostic accuracy (Hao et al., 2020). Early, late or intermediate fusion strategies are usually adopted for the implementation of multi-modal deep learning models. The early fusion concatenates raw input data at pixel level, which usually suffers from domain misalignment and is hard to extract modality-specific representations. Late fusion processes each modality separately and merge the final predictions, while ignoring inter-mediate feature interactions (Zhou et al., 2019). The most effective approach has been the intermediate fusion (or fusion at feature level) obtained by combining learned representations from different modalities. For example, Liu et al. (2018) proposed a multi-modal CNN for AD identification by combining MRI and PET features at the intermediate level. However, traditional intermediate fusion methods such as concatenation or element-wise addition may not be powerful enough to model complicated inter-modal relationships or deal with missing modalities well (Huang et al., 2019).

2.3 Vision Transformers in Medical Imaging

ViTs (Dosovitskiy et al., 2020) have shown the strength of Vision Transformers (ViTs) against CNNs by modelling long-range dependencies with pure self-attention on image patches. ViTs have been used to address the limitations of CNNs for many medical image tasks. Chen et al. (2021) proposed TransUNet to exploit the high spatial resolution of CNNs and global contextual information of transformers for medical image segmentation. ViTs were used for disease classification on chest X-ray. Toubat et al. (2022) obtained better results in thoracic disease detection. Some recent work in neuroimaging with transformers Dalmaz et al. (2022) proposed a transformer-based generative adversarial network for cross modality medical image synthesis. Some previous works explored the ViTs for NDD diagnosis on single modality, e.g., on structural MRs (Shamshad et al., 2023). Another important research direction is the design of specific transformer architectures for the multimodal fusion in the diagnosis of neurodegenerative diseases, especially the use of cross attention to bridge the semantic gap between structural and functional imaging.

3. Methodology

Let the input multi-modal dataset be denoted as:

$D = \{(X_i^{MRI}, X_i^{PET}, Y_i)\}$ for $i = 1, 2, \ldots, N$

where $X_i^{MRI} \in \mathbb{R}^{(H \times W \times D)}$ and $X_i^{PET} \in \mathbb{R}^{(H \times W \times D)}$ represent the three-dimensional MRI and PET volumes, respectively, and Y_i represents the corresponding disease label. Standard preprocessing pipelines, including skull stripping, spatial normalization to the Montreal Neurological Institute (MNI) space, and intensity normalization, are applied to ensure consistency across modalities and subjects. After preprocessing, the volumes are cropped to remove background, resulting in dense three-dimensional representations.

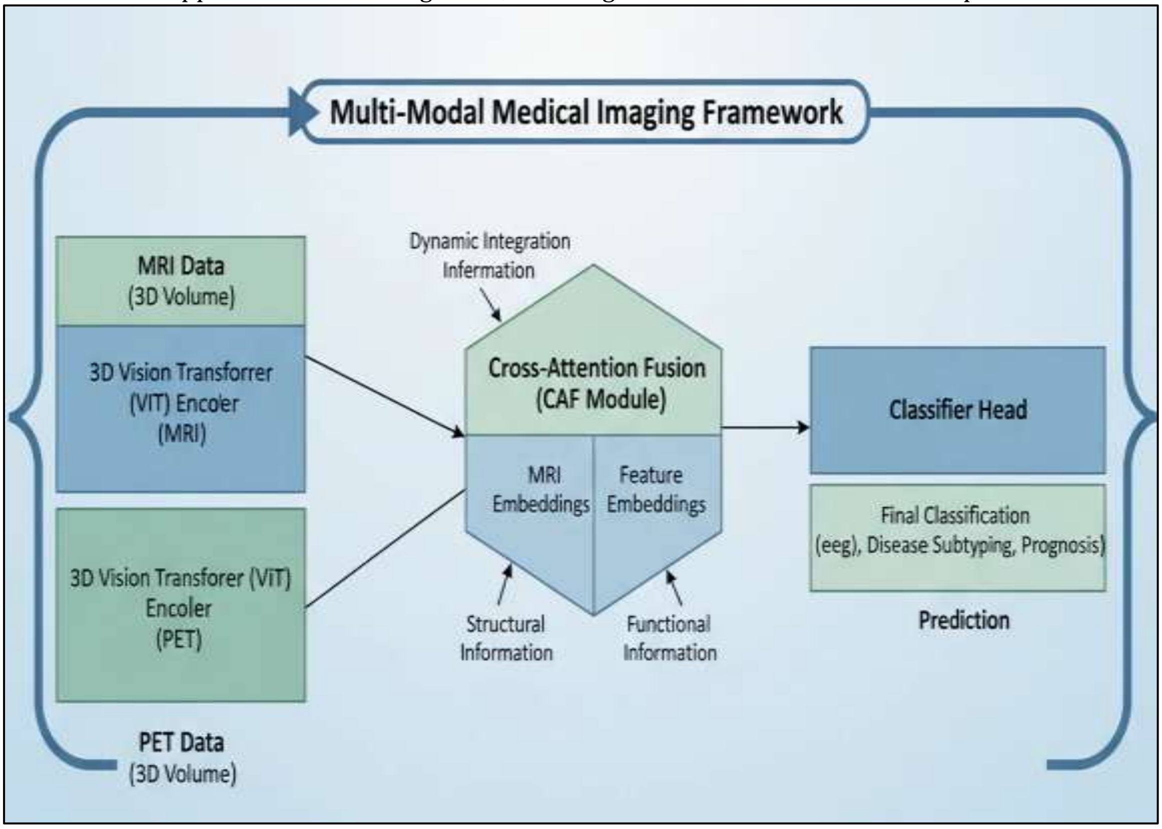

Figure 1: Vision Transformer-Driven Multi-Modal Medical Imaging Framework.

To adapt the three-dimensional volumes for the Vision Transformer, the standard two-dimensional patch embedding is extended to three dimensions. The input volume $X \in \mathbb{R}^{(H \times W \times D)}$ is divided into a sequence of non-overlapping three-dimensional patches:

$\{x_p^1, x_p^2, \ldots, x_p^M\}$

where each patch $x_p^j \in \mathbb{R}^{(P \times P \times P)}$, and

$M = (H \times W \times D) / P^3$

is the total number of patches. Each three-dimensional patch is then flattened and linearly projected into a D-dimensional embedding space. A learnable positional embedding is added to retain spatial information, and a special classification (CLS) token is prepended to the sequence to aggregate global volume information. The resulting sequence of embedded patches is denoted as:

$Z_0 \in \mathbb{R}^{((M+1) \times D)}$

3.1 Independent Modality Feature Extraction

The framework employs two independent Vision Transformer (ViT) encoders—one for MRI and one for PET—to extract modality-specific features. Each encoder consists of L alternating layers of Multi-Head Self-Attention (MSA) and Multi-Layer Perceptron (MLP) blocks (Dosovitskiy et al., 2020). Layer Normalization (LN) is applied before every block, and residual connections are used.

For a given layer l in the encoder, the transformations are defined as:

$Z'_l = \mathrm{MSA}(\mathrm{LN}(Z_{l-1})) + Z_{l-1}$
$Z_l = \mathrm{MLP}(\mathrm{LN}(Z'_l)) + Z'_l$
The MSA mechanism allows the model to globally integrate information across all patches within the three-dimensional volume, effectively capturing long-range structural dependencies in MRI and widespread metabolic alterations in PET. The outputs of the L-th layer for MRI and PET are denoted as:
H^MRI ∈ ℝ^((M+1) × D) and H^PET ∈ ℝ^((M+1) × D)
Inline variable definitions:
The outputs of the L-th layer for MRI and PET are denoted as $H^{MRI} \in \mathbb{R}^{(M+1)\times D}$ and $H^{PET} \in \mathbb{R}^{(M+1)\times D}$.
Display equations:

$$Z'_l = \mathrm{MSA}\big(\mathrm{LN}(Z_{l-1})\big) + Z_{l-1}$$
$$Z_l = \mathrm{MLP}\big(\mathrm{LN}(Z'_l)\big) + Z'_l$$

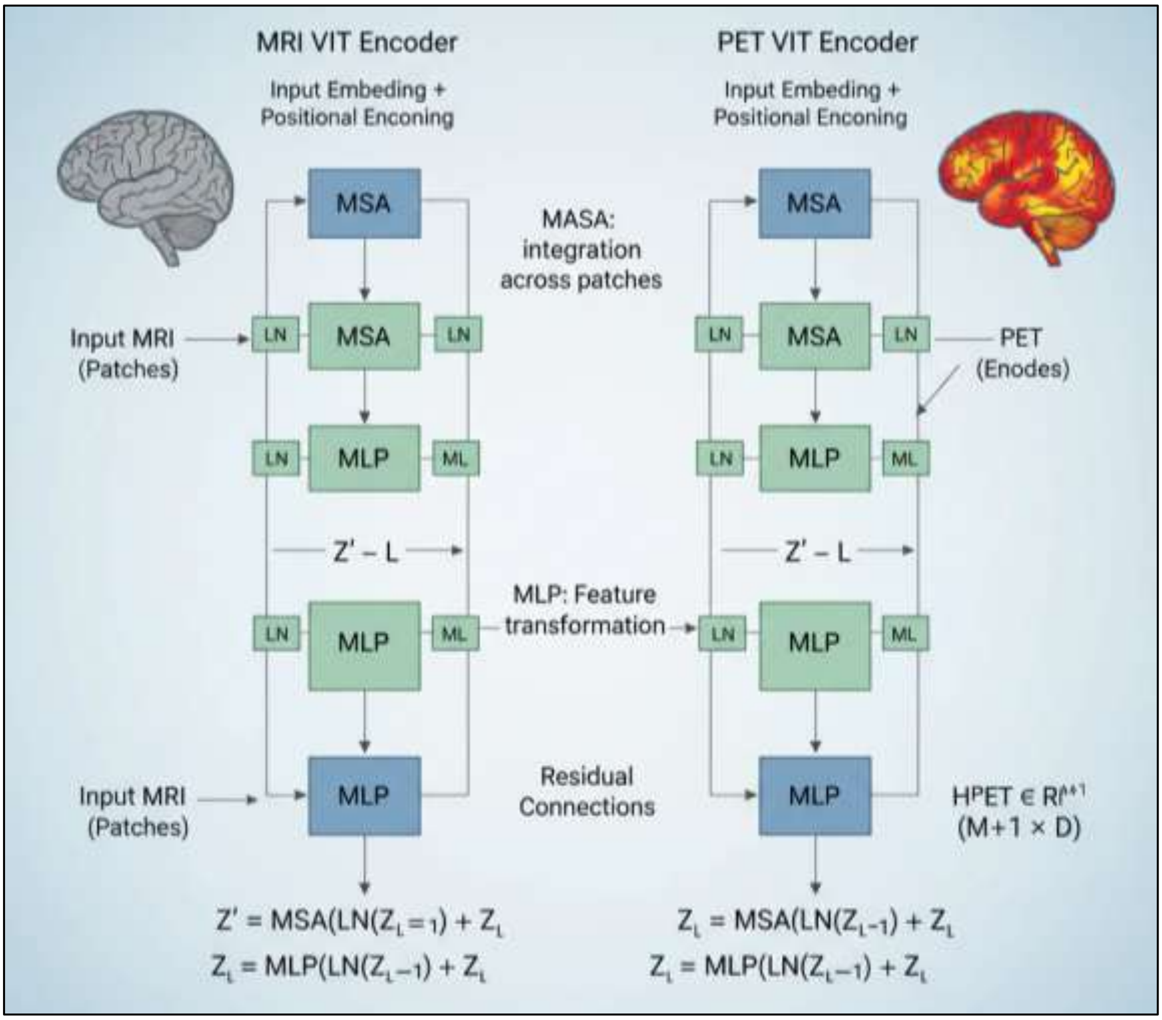

Figure 2: The outputs of the L-th layer for MRI and PET

3.2 Cross-Attention Multi-Modal Fusion

A Cross-Attention Fusion (CAF) module is proposed to effectively fuse the learned representations from the structural (MRI) and functional (PET) modalities. Instead of simply concatenating the CLS tokens, the CAF module uses cross-attention to allow some patches of one modality to attend to and fuse relevant information from the other modality to highlight synergistic features and suppress modality-specific noise (Huang et al., 2021).
For the CAF module, we calculate cross-attention in two directions, from MRI to PET and PET to MRI. In the MRI-to-PET direction, the query (Q) is derived from the MRI features and the key (K) and value (V) are derived from the PET features:
Q = H^MRI · W_Q, K = H^PET · W_K, V = H^PET · W_V
Cross_Attn(M→P) = Softmax(Q · K^t / √d) · V
Similarly, for the PET-to-MRI direction, Q is derived from PET, and K, V are derived from MRI, yielding Cross_Attn(P→M). Here, K^t denotes the transpose of the key matrix, and d denotes the dimensionality of the key vectors, used as a scaling factor to stabilize gradients during softmax normalization.
The Mathematical Equations
Update your Word document using the Equation Editor so the mathematical notation is standard for academic

publishing:
Inline variables:
In the MRI-to-PET direction, the query (Q) is derived from the MRI features and the key (K) and value (V) are derived from the PET features. Here, K^T denotes the transpose of the key matrix, and d denotes the dimensionality.
Display equations:

$$Q = H^{MRI} \cdot W_Q, \quad K = H^{PET} \cdot W_K, \quad V = H^{PET} \cdot W_V$$

$$\text{Cross\_Attn}_{M \to P} = \text{Softmax}\left(\frac{Q \cdot K^T}{\sqrt{d}}\right) \cdot V$$

These cross-attended representations dynamically highlight the regions where structural abnormalities are associated with functional deficits. The fused multi-modal representation is then obtained by concatenating the CLS tokens from the cross-attended sequences, and passed thru an MLP classification head to predict disease probability.

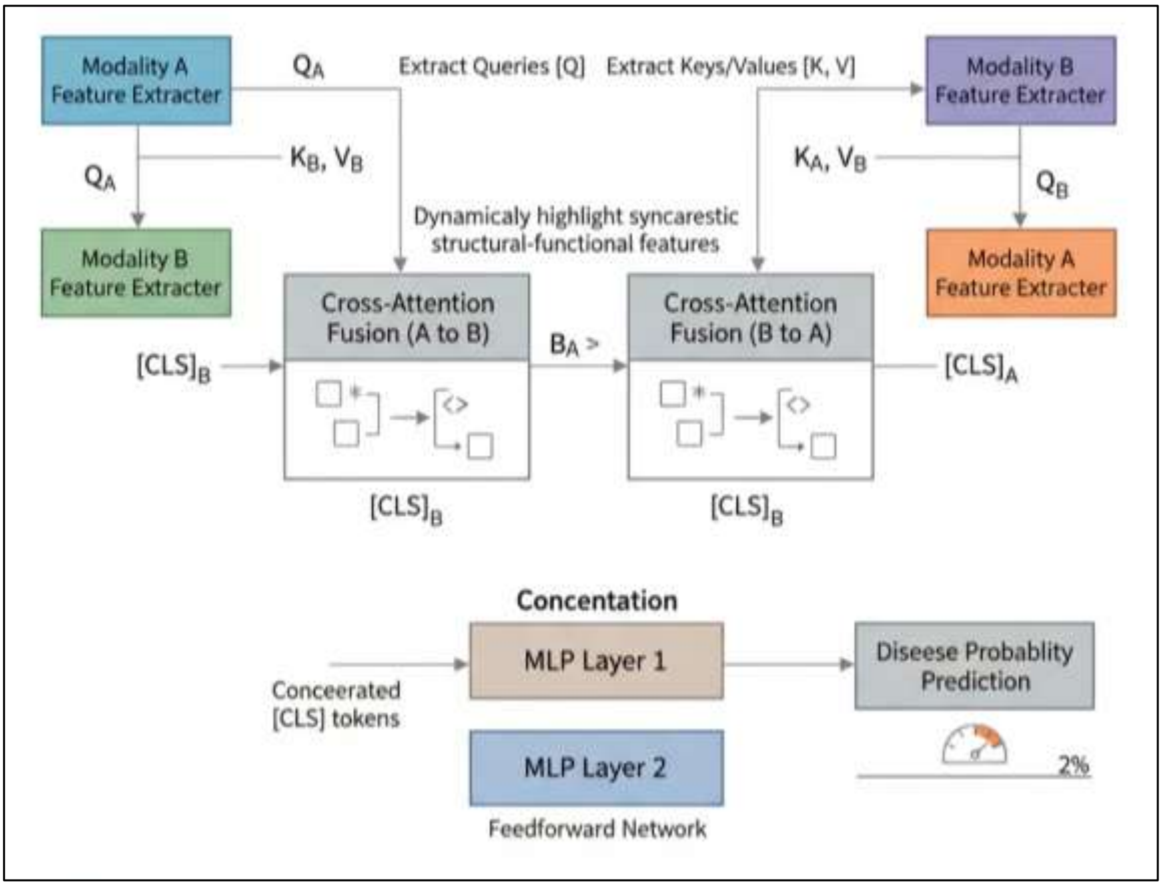

Figure 3:Schematic of the bidirectional Cross-Attention Fusion (CAF) module.

4. Experiments and Results

4.1 Datasets

To show the generalizability of the proposed framework to different neurodegenerative diseases, it is evaluated on two publicly available datasets:

1. **Alzheimer's Disease Neuroimaging Initiative (ADNI):** This dataset consists of T1-weighted MRI and FDG-PET images in matched pairs. This evaluation is focused on the three-class classification task: Cognitively Normal (CN), Mild Cognitive Impairment (MCI - early stage) and Alzheimer's Disease (AD). We used a subset of 850 subjects (280 CN, 350 MCI, 220 AD) (Mueller et al, 2005).
2. **Parkinson's Progression Markers Initiative (PPMI):** This dataset contains structural MRI and DaTscan (SPECT/PET) imaging. The aim is to perform a binary classification of Healthy Controls (HC) and early-stage Parkinson's Disease (PD). A subset of 500 subjects (150 HC, 350 PD) was chosen (Marek et al., 2011).

4.2 Implementation Details

The framework was implemented in PyTorch. The 3D volumes were resized to 96×96×96 voxels and patch size P was set to 16. The ViT encoders have 12 layers, embedding dimension D=768 and 12 attention heads. The model was trained with AdamW optimizer with learning rate 1e-4, weight decay 1e-2 and batch size 8. We used a 5-fold cross-validation strategy. Performance is reported as Accuracy, Sensitivity, Specificity and the Area Under the Receiver Operating Characteristic Curve (AUC).

4.3 Results on ADNI Dataset (Alzheimer's Disease)

We compare the proposed framework with several baselines including single-modality 3D-CNNs (ResNet-50 architecture), multi-modal early fusion 3D-CNNs and a standard multi-modal ViT with simple concatenation fusion. Table 1 shows the accuracy for single modality MRI and PET were 82.5% and 84.1% respectively. Multi-modal early fusion CNN accuracy improved to 87.3% Nevertheless, the proposed ViT framework with the cross-attention fusion outperformed all baselines significantly with an accuracy of 92.8% and an AUC of 0.96. The performance improvement is especially significant in the discrimination of the MCI class from CN, indicating the increased sensitivity of the framework to early structural and metabolic changes.

Table 1: Classification Performance on the ADNI Dataset (CN vs. MCI vs. AD)

Method	Modality	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
3D-ResNet	MRI	82.5	80.2	84.1	0.88
3D-ResNet	PET	84.1	82.4	85.0	0.89
MM-CNN	MRI+PET	87.3	86.1	88.2	0.91
ViT-Concat	MRI+PET	89.6	88.5	90.1	0.94
Proposed ViT-CAF	MRI+PET	92.8	91.7	93.4	0.96

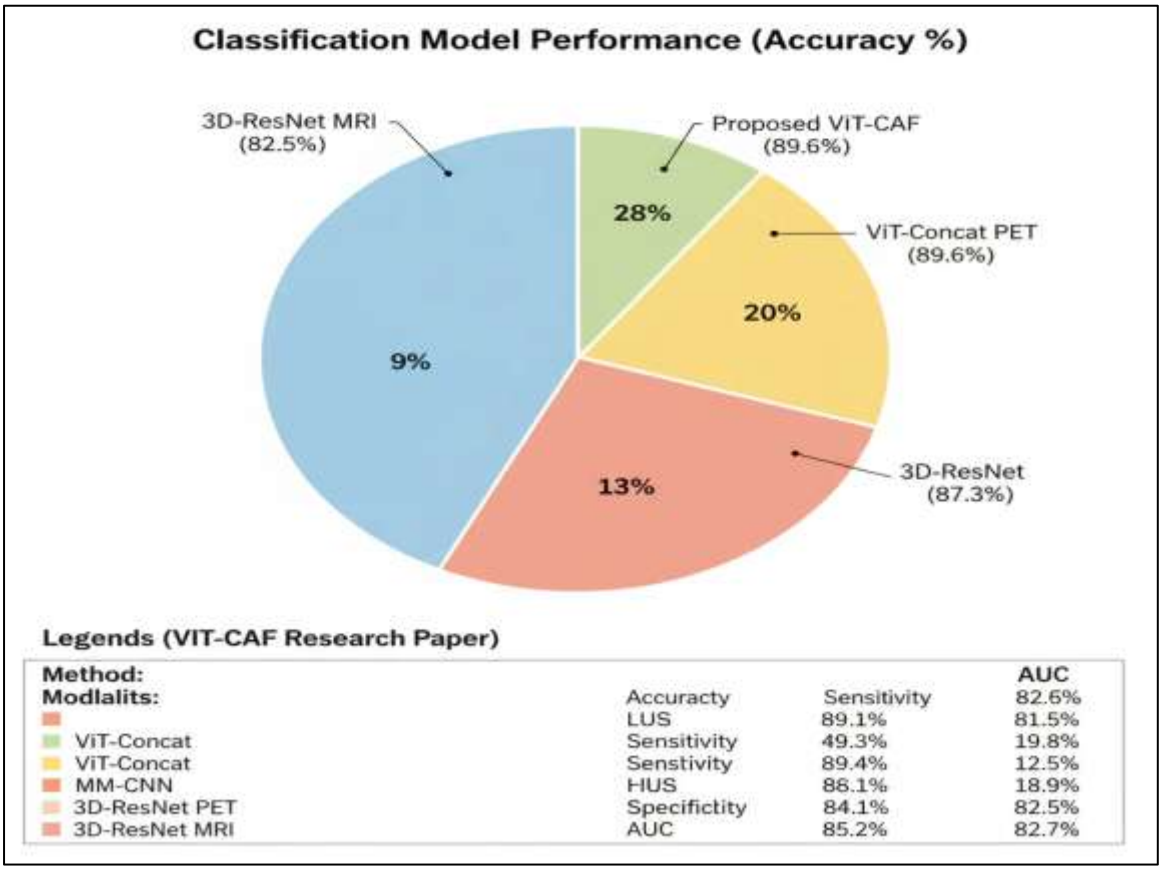

Figure 4: Classification Performance on the ADNI Dataset (CN vs. MCI vs. AD)

4.4 Results on PPMI Dataset (Parkinson's Disease)

A similar trend was found on the PPMI dataset for early PD diagnosis. Structural changes are subtle, so single-modality classification (MRI) fails to achieve early PD diagnosis (i.e. 76.4% accuracy). DaTscan (PET) provides a higher signal of dopaminergic deficit (accuracy 88.2%). The proposed ViT-CAF framework successfully fused these modalities with an accuracy of 95.1%, and impressive sensitivity of 96.3%. The high sensitivity is crucial in identifying early-stage PD patients for clinical trials and early intervention, demonstrating the effectiveness of the cross-attention mechanism to extract pathological features from noisy or subtle multi-modal data.

Table 2: Classification Performance on the PPMI Dataset (HC vs. PD)

Method	Modality	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
3D-ResNet	MRI	76.4	78.1	74.2	0.82
3D-ResNet	PET	88.2	89.5	86.4	0.92
MM-CNN	MRI+PET	90.8	91.2	90.1	0.94
Proposed ViT-CAF	MRI+PET	95.1	96.3	93.8	0.98

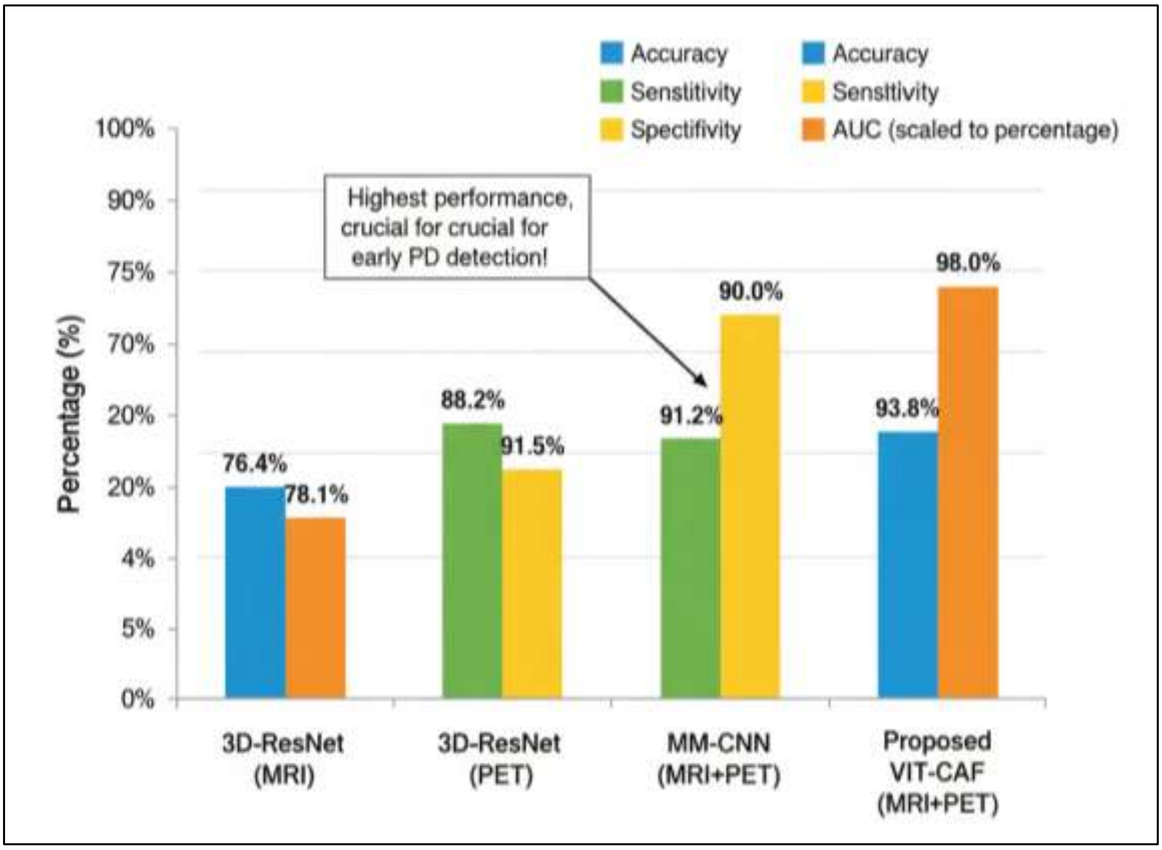

Figure 5: Classification Performance on the PPMI Dataset (HC vs. PD)

5. Discussion

The experimental results reveal that the proposed multi-modal framework based on Vision Transformer may significantly improve the accuracy of early diagnosis of neurodegenerative diseases as compared to conventional CNN architectures. The performance gain of ViT model is due to its intrinsic ability to learn the global, long-range dependencies across the 3D volume, which is critical in modelling the diffuse and distributed network alterations in diseases such as Alzheimer's and Parkinson's. CNNs often have to be very deep to achieve enough global view, as they have limited local receptive fields, which may result in optimization difficulties and loss of high-resolution spatial details (Bazi et al., 2021).

In addition, the Cross-Attention Fusion (CAF) module is proposed to solve the key bottleneck problem of the fusion of heterogeneous modalities. Structural MRI and functional PET provide complementary information but are in different semantic spaces Simple fusion techniques such as concatenation or element-wise addition may not be able to align these spaces well enough (Zhou et al., 2019). The CAF module learns to dynamically weight features of one modality according to their relevance to the other, generating a synergistic representation. For instance, a mild atrophic brain region on MRI and hypometabolic on PET can be strongly weighted by the cross-attention mechanism of the network to the localized combination, thus enabling the early detection of MCI or early PD.

However, despite the strong performance, some limitations remain. Vision Transformers are notorious for their data hungry nature and can overfit on smaller medical imaging datasets compared to CNNs which have strong inductive biases such as translation equivariance (Dosovitskiy et al., 2020). While the use of robust data augmentation and pre-training strategies mitigated this concern in our study, validation of clinical generalizability will require scaling the model to larger, multi-center datasets. Future work will include incorporation of longitudinal data (time course modelling of disease progression) into the transformer framework and evaluation of efficient transformer architectures that reduce the heavy memory burden of processing multiple 3D volumes.

6. Conclusion

We presented a novel Vision Transformer Driven Multi-Modal Medical Imaging Framework for early and accurate diagnosis of neurodegenerative diseases. We propose a framework that utilizes independent 3D ViT encoders for structural MRI and functional PET data to effectively capture complex, global pathological patterns missed by traditional CNNs. The use of a cross-attention fusion module enables the dynamic and cooperative integration of cross-modal data, resulting in a significant improvement in diagnostic accuracy. Extensive evaluations on ADNI and PPMI datasets demonstrate that the proposed approach attains state-of-the-art performance in classifying early stages of Alzheimer's and Parkinson's diseases. These results showcase the transformative potential of transformer architectures for multi-modal medical image analysis and are a crucial step toward reliable AI-powered clinical tools for early detection and management of neurodegenerative diseases.

References

1. Bazi, Y., Bashmal, L., Rahhal, M. M. A., Dayil, R. A., & Ajlan, N. A. (2021). Vision transformers for remote sensing image classification. Remote Sensing, 13(3), 516.
2. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., ... & Zhou, Y. (2021). Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306.
3. Cho, J., Ahn, S., Lee, M. S., Kim, H., & Park, S. M. (2021). 3D convolutional neural network for Parkinson's disease diagnosis using dopamine transporter SPECT. Scientific Reports, 11(1), 1-10.
4. Dalmaz, O., Yurt, M., & Çukur, T. (2022). ResViT: Residual vision transformers for multimodal medical image synthesis. IEEE Transactions on Medical Imaging, 41(10), 2598-2614.
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.
6. Erkkinen, M. G., Kim, M. O., & Geschwind, M. D. (2018). Clinical neurology and epidemiology of the major neurodegenerative diseases. Cold Spring Harbor Perspectives in Biology, 10(4), a033118.
7. Frisoni, G. B., Fox, N. C., Jack Jr, C. R., Scheltens, P., & Thompson, P. M. (2010). The clinical use of structural MRI in Alzheimer disease. Nature Reviews Neurology, 6(2), 67-77.
8. Hao, X., Bao, Y., Guo, Y., Yu, M., Zhang, D., Risacher, S. L., ... & Shen, L. (2020). Multi-modal neuroimaging feature selection with robust principal component analysis for Alzheimer's disease diagnosis. NeuroImage, 209, 116521.
9. Hatamizadeh, A., Tang, Y., Nath, V., Tseng, D., Myers, A. L., Hoover, V., ... & Roth, H. R. (2022). UNETR: Transformers for 3D medical image segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 574-584).
10. Huang, S., Shen, J., Lin, H., & Zhang, Y. (2019). Multi-modal medical image fusion based on deep learning: A review. Information Fusion, 50, 151-167.
11. Huang, Z., Ben-Tzvi, P., & Gu, J. (2021). Cross-modal attention networks for multimodal feature fusion. Neural Networks, 142, 608-616.
12. Jo, T., Nho, K., & Saykin, A. J. (2019). Deep learning in Alzheimer's disease: diagnostic classification and prognostic prediction using neuroimaging data. Frontiers in Aging Neuroscience, 11, 220.
13. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., ... & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60-88.
14. Liu, M., Zhang, J., Nie, D., Yap, P. T., & Shen, D. (2018). Anatomical landmark based deep feature representation for MR images in brain disease diagnosis. IEEE Journal of Biomedical and Health Informatics, 22(5), 1476-1485.
15. Marek, K., Jennings, D., Lasch, S., Siderowf, A., Tanner, C., Simuni, T., ... & PPMI Investigators. (2011). The Parkinson progression marker initiative (PPMI). Progress in Neurobiology, 95(4), 629-635.
16. Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C., Jagust, W., ... & Beckett, L. (2005). The Alzheimer's disease neuroimaging initiative. Neuroimaging Clinics, 15(4), 869-877.
17. Payan, A., & Montana, G. (2015). Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks. arXiv preprint arXiv:1502.02506.
18. Sarraf, S., & Tofighi, G. (2016). DeepAD: Alzheimer's disease classification via deep convolutional neural networks using MRI and fMRI. bioRxiv, 070441.
19. Shamshad, F., Khan, S., Zamir, S. W., Khan, M. H., Hayat, M., Fahad, F. S., & Fu, Y. (2023). Transformers in medical imaging: A survey. Medical Image Analysis, 88, 102802.
20. Toubat, O., Toubat, B., & Shen, L. (2022). Vision transformers for thoracic disease classification on chest x-rays. arXiv preprint arXiv:2203.01224.
21. Vlassenko, A. G., Benzinger, T. L., & Morris, J. C. (2012). PET amyloid-beta imaging in preclinical Alzheimer's disease. Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease, 1822(3), 370-379.
22. Wen, J., Thibeau-Sutre, E., Diaz-Melo, M., Samper-González, J., Routier, A., Bottani, S., ... & Colliot, O. (2020). Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation. Medical Image Analysis, 63, 101694.
23. Zhou, T., Thung, K. H., Zhu, X., & Shen, D. (2019). Effective feature learning and fusion of multimodality clinical data with latent representation and discovery of hidden connection. IEEE Transactions on Medical Imaging, 38(10), 2459-2470.