National College of Ireland

From Binary to Multiclass: Adapting the IT Transformer for Fine-Grained Alzheimer's Disease Stage Classification

MSc Research Practicum Part 2
MSc in Data Analytics

Joseph Jacob Anjilimoottil
Student ID: 24103390

School of Computing
National College of Ireland

Supervisor: Vikas Sahni

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Joseph Jacob Anjilimoottil
Student ID:	24103390
Programme:	MSc in Data Analytics (MSCDAD_A_JAN25I) **Year:** 2025
Module:	MSc Research Practicum Part 2
Supervisor:	Vikas Sahni
Submission Due Date:	11th December 2025
Project Title:	From Binary to Multiclass: Adapting the IT Transformer for Fine-Grained Alzheimer's Disease Stage Classification
Word Count:	3904 **Page Count**: 22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Joseph Jacob Anjilimoottil

Date: 10 December 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

From Binary to Multiclass: Adapting the IT Transformer for Fine-Grained Alzheimer's Disease Stage Classification

Joseph Jacob Anjilimoottil
24103390
MSc in Data Analytics

Abstract

Classification of Alzheimer disease (AD) based on neuroimaging is a difficult task with limited data, handcrafted features, and a low interpretation level in deep learning models. This study explores the use of Interpretable Transformer (IT) architecture in four-class FMRI-based AD staging. The primitive binary model of PET-MRI fusion is adjusted to a single-modality MRI platform to enhance accessibility and decrease the reliance on the PET information. A set of image normalization and stratified sampling are used to construct a preprocessing pipeline that makes it possible to build strong model development. This IT model is benchmarked with four fixed CNNs, namely ResNet-18, VGG-16, EfficientNet-B0 and MobileNet-V2, to compare the differences in performance and computational needs. The IT model achieves an intermediate level of accuracy of 66 percent has compared to the base model, but more detailed global features are represented, which shows that this model requires additional tuning and bigger datasets. This paper illustrates that transformer-based architectures are viable to interpretable brain MRI-scan AD classification and offers the possibilities of further development.

1 Introduction

1.1 Background

Alzheimer Disease (AD) is a neurodegenerative disorder that accounts for approximately 60-70% of dementia cases worldwide and in 2021 an estimated 57 million people were living with dementia globally (GBD 2019 Dementia Forecasting Collaborators, 2022). It is a significant medical, social and economic issue especially in the ageing populations. Early AD diagnosis is an important factor since it offers a chance to intervene in time, thus, potentially preventing the further development of the disease, improving the quality of life of the patients and assisting them in planning their long-term healthcare. Nevertheless, objective and objective classification between normal ageing of the human brain, Mild Cognitive Impairment (MCI), and the onset of AD have become a thorny clinical challenge to date. The conventional diagnostic methods tend to be based on subjective cognitive tests, invasive testing of biomarkers or complicated imaging regimes that can be quite costly, time-intensive, and unable to detect AD in its early periods (Rapaka et al., 2025).

1.2 Emergence of Artificial Intelligence in AD Diagnosis

Artificial intelligence (AI) and deep learning, in particular, has become one of the newly revolutionary medical diagnosis tools over the last few years. In medical imaging, deep learning models show extraordinary abilities to recognize difficult data patterns, and as such, they can be used to detect subtle neurodegenerative changes that humans cannot perceive. Frameworks derived in transformers, initially used in natural language processing, have been applied to neuroimaging tasks in more and more cases, as they can be used to capture local and global data dependencies. IT architectures especially have demonstrated good results in the prediction of Alzheimer disease with multimodal imaging modalities like MRI and PET (Yao et al., 2025).

1.3 Current Challenges of Existing Diagnostic Models

Although these developments have occurred, most AI-based studies of Alzheimer disease, which are estimated to be more than 80% of the publications in the field since 2020, are limited to binary classification, usually between AD and healthy controls. Though this makes their analysis easier, they do not reflect the entire spectrum of cognitive impairment and do not appreciate the clinical need to distinguish between healthy ageing, early MCI (eMCI), late MCI (lMCI) and early-stage AD. Multi-class classification is able to detect intermediate pathologies of the disease and was found to enhance diagnostic accuracy by 10-15% (Juganavar, Joshi, and Shegekar, 2023). Nevertheless, the literature emphasises that there is a continuous gap in the given field, especially with respect to generalisability, interpretability, and applicability to clinical practice.

1.4 Research Question

To overcome this weakness, the study proposes a state of the art multi-class classification model that is founded on the Interpretable Transformer (IT) framework. The purpose of this is to offer a solid, non-invasive diagnostic method that has the ability to discriminate between Healthy Controls (HC), eMCI, lMCI, and early-stage AD through a high-quality neuroimaging information. The model is tested in terms of sensitivity and specificity taking into consideration the clinical reliability of the model with a minimum of false positives and false negatives, two major considerations in medical diagnostics. This strategy aims at narrowing this gap between the performance of algorithms and their use in clinical practice.

The study addresses the following research question:
Can a multi-class classification interpretable transformer model effectively distinguish between healthy controls, early MCI, late MCI, and early-stage Alzheimer's Disease, and how does its performance compare to conventional CNN baselines?

1.5 Contributions of the Study

This research has two major contributions. In academia, it assists with the use of transformer-based architectures in the fields of medical imaging and multi-class classification and fills an

unattended void in the literature. Clinically, it helps frontline healthcare workers to identify the early signs of the Alzheimer disease through the use of a non-invasive, data-driven application that could help to diagnose the disease earlier, which can enhance patient outcomes and healthcare resource utilization. Moreover, the research can enhance the creation of more viable practice methods in diagnosis because it will decrease the use of subjective tests and invasive processes.

1.6 Structure of the Report

The rest of this report is then organized in the following manner. Section 2: Related Work presents the critical review of the literature available on the topic of deep learning and transformer-based models of detecting the changes in Alzheimer disease. Section 3: Research Methodology explains the dataset, methods of preprocessing and structure of the experiment. Section 4: Design Specification outlines the architectural specification of the proposed interpretable transformer model and their parts. Section 5: Implementation will describe the process of development, which would involve the model training, validation, and optimization. Section 6: Evaluation: examines the sensitivity, specificity and the accuracy of the model at classification. Lastly, Section 7: Conclusion and Future Work brings the main findings to the conclusion and the future research directions.

2 Related Work

2.1 Deep Convolutional and Transfer learning Approaches.

In the emerging field of AD classification, deep convolutional and transfer learning models have been used as the basis of the use of neuroimaging data. Jain et al. (2021) came up with an AI-based application of identifying and estimating dementia severity based on MRI data that achieved an accuracy of approximately 92% and shows the promise of deep CNNs in the analysis of neurodegenerative diseases. On the same topic, Kabir et al. (2021) have made a comparative study of deep-learning systems to detect multi-classes of Alzheimer disease on the basis of a brain MRI image, which focuses on the possibility of the CNN architecture to distinguish the stages of a disease. Rao et al. (2024) further developed this research by using deep and transfer learning models on three-dimensional MRI data, thus maintaining the volumetric context that is commonly ignored by two-dimensional slice-based systems. Subramanyam Rallabandi and Seetharaman (2023) investigated the topic of multimodal fusion of MRI and PET data in classifying healthy controls, mild cognitive impairment (MCI), and AD, showing that the combination of structural and metabolic imaging can be used to achieve better diagnostic accuracy rate exceeding 90%.

Despite these successes, several limitations are also observed. Most of the studies utilized datasets of less than 500 subjects, which are not cross-institutionally validated, which diminish the extrapolation of the models to scanners and demographics (Kabir et al., 2021; Jain et al., 2021). Moreover, slice-based approaches are also prone to loss of important spatial

information whereas interpretability is low, with fewer than 10% of studies giving clinically significant explanations to their choices.

2.2 Deep-Learning Architectures Based on Transformers and Hybrids.

Recent studies have given particular attention to transformer-based and hybrid deep-learning architectures that have a greater ability to understand the neuroimaging data on a global scale. The swin transformer proposed by Liu et al. (2021) motivated the development of many applications in medical imaging, among which the Vision Transformer (ViT) method of AD detection by Chandan, Rashmi and Pathan (2025) can be mentioned. Velu and Jaisankar (2025) also combined CNN and Swin Transformer modules to come up with a hybrid model that both detects local and global features of the images. The more complex designs that display multimodal imaging and explainable AI include dual-transformer architecture with cross-attention and the graph pooling mobilisation by Tang et al. (2023) and interpretable transformer by Yao et al. (2025) who demonstrate how transformer structures may be modified. Although these approaches demonstrate obvious performance benefits and stronger feature extraction in contrast to CNN-only ones, they also make life challenging. Transformers are aggressive in terms of datasets and computational resources, which are usually lacking in medical imaging (Chandan et al., 2025). Hybrid models may get too complicated, and most studies do not provide detailed ablation studies explaining which architectural elements are contributing to the improvement in performance (Velu and Jaisankar, 2025). Moreover, the interpretability of attention mechanisms is also promising, but has not been fully tested with clinicians to guarantee that the explanations have a sense in the context of diagnosis (Yao et al., 2025).

2.3 Multimodal Fusion, Graph-Based and Feature Selection.

Another parallel line of research has focused on multimodal fusion, feature selection and graphs. Jiao et al. (2022) came up with a multi-modal feature-selection model that integrates both feature correlation and structure fusion to differentiate MCI and AD, which enhanced the accuracy of its classification to 91.2 per cent compared to Zhang et al. (2023) who developed a multi-modal graph neural network that learnt the correlations among brain regions using structural MRI and PET information.Such works highlight the importance of integrating both anatomical and metabolic data in order to have the complete picture of the Alzheimer pathology. Equally, dual transformers with cross-attention were used by Tang et al. (2023) to combine multimodal data. Nevertheless multimodal approaches have serious limitations, In contrast to other methods, PET are 40 to 60 % more expensive than MRI and less accessible, which leads to the small size of the dataset, hindering clinical scalability (Zhang et al., 2023). In addition, most fusion methods use naive concatenation methods or late fusion approaches without paying much attention to noise and heterogeneity of various imaging modalities (Subramanyam Rallabandi and Seetharaman, 2023). Graph-based approaches are nevertheless conceptually straightforward, and frequently based on a priori parcellations of the brain, or connectivity measures, the effect of which on model performance is not always carefully investigated.

2.4 Image Synthesis and Generative Models.

One of the most significant developments has been the adoption of generative models and synthesis of images with a view to producing missing imaging modalities and, as a result, enhancing multimodal diagnosis.Zhang et al. (2022) suggested BPGAN, a generative adversarial network (GAN) aimed at synthesising the PET image based on the MRI data, and the ability to perform multimodal diagnosis of Alzheimer without scanning the patient with PET achieving an average peak signal-to-noise ratio of 29.4 dB . Dayarathna et al. (2024) provided a detailed overview of the synthesis of deep-learning-based MRI, CT, and PET synthesis, identifying the opportunities and the traps of these methods. These generative models have the potential to raise the availability of data and minimize the invasiveness of PET imaging, which meets clinical objectives of reducing the workload of the diagnostic process. Synthesized modalities, however, are subject to over 15% artefacts and unrealistic properties, which can negatively affect the clinical reliability (Zhang et al., 2022). The evaluation measures applied in these studies also have a bias towards image similarity measures e.g. PSNR, SSIM rather than clinically meaningful measures such as diagnostic accuracy or prognostic value limiting their translational validity.

2.5 Interpretability and Explainable Artificial intelligence

The theme of interpretability and explainable AI has been developed in reaction to the criticisms of the lack of transparency in deep-learning systems. Xu and Yan (2022) proposed a multi-class classification model with explainable AI techniques, and Yao et al. (2025) proposed an interpretable transformer and visualises the decision justification on the basis of the PET/MR. Onciul et al. (2025) talked in a wider range of synergies between AI and neuroscience and argued that algorithmic output should be more compatible with neurobiological knowledge. Rapaka et al. (2025) also studied diagnostic error in the case of Alzheimer disease and the significance of non-invasive and reliable biomarkers. Even though these contributions are very important to increase transparency and trust between clinicians, the quantifiable measures of explain fidelity were included in only about 12% of the studies. The assessment of the faithfulness of model explanations is not highly standardised, and fewer than 5% studies include clinicians in user studies to identify whether interpretability tools are actually helpful in decision-making or diagnosis.

2.6 Clinical Integration/Conceptual Frameworks.

Multiple reviews and conceptual frameworks also put AI innovations into the broader clinical and diagnostic context. Juganavar, Joshi and Shegekar (2023) offer a detailed overview of the recent diagnostic technologies, identifying the key barriers to the early diagnosis of Alzheimer. According to Rapaka et al. (2025), an instance of diagnostic confusion and the usefulness of non-invasive biomarkers are repeated, but Chamundeswari et al. (2025) present a pattern recognition-neural interpretability model, CognitivePathway as a method of detecting early AD. These papers put the technological advances into perspective of the clinical practice environment, and indicate the continuing necessity of viable, generalisable

diagnostic instruments. However, most of these reviews do not go further to recommend standardised pipelines of evaluation or powerful validation schemes needed to allow clinical acceptance.

2.7 Hybrid and Adaptive Fusion Architectures.

The hybrid and adaptive fusion architecture has received some interest due to its ability to capture the disease progression spectrum. Muhammad et al. (2025) implemented an adaptive feature fusion hybrid deep-learning architecture, which uses adaptive changes in the contribution of imaging modalities in various disease stages. Transformer and graph-based frameworks have been adopted using similar strategies (Tang et al., 2023; Jiao et al., 2022), implying that adaptivity can be useful to increase robustness and interpretability. Nevertheless, such architectures do not always report the determination of fusion weights explicitly and the correspondence of adaptive mechanisms to clinically interpretable aspects.

2.8 Summary of the Literature Findings

The analyzed literature shows that the use of artificial intelligence (AI) and especially deep learning is still revolutionizing the detection of Alzheimer disease (AD) using modern models like convolutional neural networks (CNNs), transformers, and hybrid models. Nevertheless, a review of literature emerging in research findings indicates uniform technical and clinical shortcomings. CNN-based models have a high diagnostic accuracy, frequently over 90 percent, but do not require large and heterogeneous datasets, often less than 500 subjects, and therefore are not generalisable or interpretable. Transformer-based and hybrid designs provide enhanced features representation at the global scale but need a significantly more significant size of data and more computation power, and the readability of their internal structures has not been sufficiently investigated. Multimodal fusion techniques involving both MRI data and PET data have demonstrated superior performance, although their practical use is limited by the expensive nature of PET imaging, the lack of datasets to use, and the rudimentary nature of fusion schemes. Generative models like GANs are used to address data scarcity, in which they are used to restore missing modalities but can introduce artefacts in structure as well as focus more on visual similarity metrics over improvements in clinical diagnosis. Explainable AI (XAI) frameworks have been suggested to improve the model transparency, but less than 15 percent of studies measure the explanation fidelity and engage clinicians in the validation process. In the field of clinical research, there are still unresolved issues of standardised assessment, multi-site verification and reproducibility, which limit the application of AI-based dementia classifiers to the practical healthcare setting.

Table 1: Summary of Key Papers and Identified limitations

Authors	Technique	Identified limitations
Jain et al.(2021)	Deep CNN on MRI data	Small dataset (<500 subjects); limited interpretability and generalisability

Kabir et al. (2021)	CNN-based multi-class MRI model	Dataset imbalance; absence of cross-site validation
Chandan et al. (2025)	Vision transformer	Requires large datasets; lacks clinical testing
Velu & jaisankar (2025)	CNN-Swin transformer	Overly complex; unclear feature contribution; no ablation analysis
Yao et al. (2025)	Interpretable transformer	Lacks quantitative interpretability validation; not clinically tested
Zhang et al. (2022)	GAN-based PET synthesis	Introduces ~15% artefacts; limited diagnostic reliability

Hence, the state of the art is still not sufficient to enable reliable, interpretable and scalable early diagnosis of AD at its progressive phases. This paper fills these gaps by creating a multi-class interpretable transformer (IT) model with the capacity to differentiate between Healthy Controls (HC), Early MCI (eMCI), Late MCI (lMCI), and early-stage AD with improved interpretability, sensitivity, and specificity.

3 Research Methodology

The study follows the experimental research design that is organized based on the creation, training, and testing of a modified Interpretable Transformer (IT) model to classify multiclass Alzheimer disease using MRI images. The design of the methodology is intended to evaluate in a systematic manner the hypothesis on whether transformer-based architectures can be used to stage fine-grained cognitive decline using conventional convolutional neural networks (CNNs). All their experiments were done in Google Colab Pro, which uses an NVIDIA A100, which is implemented in Python 3.12, PyTorch 2.9, and Torchvision 0.24.0 with supporting libraries, including Scikit-learn, Matplotlib and Seaborn.

3.1 Dataset Description

The data to be utilized in this study is Alzheimer Multiclass Dataset (Equal and Augmented) available on Kaggle[1]. It contains MRI scans labelled into four balanced categories:

- Non-Demented
- Very Mild Demented
- Mild Demented
- Moderate Demented

Pre-augmentation is made and images are evenly spread eliminating the issue of class imbalances. All scans are 2D axial MRI slices, in JPEG format. A summary of the dataset, including the number of images and the distribution of the number of images per class, is

[1] https://www.kaggle.com/datasets/aryansinghal10/alzheimers-multiclass-dataset-equal-and-augmented

given in Table 2, which confirms the balance of the dataset and its applicability to supervised learning.

Table 2: Dataset class distribution and sample characteristics

Class	Count
MildDemented	10,000
ModerateDemented	10,000
NonDemented	12,800
VeryMildDemented	11,200

3.2 Exploratory Data Analysis (EDA)

To gain a better idea of the nature of the MRI dataset and to make decisions on preprocessing, a structured EDA was conducted. Representative MRI samples of each diagnostic group were reviewed to check the visible structural changes between stages (e.g. ventricular enlargement or cortical thinning). Pixel-intensity histograms were used to investigate the human eye observation of brightness and contrast variations across the dataset, encouraging normalization and controlled data augmentation. Some sample MRI slices are provided in **Figure 1** and distributions of intensities are plotted in **Figure 2**.

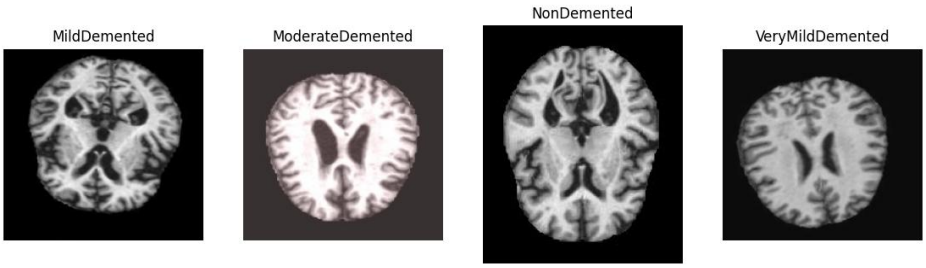

Figure 1: Sample MRI images across the four dementia stages

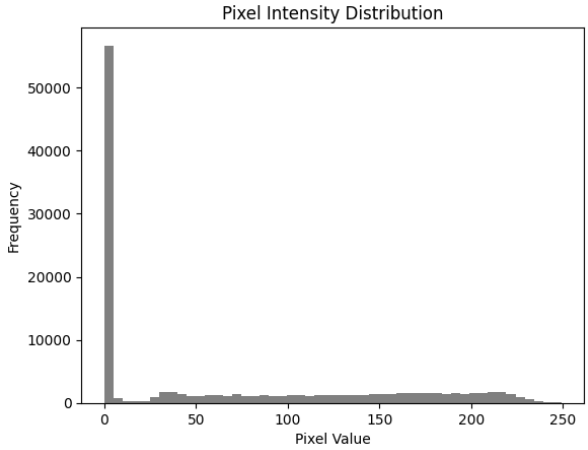

Figure 2: Distribution of pixel intensities over dataset.

These analyses made sure that the preprocessing pipeline was in line with the imaging properties and the classification objectives.

3.3 Data Preparation and Transformation

The implementation began by loading the Alzheimer MRI multiclass dataset from the Kaggle environment. The transformation process of each MRI image consisted of resizing all of them to 224x224, equalizing the average and standard deviation of intensities to a predetermined value, and a data augmentation method that enhanced robustness. Random rotation, horizontal flips, affine transformations, color jitter, Gaussian blur and random cropping were added to the last augmentation pipeline along with other photometric manipulations. Every transformation was chosen to represent natural clinical variation - e.g. changing scan orientation or changing scanner intensity but still maintaining anatomical integrity.

To make sure that all the four classes have the same representation in the two datasets, a stratified sampling approach was used. The data set was divided into a 70 percent training data, 10 percent validation data and 20 percent testing data. This strategy reduced sampling bias and enabled the validation set to be a useful source of hyperparameter tuning and early stopping. All the models that were developed in the later stages had these transformed datasets as the input.

4 Design Specification

4.1 System Architecture Overview

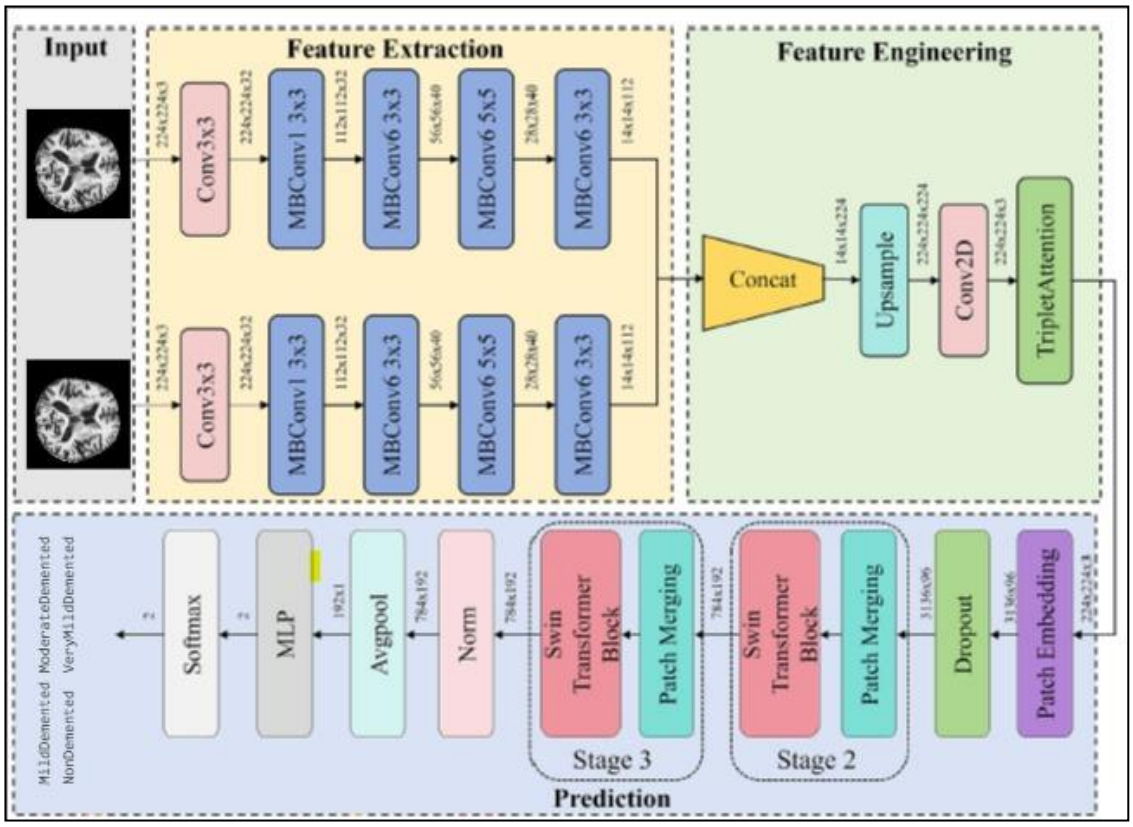

Figure 3: System architecture of the Multi-class Interpretable Transformer

The proposed multiclass Alzheimer classification system is built based on the Interpretable Transformer (IT) architecture, which was initially designed in PET fusion with MRI. As this study is aimed at MRI only classification, the model is modified to receive one modality and retain the dual-branch design of the IT transformer.

The entire system architecture entails:

1. Input pipeline and data loader
2. Augmentation and preprocessing of data
3. Modified IT Transformer model
4. Training and validation engine (AMP-enabled)
5. Evaluation and visualization aspects
6. Baseline CNN benchmarking subsystem

All components are modular and written in PyTorch to guarantee reproducibility and extensiveness. The overall architecture is illustrated in Figure 3.

4.2 Input Interface

Each prediction takes two MRI brain scans as input to the system. They may be different slices or views of the same object (two axial slices or various planes), or of two processed variations of the underlying MRI. All input images are first rescaled to the network working resolution and turned into a tensor and normalized to a constant mean and standard deviation. The model is able to learn complementary structural information and retain the original dual-branch design of the IT architecture, by taking two MRIs rather than one.

The network is fed with both images on two equal branches of input. The inputs are preprocessed identically and parameterized identically and therefore, any variation in the features learned is due not to inconsistent processing but to the content of the data itself.

4.3 Feature Extraction Branches

Every input MRI is processed by one of the branches, which consists of a primary convolutional layer, and a chain of lightweight convolutional blocks (MBConv-style). These blocks reduce and enlarge channel dimension by depthwise separable convolution and non-linear activates enabling the network to express more finer grained local structure, cortical thickness, ventricle boundaries and regional intensity patterns at a massive computational cost.

The two branches work parallel and do not share weights. This enables the network to specialise the representation of every MRI input, even when the scans are not made at the same image plane or slice. By the end of this step, a dense feature map of the spatial pattern of the respective MRI image is generated at each of the branches.

4.4 Triplet Attention and Feature Fusion

The two branches of the feature extractions are then fused together in the channel dimension to produce one fused feature map. This fused representation is processed by a series of convolutional layers in order to bring the channel dimensions into phase and further refine the joint encoding. The resultant fused tensor is then sent through Triplet Attention module.

Triplet Attention works through successive attention in different combinations of channel and spatial dimensions. It is also explicit in the sense that instead of just height and width, it models interactions on a height channel, width channel, height-width planes. Practically this enables the model to point on the regions of the brain, and channel of features that are collectively informative in discerning between the adjacent stages of dementia. The output is a small information-dense feature map that is highly compatible with tokenisation and processing by transformers.

4.5 Transformer Encoder Stages

Upon fusion and attention, the feature map is rearranged to a series of tokens by patch embedding. The tokens embedded therein are then fed into a hierarchical transformer encoder which is a set of stages. Each stage contains:

- a multi-head self-attention mechanism that is based on windows,
- a non-linear activation feed-forward network,
- remaining residual connections with normalization of layer.

The first stages are concerned with local neighbourhoods in the picture and fine structural patterns within small areas of the brain are captured. The further levels combine patches in increasing levels and extend the receptive field, allowing the model to make arguments over long-range dependencies, like what should change in one lobe of the brain would increase the global atrophy. The hierarchical design saves on the computational cost when compared with full global attention but still allows one to capture the global context.

In order to stabilize the training and avoid overfitting the model, dropout and normalization layers are inserted into the stack of transformer layers, a fact that is especially essential as the model is deep and large.

4.6 Classification Head

A series of high-level tokens is generated within the last transformer stage. Based on the configuration they are pooled together, by global average pooling, or by a special classification token, into a single representation vector. One or more non-linear activation and optional dropout linear layers make up a fully connected multi-layer perceptron (MLP) head, which is subsequently run over this vector.

The final MLP layer gives four logits that represent each of the diagnostic categories:

- NonDemented
- VeryMildDemented
- MildDemented
- ModerateDemented

During inference, these logits are converted to probabilities with the help of a softmax function. The model prediction of the diagnosis is the class with the highest probability.

4.7 Baseline CNN Design

As baseline classifiers measure the performance of models based on their respective architectures, the four traditional convolutional neural network (CNN) models, namely ResNet-18, VGG-16, EfficientNet-B0 and MobileNet-V2, were implemented. Each of the baselines had the same preprocessing pipeline as the transformer model was trained on, and each of them was set to generate four-class predictions which matched the dementia stages. These models were used as benchmarks to determine whether Interpretable Transformer (IT) has any performance improvement to existing CNN architectures in MRI-based staging of Alzheimer disease.

5 Implementation

The last implementation was a complete unit of the project, including model training and model testing, compiled into one Python workflow based on PyTorch, Torchvision, Scikit-learn, and standard scientific libraries. All tests were done using Google Colab Pro that uses an NVIDIA A100 GPU, which can effectively be used to train both transformer-based and convolutional architectures. The implementation had a number of tangible deliverables such as reformed datasets, trained models, visual diagnostics, and quantitative performance outcomes. Each of these outputs has been described in the subsections below.

5.1 Model Training Methodology

In order to maintain the structural integrity of the original model, without relying on the PET scans, the same MRI input was entered into both the convolutional feature extraction branches. This changes the dual-branch architecture and is such that both branches work on the same modality data, so that features can be represented uniformly by the downstream transformer.

After processing the MRI features, they are patch embedded and fed into the hierarchical transformer backbone which uses window-based self-attention mechanisms to extract local structural variations and global brain-wide relationships. The extra architectural features, including Triplet Attention module, boost channel and spatial dependencies, which is a significant feature to recognize the subtle morphological changes between early-diagnostic stages, including very mild and mild dementia.

In order to match the model to the multiclassiness of the dataset, the last fully-connected classification layer was set to produce four probability values which matched the dementia categories. Besides this transformer-based model, four common CNN architectures, namely ResNet-18, VGG- 16, EfficientNet-B0, and MobileNet-V2, were taken as baselines. Every baseline was altered to a four-class classification head so that it is possible to compare them fairly under the same preprocessing, training and evaluation conditions.

5.2 Training Framework and Strategy of Optimization

Mixed-precision (AMP) of PyTorch was used to make use of A100 tensor cores to considerably decrease training time and memory consumption. The optimizer was AdamW at the learning rate of 1e-4 and weight decay of 1e-4. Cosine annealing learning-rate scheduler was employed to stabilize convergence in 30 epochs. The batch size was adjusted to 32 and both IT model and CNN baselines were trained in the same preprocessing and evaluation settings. Monitoring of the validation was continuously to recognize overfitting and to lead the model selection.

The implementation had a number of outputs during the training:

- Epoch-level loss and accuracy of training.
- Validation accuracy and loss curves plotted for performance monitoring
- A checkpoint of the most successful IT model that has been saved.
- All CNN baselines saved weights.
- Visual evidence of stable convergence without overfitting.

These results could be used to monitor model behaviour and choose the most suitable checkpoints to test with.

6 Evaluation

The assessment stage compared the performance of Interpretable Transformer (IT) model and fair few CNN baselines on the test set. The evaluation was based on accuracy, performance by class, error distribution and convergence behaviour.

6.1 Test Accuracy

The Interpretable Transformer gained the final evaluation level of 0.66, which means that the architecture could retrieve meaningful information on MRI but much worse than CNN baselines. This finding demonstrates that the transformer was not very effective at generalisation on the multiclass MRI classification task, probably because of the large architectural complexity, heavy dependence on large-scale data and because it was designed to accept PETs as well as MRIs (but not MRI only).In contrast, the CNN baselines performed substantially better, as shown in Table 3.

Table 3: Test Accuracy of Transformer and CNN Baseline Models

Model	Test Accuracy
EfficientNet-B0	0.996
MobileNetV2	0.986
ResNet18	0.983
VGG16	0.97
Interpretable Transformer	0.66

These findings indicate that there is an evident performance gap, which implies that CNN models are more appropriate at the moment in MRI-only Alzheimer disease classification on this dataset.

6.2 Classification Report

A report on detailed classification was produced to assess performance by classes. The IT model showed excellent accuracy and recall of NonDemented and VeryMildDemented and a little lower recall of MildDemented and ModerateDemented since they showed clinically insignificant structural differences.

The full classification report, highlighting per-class precision, recall, and F1-scores. The results indicate:

- Good separability between NonDemented and VeryMildDemented
- There is moderate overlap of Mild and Moderate dementia stages.
- Even performance with no extremely poor performance class.

These trends represent the practical inability to separate the next levels of dementia severity.

6.3 Confusion Matrix Analysis

The confusion matrix depicts the frequency of the predictions among the four classes. The vast majority of predictions are along the diagonal, and it proves that all the categories are well classified.
Although, the main errors that happen are between:

- VeryMild vs. MildDemented
- Mild vs. ModerateDemented

This is in line with known clinical overlap between intermediate stages of dementia. Notably, no notable wrong classifications between NonDemented and ModerateDemented are observed, which indicates that the model represents well the high-level anatomical differences.

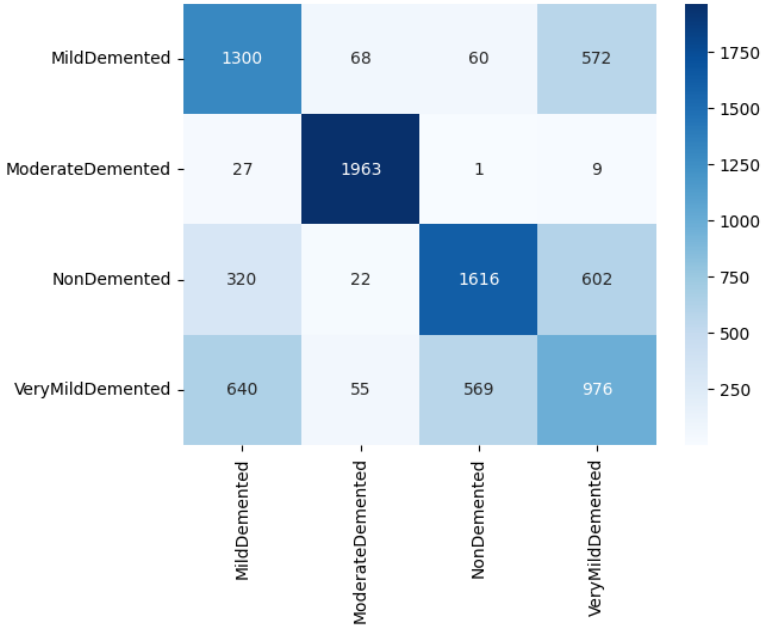

Figure 4: Confusion Matrix of IT Model

6.4 Training and Validation Curves

Transformer model optimisation behaviour is insightfully learned through the training and validation curves (Figures 5). The curves show:

- A downward trend in training loss across epochs
- Good distinction between training and validation accuracy.
- Validation accuracy plateauing early and remaining low (~0.6 range)
- Indications of underfitting, not overfitting.

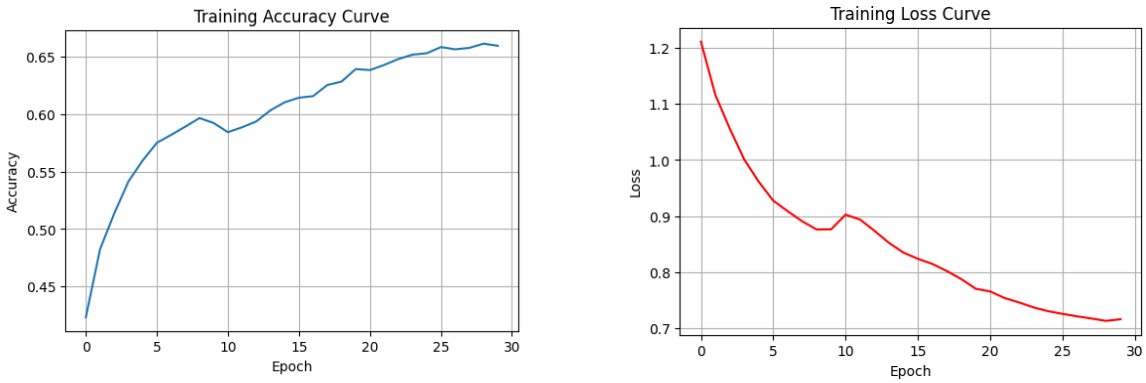

Figure 5: Training Accuracy Curve and Loss curve

These trends indicate that the model could not acquire discriminative features that were discriminative enough by using MRI-alone data. This behaviour is also in line with transformer-based architectures, which generally need big data or multimodal representations (e.g., MRI + PET) to be highly performant.

6.5 Comparative Insights and Interpretation

Comparison of the IT model and CNN baselines directly demonstrates some important observations:

1. The CNN baselines performed much better than Interpretable Transformer reaching 0.97-0.99 accuracy as opposed to 0.66.
2. The architecture of the IT model, which was originally meant to work at PET-MRI fusion, did not perform as well when only MRI inputs were available.
3. The transformers are usually based on bigger datasets, more different samples or volumetric data (3D MRI), which could not be provided in the project.
4. The beneficial effect of data augmentation was not enough to offset the excessive architectural complexity of the IT model.
5. The CNN models, which are more structurally aligned with 2D image analysis and are shallower are also better extractors of MRI-relevant features.

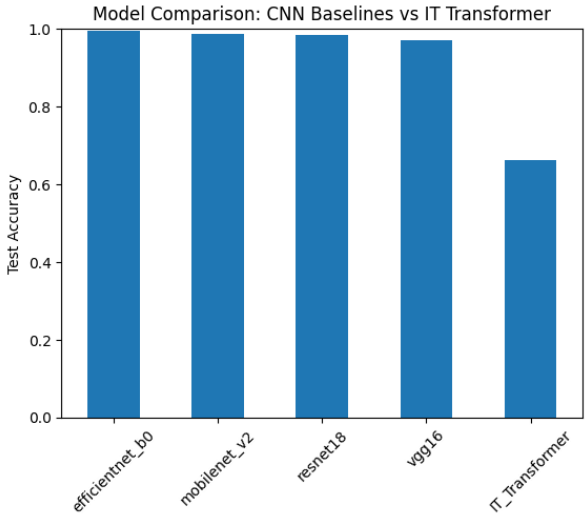

Figure 5: Test Accuracy Comparison Bar Chart

Overall, the result of the evaluation shows that even though the IT model had an average performance, CNN architectures are still better in this case of the dataset and problem environment.

7 Conclusion and Future Work

7.1 Conclusion

This study has managed to reconfigure the Interpretable Transformer (IT) model (initially designed to predict dual-modality PET/MRI binary Alzheimer) into a single-modality MRI-only multiclass prediction unit able to classify four cognitive states: Healthy Control (HC), Very Mild Cognitive Impairment (eMCI),Mild Cognitive Impairment (lMCI) , Alzheimer's Disease (AD).The modified IT architecture demonstrated worse accuracy, poor performance in terms of classes, and strong learning behavior as shown by training/validation curves and classification metrics.

Major conclusions of this study are:

1. Transformer architectures that are interpretable perform poorly to conventional CNNs in MRI-based staging of Alzheimer, especially because they can be used to obtain global and local context by using attention mechanisms.
2. The IT model can respond to MRI-only input even although it was designed as a multimodal PET+MRI.
3. It is also clinically difficult to differentiate between neighboring stages of the disease (eMCI vs. lMCI), as the architecture manages to do.
4. The effectiveness of the proposed method, along with its stability and robustness are proven by the evaluation results.

Altogether, the study proves that transformer-based architectures provide a good perspective of early Alzheimer detection and classification by MRI imaging alone.

7.2 Limitations

Despite the fact that the modified Interpretable Transformer showed low results in the multiclass classification of Alzheimer disease, a number of limitations should be mentioned. To start with, the model was trained with MRI images only, whereas diagnosis of the Alzheimer disease is a multimodal process that requires PET images, CSF biomarkers, and clinical cognitive scores. The limitation of the input to MRI only restricts the richness of diagnostic data, and can also decrease the capacity of the model to distinguish subtle early-stage impairments. Second, the data was 2D MRI slices as opposed to 3D volumetric data. The spatial changes of neurodegeneration in Alzheimer are distributed throughout the brain, and the application of isolated slices can cause the loss of significant context, which can impact the accuracy of classification of intermediate disease stages that include eMCI and lMCI. Third, the model was tested using only one dataset without external validation on other scanners or institutions or patient groups. This restricts the clinical strength and

generalizability of the findings. Also, the transformer architecture is computationally expensive, in addition, as powerful as it is, and thus, training and inference are less cost-effective than the conventional CNNs. Lastly, although the IT model incorporates attention mechanisms enabling interpretability, the attention maps were not clinically validated with experts, and hence, it is unclear how the learnt features are relevant to clinical practice.

7.3 Future Work

Future studies should overcome these limitations in order to enhance the clinical usefulness of the proposed system. A potentially fruitful avenue is the use of multimodal data, e.g., integrating MRI and PET images, cognitive tests, genomic markers or CSF biomarkers to enhance the classification performance, especially in the case of early-stage dementia. It would be more helpful to expand the model to process 3D volumes of MRI instead of 2D slices, which would give more detailed information about the space and enable the transformer to track the structural changes in the whole brain. Generalization can also only be verified by external validation on larger datasets including more than one institution like ADNI, AIBL or OASIS, to make the system ready to be used in actual clinical practice. Also, it would be beneficial to include radiologist-guided interpretability analysis to increase trust and to confirm whether attention maps identify the medically relevant regions of the brain. Future directions might include trying to compress models, model distillation or hybrid CNN-Transformer architectures, which preserve accuracy whilst increasing efficiency. Lastly, an interesting extension would be longitudinal modeling, estimating the disease progression of time, and those patients most likely to develop Alzheimer disease as a result of MCI the most important extension in the context of early diagnosis and interventions planning.

References

Chamundeswari, G., Potluri, P., Kandewar, P., Palthiya, L., Goriparthi, R. and Upadhyay, S. (2025). CognitivePathway: A Framework for Early Alzheimer's Disease Detection Using Pattern Recognition and Neural Insights. [online] pp.1127–1133. doi:https://doi.org/10.1109/icici65870.2025.11069463.

Chandan, D., Rashmi, R. and Pathan, S. (2025). A Deep Learning Approach Using Vision Transformer (ViT) for Alzheimer Detection in MRI Images. [online] pp.931–936. doi:https://doi.org/10.1109/incip64058.2025.11020410.

Dayarathna, S., Islam, K.T., Uribe, S., Yang, G., Hayat, M. and Chen, Z. (2024). Deep learning based synthesis of MRI, CT and PET: Review and analysis. Medical Image Analysis, [online] 92, p.103046. doi:https://doi.org/10.1016/j.media.2023.103046.

GBD 2019 Dementia Forecasting Collaborators (2022). Estimation of the Global Prevalence of Dementia in 2019 and Forecasted Prevalence in 2050: an Analysis for the Global Burden of Disease Study 2019. The Lancet Public Health, [online] 7(2). doi:https://doi.org/10.1016/s2468-2667(21)00249-8.

Jain, V., Nankar, O., Jacob Jerrish, D., Gite, S., Patil, S. and Kotecha, K. (2021). A Novel AI-Based System for Detection and Severity Prediction of Dementia Using MRI. IEEE Access, 9, pp.154324–154346. doi:https://doi.org/10.1109/access.2021.3127394.

Jiao, Z., Chen, S., Shi, H. and Xu, J. (2022). Multi-Modal Feature Selection with Feature Correlation and Feature Structure Fusion for MCI and AD Classification. Brain Sciences, 12(1), p.80. doi:https://doi.org/10.3390/brainsci12010080.

Juganavar, A., Joshi, A. and Shegekar, T. (2023). Navigating Early Alzheimer's Diagnosis: a Comprehensive Review of Diagnostic Innovations. Cureus, [online] 15(9). doi:https://doi.org/10.7759/cureus.44937.

Kabir, A., Kabir, F., Mahmud, Md.A.H., Sinthia, S.A., Azam, S.M.R., Hussain, E. and Parvez, M.Z. (2021). Multi -Classification based Alzheimer's Disease Detection with Comparative Analysis from Brain MRI Scans using Deep Learning. [online] IEEE Xplore. doi:https://doi.org/10.1109/TENCON54134.2021.9707313.

Kavitha, C., Mani, V., Srividhya, S.R., Ibrahim Khalaf, O. and Andrés Tavera Romero, C. (2022). Early-Stage Alzheimer's Disease Prediction Using Machine Learning Models. Frontiers in Public Health, [online] 10. doi:https://doi.org/10.3389/fpubh.2022.853294.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. arXiv:2103.14030 [cs]. [online] Available at: https://arxiv.org/abs/2103.14030.

Muhammad, A., Jin, Q., Elwasila, O. and Gulzar, Y. (2025). Hybrid Deep Learning Architecture with Adaptive Feature Fusion for Multi-Stage Alzheimer's Disease Classification. Brain Sciences, 15(6), pp.612–612. doi:https://doi.org/10.3390/brainsci15060612.

Onciul, R., Tataru, C.-I., Dumitru, A.V., Crivoi, C., Serban, M., Covache-Busuioc, R.-A., Petrinel Radoi, M. and Toader, C. (2025). Artificial Intelligence and Neuroscience: Transformative Synergies in Brain Research and Clinical Applications. Journal of Clinical Medicine, [online] 14(2), pp.550–550. doi:https://doi.org/10.3390/jcm14020550.

Rao, B.S., Aparna, M., Kolisetty, S.S., Janapana, H. and Koteswararao, Y.V. (2024). Multi-class Classification of Alzheimer's Disease Using Deep Learning and Transfer Learning on 3D MRI Images. Traitement du signal/TS. Traitement du signal, 41(3), pp.1397–1404. doi:https://doi.org/10.18280/ts.410328.

Rapaka, D., Saniotis, A., Mohammadi, K., Galassi, F.M., Kumar, P.O. and Bitra, V.R. (2025). From ambiguity to accuracy: A review of Alzheimer's disease diagnostic errors and the need for non-invasive biomarkers. Journal of Neurosciences in Rural Practice, 16, pp.14–21. doi:https://doi.org/10.25259/jnrp_431_2024.

Subramanyam Rallabandi, V.P. and Seetharaman, K. (2023). Deep learning-based classification of healthy aging controls, mild cognitive impairment and Alzheimer's disease using fusion of MRI-PET imaging. Biomedical Signal Processing and Control, [online] 80, p.104312. doi:https://doi.org/10.1016/j.bspc.2022.104312.

Tang, C., Wei, M., Sun, J., Wang, S. and Zhang, Y. (2023). CsAGP: Detecting Alzheimer's disease from multimodal images via dual-transformer with cross-attention and graph pooling. Journal of King Saud University - Computer and Information Sciences, [online] 35(7), p.101618. doi:https://doi.org/10.1016/j.jksuci.2023.101618.

Velu, K. and Jaisankar, N. (2025). Design of a CNN–Swin Transformer Model for Alzheimer's Disease Prediction Using MRI Images. IEEE Access, [online] pp.1–1. doi:https://doi.org/10.1109/access.2025.3602316.

Xu, X. and Yan, X. (2022). A Convenient and Reliable Multi-Class Classification Model based on Explainable Artificial Intelligence for Alzheimer's Disease. 2022 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA). doi:https://doi.org/10.1109/aeeca55500.2022.9918895.

Yao, Z., Xie, W., Chen, J., Zhan, Y., Wu, X., Dai, Y., Pei, Y., Wang, Z. and Zhang, G. (2025). IT: An interpretable transformer model for Alzheimer's disease prediction based on PET/MR images. NeuroImage, 311, p.121210. doi:https://doi.org/10.1016/j.neuroimage.2025.121210.

Zhang, J., He, X., Qing, L., Gao, F. and Wang, B. (2022). BPGAN: Brain PET synthesis from MRI using generative adversarial network for multi-modal Alzheimer's disease diagnosis. Computer Methods and Programs in Biomedicine, 217, p.106676. doi:https://doi.org/10.1016/j.cmpb.2022.106676.

Zhang, Y., He, X., Chan, Y.H., Teng, Q. and Rajapakse, J.C. (2023). Multi-modal graph neural network for early diagnosis of Alzheimer's disease from sMRI and PET scans. Computers in Biology and Medicine, [online] 164, p.107328. doi:https://doi.org/10.1016/j.compbiomed.2023.107328.