Multimodal Transformer Fusion for Alzheimer's Disease Detection: Comparative Analysis of Image Modality Performance and Architectural Efficacy

G077 (s2134605, s2497228, s2260856)

Abstract

Alzheimer's disease (AD) poses a significant global health challenge, and demands improved diagnostic and therapeutic interventions as the elderly population grows worldwide. Traditional neuroimaging, and uni-modal models falls short in understanding AD's complexity, leading to interest in new multimodal approaches. We propose a novel architecture to combine PET, MRI, CT scan data and using CNN and UNet encoders and intermediate fusion using transformers to improve the diagnostic accuracy. Our study aims to combine these techniques for comprehensive AD diagnosis, resulting in a new multimodal fusion approach which enhances diagnosis accuracy, potentially improving patient outcome.

1. Introduction

Alzheimer's disease (AD) is a serious global health concern that affects about 55 million people and is primarily responsible for 60–70% of dementia cases in the elderly people (WHO, 2023). As the number of senior citizens grows worldwide, the need for efficient diagnostic and treatment interventions grows increasingly pressing. Even with improvements in our knowledge of the intricate neuropathology of AD, prompt diagnosis and treatment are still essential to altering the progression of the disease.

Hence, identifying AD early and accurately is imperative due to its complex nature. However, reliance on single-modal neuroimaging methods like Positron Emission Tomography (PET) and Magnetic Resonance Imaging (MRI) has limitations, often providing only a fragmented understanding of AD pathophysiology(Gharaibeh et al., 2022). This shortfall has caused the exploration of multimodal neuroimaging approaches.

Our study introduces a novel fusion of transformers and deep learning models within a multimodal neuroimaging framework for diagnosing Alzheimer's disease (AD). By leveraging MRI, CT, PET scans, and clinical assessment, we seek to advance beyond previous methods, gaining a deeper understanding of AD pathology and addressing current limitations. Our approach aims to offer a comprehensive diagnostic architecture by synthesizing high-dimensional data from multiple imaging modalities, thus providing a more complete picture of AD-related neurodegenerative alterations.

Deep learning models and transformers were chosen due to their demonstrated effectiveness in managing high-dimensional datasets and their ability to extract and synthesise features from a variety of data formats. These approaches are especially well-suited to the difficulties posed by multimodal neuroimaging data in AD research, and they have demonstrated remarkable potential in a number of medical imaging fields.

Our research is motivated by the following research question:

Enhancing Diagnosis Accuracy

Research Question: How might the accuracy and thoroughness of AD diagnosis be enhanced by integrating MRI scans, CT scans, PET scans, using attention mechanisms and clinical assessments using sophisticated deep learning models?

Objective: The goal is to create and evaluate a deep learning-based framework that combines MRI, CT, PET, and clinical assessment data in a synergistic way to provide a more comprehensive and accurate diagnosis of Alzheimer's disease.

Our paper is organised as follows: **Section 2** describes the evaluation techniques, data processing, and dataset in depth. Our study technique is described in **Section 3**. We describe our experimental setup and findings in **Section 4**. To put our work in context, we review relevant literature in **Section 5**. The final section, **Section 6**, highlights the field-wide implications of our study by summarising our contributions and outlining potential avenues for future research.

2. Data set and task

2.1. Dataset Selection

The data sets that will be used for this project are:

OASIS Brains (Open Access Series of Imaging Studies): The Oasis project provides multiple open source datasets (OASIS 1-4) which give longitudinal and cross-sectional multimodal neuroimaging data, clinical information, cognitive scores, and more. These datasets are intended for targeting different aspects of brain research, including aging, dementia, and Alzheimer's disease. For our study, we exclusively utilize the OASIS datasets. Table 1 1. provides an overview of the demographic characteristics of the subjects included in these datasets.

OASIS-1: Cross-sectional data from 416 right-handed people, ages 18 to 96, who underwent three or four T1-weighted MRI scans in a single session, are included in the collection. It consists of 20 nondemented people who were rescanned within 90 days and 100 adults over 60 with very mild to severe Alzheimer's disease (AD). Research on adult neurodegeneration, early identification of AD, and brain ageing would benefit greatly from this dataset (Marcus et al., 2007).

OASIS-2: The dataset includes three or four T1-weighted MRI scans each session for 150 right-handed people between the ages of 60 and 96. 373 imaging sessions were performed on the subjects at yearly intervals. Of these, 64 remained nondemented (51 with mild to moderate Alzheimer's), 14 changed from nondemented to demented, and 64 remained demented consistently. Age-related changes in brain structure and the course of Alzheimer's disease are good subjects for OASIS-2 study (Marcus et al., 2010).

OASIS-3: With the help of WUSTL Knight ADRC, OASIS-3 provides longitudinal multimodal neuroimaging by merging PET and MRI data from 1,378 subjects over a 30-year period. There are 622 people with cognitive deficits and 755 adults with normal cognitive function. There are 2,842 MRI sessions (T1w, T2w, T2star), 1,472 CT scans, and 2,157 PET (AV45, PIB, FDG) scans. This dataset is essential to the study of Alzheimer's disease progression and brain ageing (LaMontagne et al., 2019).

OASIS-4: Distinct from OASIS-3, the Clinical Cohort dataset is centred on clinical research and consists of 663 patients between the ages of 21 and 94 who were evaluated for dementia and memory problems using neuropsychometric, clinical, and CSF examinations. It backs up long-term studies on the course of Alzheimer's, alterations in biomarkers, and the effects of medication (Koenig et al., 2020).

Data	Class	MRI	PET	CT	Clinical
OAS1	Number	436	-	-	-
	Gender(M/F)	168/268	-	-	-
	Age	18-96 mean: 51.36			
OAS2	Number	373	-	-	373
	Gender(M/F)	160/213	-	-	160/213
	Age	60-98 mean: 77.01			
OAS3	Number	1376	888	1063	1378
	Gender(M/F)	487/611	269/353	344/444	487/611
	Age	42-95			
OAS4	Number	688	-	-	663
	Gender(M/F)	330/338	-	-	330/338
	Age	21-94			

Table 1. **The demographic characteristics of our dataset**

2.2. Modalities Utilized

The OASIS datasets' inclusion modalities can differ, due to the different research objectives of each dataset. An outline of the main modalities employed in this study is provided in Table 2.

Modality	Dataset of Modality	Advantages	Disadvantages
MRI	*All*	- High resolution and contrast of brain structures - Non-invasive - Effective for identifying brain atrophy and structural changes.	- Time-consuming - Expensive - May not detect early microscopic changes.
PET	*OASIS3*	- High sensitivity for detecting metabolic changes can identify early functional changes before structural abnormalities appear.	- Expensive - Invasive due to radioactive tracer - Limited availability.
CT	*OASIS3*	- Widely available - Faster and cheaper than MRI and PET.	- Lower resolution compared to MRI and PET - Less effective for early detection of AD due to poor contrast in soft tissues.
Clinical	*All*	- Non-invasive - Can be performed quickly provides immediate feedback on cognitive function.	- Subjective - Performance can be influenced by patient's mood, education, and other factors - Not as sensitive as neuroimaging in early stages of AD.

Table 2. **Comparison of imaging modalities for Alzheimer's Disease detection (Aramadaka et al., 2023) (Kim et al., 2022) (Zhang et al., 2020)**

2.3. Data Preprocessing

2.3.1. Preprocessing and Feature Extraction

The modalities were divided into two categories: neuroimaging modalities and text modalities.

***A.* Neuroimaging Modalities (MRI, PET, CT Scans):** Preprocessing is essential for each of these imaging modalities in order to guarantee clean, participant-level data. This entails actions such as neuroimage conversion, reshaping image, and filtering.

***A.1.* For the preprocessing of neuroimaging data for Alzheimer's disease detection, the following steps were taken:**

1. **Conversion of NII scans to PNG format:**
   For MRI and CT: In order to capture important Alzheimer's regions like the ventricles, entorhinal cortex, cerebral cortex, and hippocampus—regions important for memory and early disease markers—NII scans were converted to PNG. Slices 100 to 160 were specifically targeted. When a condition progresses, axial slices can be used to identify cortical alterations and ventricular enlargement.

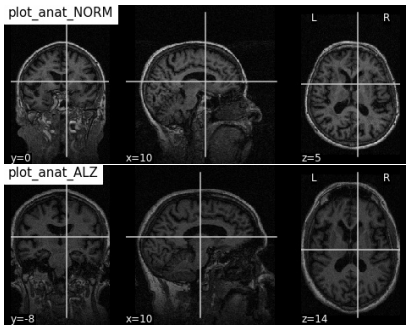

Figure 1. **Comparison of brain images depicting various planes between a healthy brain and one affected by Alzheimer's disease (AD) (Palko, 2016)**

For PET: Targeting amyloid plaques in the brain, AV45 (florbetapir) and PIB (Pittsburgh Compound B) PET Scans are applied, with middle to upper slices (40–70% of the scan) being the primary focus. Reduced FDG uptake, which requires 30% to 80% slice coverage, indicates Alzheimer's-related hypometabolism, particularly in the temporal and parietal lobes and posterior cingulate cortex. FDG measures brain metabolic activity. In order to facilitate additional analysis such as edge identification and feature extraction for machine learning, images are smoothed with a Gaussian filter to reduce noise, using weighted averages for pixel intensity (Bharathi & Arunachalam, 2021).

2. **Utilization of the Clinical Dementia Rating (CDR) Scale from Clinical Assesment:** The CSV files accompanying the slices contained data on the Clinical Dementia Rating (CDR) Scale, a tool used to evaluate the degree of dementia. Scores range from 0 (indicating normal cognitive function) to 2 (indicating substantial dementia). By categorizing the images based on participants' cognitive performance using the CDR scale, a more comprehensive study was conducted, considering the extent of cognitive decline in the analysis.

3. **Standardization of image size:** Each image was resized to a consistent 128x128 pixel dimension. Standardizing the image size ensures uniformity in the data fed into the model, optimizing conditions for precise analysis and effective model training. This standardization minimizes the impact of image size variations on the functionality of the model, thereby enhancing the accuracy of the Alzheimer's disease diagnosis method.

4. **Train-test split:** 80% of the data were used in the train-test split for training, with the remaining 20% set aside for testing and confirming the model's performance.

Table 3 shows the total images for each stage of the AD for each modality i.e. after the first two processing steps have been done.

AD Stage	MRI	CT	PET
Non-demented	411,631	64,899	76,242
Very mild dementia	122,133	7,852	8,296
Mild dementia	43,366	3,480	4,585
Moderate dementia	14,568	2,520	4,261

Table 3. **Total images for each AD stage per modality**

B. **Text Modalities(Clinical Assessments):** Standardisation of clinical assessment scores to ensure comparability of cognitive scores by taking participant baseline differences into consideration.

B.1. **For the text modality, the following step was taken:** Determine which characteristics—such as CDR and date of session—are most important in the diagnosis of Alzheimer's disease. This lowers the number of dimensions and concentrates the investigation on the most instructive elements.

2.3.2. Integration and Analysis

Multimodal Integration: To combine features from different modalities, multimodal deep learning fusion techniques are used. This method may improve the model's capacity to represent the intricacy of Alzheimer's disease.

2.4. Evaluation Metrics

Performance Metrics:
To guarantee a thorough grasp of the model's efficacy, it is crucial to assess it using a variety of performance criteria. The accurate categorization of AD in comparison to MCI or healthy controls is measured by accuracy. Specificity assesses the model's capacity to accurately identify healthy or non-AD persons, and sensitivity (recall) measures the model's efficacy in recognising actual AD cases. Through the combination of sensitivity and specificity, the Area Under the ROC Curve (AUC) provides a comprehensive performance metric. To ensure that the model is accurate in recognising real AD instances, precision computes the percentage of accurately predicted AD cases among all AD predictions. The combination of these measures yields a detailed assessment of the diagnostic precision of the model.
Comparison with Baseline Model:
On the same dataset, we evaluate our multimodal system's performance against existing multimodal approaches and well-established single-modality methods to identify any notable gains in specificity, sensitivity, or accuracy that our system makes.

3. Methodology

3.1. Background

3.1.1. Stages of the Alzheimer's Disease

This is a broad summary of the stages involved in detecting Alzheimer's disease, emphasizing the use of multimodal approaches, which integrate various data sets such as biomarkers, neuroimaging, clinical assessments, and occasionally genetic data to improve diagnostic accuracy, particularly in the early stages of the illness when therapies can be more successful (AlMansoori et al., 2024):

1. **Preclinical Alzheimer's Disease:** Blood tests, CSF analysis, and PET/MRI revealed biological indicators in the absence of symptoms.

2. **Very-Mild Cognitive Impairment (MCI):** Slight cognitive alterations with minimal daily impact. Can be tested by MRIs, biomarker analyses, and cognitive tests.

3. **Mild AD:** Obvious cognitive and memory problems affecting day-to-day functioning, identified by PET/MRI, biomarkers, and clinical examinations.

4. **Moderate to Severe AD:** Significant cognitive impairment that needs daily attention; determined by biomarkers, neuroimaging, and clinical assessments for modifications to treatment.

Multimodal AD detection integrates various data for precise diagnosis and treatment, enhancing life quality for patients and families, with ongoing research improving detection capabilities (Fang et al., 2023).

3.1.2. Multimodal Data Fusion

Multimodal Data Fusion is the combining of data from several sources or formats (such as text, graphics, or sound) to enhance comprehension, learning, or decision-making inside a system. Benefits include higher resistance to noise or missing data in one or more modalities, improved reliability through cross-validation among modalities, and improved performance accuracy by utilising complementary information. More thorough analysis and insights can be obtained with this method than with single-modality data processing (Lahat et al., 2015).

Multimodal Approaches:

A. Traditional approaches for multimodal fusion

Multimodal fusion frequently involves early, late, and intermediate fusion. Early fusion facilitates early learning of inter-modal interactions by merging modalities at the input level. Decisions are combined via late fusion once modalities have been processed separately. Partially combining the properties of different modalities, intermediate fusion strikes a balance between intricate linkages and modality-specific data. These approaches use statistical methods, decision trees, neural networks, and other machine learning techniques to apply to a variety of tasks and data types. Figure 3 shows the different fusion procedures.(Sharma & Mandal, 2022)

B. Deep learning-based multimodal fusion (CNN, Autoencoders, GAN, transformers)

Transformers, autoencoders, GANs, and CNNs are used in deep learning-based multimodal fusion to combine inputs from several sources (Kalamkar & A., 2023). While autoencoders lower dimensionality and facilitate simpler fusion, CNNs handle spatial data in photos and videos. Synthetic data generated by GANs improves the resilience of fusion (Ye et al., 2023). Transformers are helpful for deriving context from many data kinds because they employ self-attention to record dependencies across modalities (Lin et al., 2022). The employment of transformers for multimodal fusion is the main topic of this study.

Multimodal fusion using transformers

Transformer architectures enable multimodal fusion, combining text, image, and audio data. It handles sequential data and determines the relative relevance of various data elements by utilising the transformer's self-attention mechanism. For applications like voice recognition, video interpretation, and picture captioning, this method allows for efficient data integration, improving model robustness and accuracy. It captures intricate intermodal linkages, which enhances the model's performance on multimodal tasks (Prakash et al., 2021). We are proposing the architectures mentioned in the next section for our project.

3.1.3. Deep learning CNN Models

A fundamental problem in Alzheimer's disease identification using different brain imaging scans is performing an effective feature extraction to discriminate between different brain areas. A straightforward technique would be to use a CNN to extract initial features before putting them into a transformer layer for fusion and an MLP layer for classification. However, this if not done properly would result in bad feature extraction and also increase the model complexity, resulting in longer training and the need for large amounts of labelled data, which is limited in medical imaging domains. To address this issue, we utilized a pre-trained U-Net model for feature extraction. The U-Net, which was created for brain tumour segmentation (Buda et al., 2019), recovers relevant information from a 3-channel MRI slice as input and generates a 1-channel probability map that highlights possible abnormality regions. This strategy uses transfer learning from a comparable MRI segmentation model which will reduce model complexity and training time.

3.2. Our Proposed Hybrid Deep Learning Architectures

This study makes use of a brand-new multimodal deep learning framework designed especially for integrating and analysing various medical imaging data sets. Our model uses imaging data from various modalities such as T1 and T2-weighted MRI scans, PET scans (PIB, AV45 and FGD), and CT scans. To ensure consistency throughout the dataset, every imaging modality goes through preliminary processing to standardise image dimensions and format. After processing the scans, each modality is fed to an encoder, which is the downsampling path of a Unet (Ronneberger et al., 2015). Fundamentally, the architecture uses a *pre-trained UNet* model for each modality to extract features, using its handling of spatial hierarchies and details to capitalise on UNet's demonstrated expertise in medical picture analysis. Initially, a non-trained UNet was used, but the resulting accuracy was low, so a trained UNet was used instead. This captures essential information and transform the images into a lower dimensional representation. In order to encode these extracted features into a single, lower-dimensional space suitable for transformer processing, each one is thereafter independently run through a *Multi-Layer Perceptron (MLP)*.

The architecture uses a Transformer Encoder to process the concatenated features from all modalities after feature encoding. In order for the model to build more robust representations that are sensitive to the subtle variations between the stages of dementia, this step is essential for capturing complex inter-modal interactions and dependencies. One of the four predetermined phases of dementia is the final classification that an additional MLP classifier processes from the Transformer's output. To ensure that the final predictions are probability distributions over the available classes, this MLP is composed of linear layers with ReLU activation functions for the intermediate levels and a softmax activation for the output layer.

Further information regarding the model is in 4, and our suggested architecture is shown in **Figure 2**.

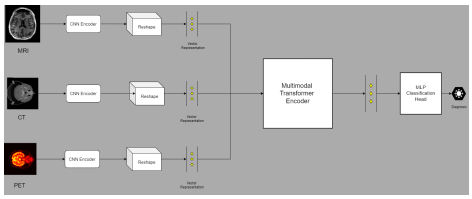

Figure 2. **Our Proposed Architecture**

1. **Modality-Specific Encoders:** Each modality is first processed by particular encoders made to record features unique to that modality. Two approaches were tried for the encoder:
   - CNNs: CNNs are used as the encoders.
   - UNets: The downsampling paths of Unets are used as these encoders.
   Both approaches process the input images and convert them into a standardised feature representation.
2. **Multimodal Transformer Encoder:** A Transformer encoder receives the concatenated outputs from the modality-specific encoders. By making use of the Transformer's capacity to record intricate relationships between modalities, this encoder further unifies the data from several modalities. The Transformer encoder is made up of multiple layers of Transformer encoder blocks, each of which has multi-head self-attention mechanisms and position-wise completely coupled feed-forward networks.

The final disease classification or diagnostic prognosis is obtained by passing the output from the multimodal Transformer encoder via a number of fully linked layers and a classification layer.

Modifications and Innovations:
We present a novel approach to multimodal medical imaging by expanding the Transformer model, which has been used primarily for sequence-based data. This modification makes it possible to analyse MRI, CT, and PET scans simultaneously—a combination that has never been done before in the field of medical diagnosis. With the help of the UNet architecture for the first modality-specific feature extraction and a multimodal Transformer encoder for the integration of these features, our model is able to capture both the complex inter-modal dynamics and the fine-grained spatial relationships within individual images. This innovative strategy makes it easier to comprehend the data at a deeper level and may reveal underlying relationships between modalities that are missed by more conventional approaches. Medical imaging has advanced when it can now process and synthesise data from several imaging sources in a single framework.

4. Experiments

4.0.1. Results

MRI: With a test accuracy of 68.52% and a rather high test loss of 0.903, the assessment of the pretrained model's second iteration for dementia stage classification shows mixed results, indicating potential for further development in the model's confidence and generalisation. With a precision of 0.81 and recall of 0.89, the model shows a

Algorithm 1 Training a Multimodal Transformer Model

Input: Multimodal data x_i, labels y_i, batch size b, learning rate lr, epochs E
Initialize UNet-based Encoders $Encoder_{UNet}$ for MRI, CT, and PET modalities
Initialize TransformerEncoder $T_{encoder}$ and MLP classifier C for final classification
Output: Trained Multimodal Transformer Model

for epoch = 1 **to** E **do**
  Initialize epoch loss $L_{epoch} = 0$ & accuracy $A_{epoch} = 0$
  for all batches (x_{batch}, y_{batch}) in dataset **do**
    Split x_{batch} into modalities (MRI, CT, PET)
    Process each modality through $Encoder_{UNet}$ to obtain encoded features e_{MRI}, e_{CT}, e_{PET}
    Flatten and transform encoded features through MLP to match transformer input size
    Concatenate transformed features: $e_{combined} = Concat[e_{MRI}, e_{CT}, e_{PET}]$
    Process combined features through $T_{encoder}$: $t_{features} = T_{encoder}(e_{combined})$
    Classify with MLP: $logits = C(t_{features})$
    Calculate loss: $L = \text{CrossEntropyLoss}(logits, y_{batch})$
    Backpropagate error and update model parameters
    Update epoch loss $L_{epoch} += L$ and compute batch accuracy
  end for
  Calculate average epoch loss and accuracy
  Print epoch summary including loss and accuracy
end for

Parameter	Description
Convolutional Layers in Encoder	3 Layers with filters (32, 64, 128) and kernel size 3
Pooling Layers in Encoder	MaxPooling with size 2
Encoder Output Size	Varies (depends on input image size)
Pre-trained UNet	Loaded from 'mateuszbuda/brain-segmentation-pytorch', frozen parameters
Transformer Heads	4
Transformer Hidden Size	64
Transformer Layers	1
Transformer Dropout	0.1
MLP Classifier Layers	Linear layers with sizes [Transformer Input Size * 3, 64, 32, Num Classes]
Activation Functions	ReLU for intermediate layers Softmax for output layer
Loss Function	CrossEntropyLoss
Optimizer	Adam

Table 4. **Parameters of the Multimodal Transformer Model**

Aspect	Description
Model Architecture	ResNet50 adapted for 3D inputs.
Input Shape	(128, 128, 128, 1)
Output Layer	Dense layer with 4 units and softmax activation function.

Table 5. **Model Parameters Overview**

Aspect	Description
Optimizer	Adam
Learning Rate	0.00005 (5e-5)
Loss Function	Categorical Crossentropy
Batch Size	32
Epochs	10 - The number of epochs was intentionally kept low to reduce computation time.
Train-Test Split	80% training, 20% validation - Random sampling with a fixed random state for reproducibility.

Table 6. **Hyperparameters Overview**

great capacity to identify cases of moderate dementia, resulting in an F1-score of 0.85. Lower precision and recall levels in the *non_demented* and *mild_dementia* stages, however, suggest that it has difficulty with these stages. The *very_mild_dementia* category has lesser precision but a comparatively superior recall. Macro and weighted averages show that overall performance is constant but still able to be improved across classes, with a focus on identifying moderately demented stages. The precision-recall, ROC, and classification report are displayed in **Appendix B.1.1**.

PET: The trained model, with a test loss of 0.958 and an accuracy of 67.4%, performed well but this accuracy could be improved. Although the model has a tendancy to overclassify images as *non-demented* (high recall of 0.84 but lower precision), it does exceptionally well in recognising mildly demented cases (precision of 0.90 and recall of 0.67). The difficulty of the small sample sizes is reflected in the lack of representation of the *moderate_dementia* category in the model. The balanced statistics show that the performance of individual classes is relatively consistent, although improvement is desirable. The precision-recall, ROC, and classification report are displayed in **Appendix B.1.2**.

CT: When it comes to dividing brain images into four stages of dementia, the pretrained model performs well, with a test loss of 0.378 and an impressive test accuracy of 87.47%. The model's maximum precision (0.97) and F1-score (0.93) are found when classifying *mild_dementia* patients, demonstrating the model's strong efficacy in diagnosing mild dementia. The *moderate_dementia* category had an impressive recall of 0.97, but a lower precision of 0.62, indicating some overclassification but also a good ability to detect most cases with moderate dementia. The macro and weighted average F1-scores, which are 0.85 and 0.88, respectively, together with other measures like MMCE (0.125) and log loss (0.320), demonstrate that the model is well-calibrated and maintains a balanced performance across classes. The precision-recall, ROC, and classification report are displayed in **Appendix B.1.3**.

Modality	Accuracy	Loss	Precision	Recall	F1 Score
MRI	68.52%	0.903	0.72	0.69	0.69
PET	67.4%	0.958	0.73	0.67	0.67
CT	87.5%	0.378	0.88	0.87	0.88
Simple CNN model- Train From Scratch	25.43%	1.386	0.06	0.25	0.10
Our Unet Encoder Model- Train From Scratch	69.73%	1.03	0.70	0.70	0.70
Our Model - Pre-trained	81.4%	0.925	0.81	0.81	0.81

Table 7. **Evaluation Metrics for each modality**

The comparative analysis of the three single-modality basline models models for the classification of dementia stages highlights the potential for major improvements using different approaches, such as targeted data addition, class-weight adjustments, model architecture optimisation, and the incorporation of sophisticated neural architectures. Enhancements like these are meant to increase accuracy, guarantee learning that is fair to all stages of dementia, and improve the model's diagnostic usefulness. Although the models demonstrate some promising capabilities, especially in detecting the non_demented category, the results indicate that focused hyperparameter tuning in conjunction with data improvement methods could significantly improve the models' accuracy. Improving the models' performance—particularly in precisely classifying difficult categories like *moderate_dementia*—is essential to realising their full potential in clinical diagnostic procedures, suggesting a multifaceted strategy that takes into account learning dynamics, model complexity, and data quality to accomplish this.

4.1. Our Proposed Model

4.1.1. Experiment

The main objective was to assess how deep learning transformers and models that integrate clinical data, MRI, CT, and PET improve AD diagnosis. Our aim was to ascertain the joint impact of structural and functional neuroimaging on gaining a more comprehensive and precise understanding of the disease's pathogenesis. We used comparative analysis to test our multimodal fusion approach's added value above baselines (individual models for MRI, CT, PET, and Clinical Assessment) and compare its performance with them.

Using hyperparameters derived from pilot testing and industry best practices mentioned in **Table 4**, the experiments were created with repeatability in mind. Together with statistical analysis, the results are displayed with measures like accuracy, precision, recall, and F1-scores. The in-

terpretation process takes into account the goals of each experiment, analysing its effects, relationship to the conceptual framework, advancements over previous studies, and any additional data insights.

4.2. Baseline Models

4.2.1. Encoder Choice

ResNet50, a 50-layer CNN, tackles the vanishing gradient problem with "residual blocks" or "skip connections" to maintain performance in deeper networks. This feature greatly reduces training times by streamlining training and improving accuracy for computer vision applications. ResNet50's capacity to use pre-trained characteristics and adapt through weight tweaks improves its efficacy for medical image analysis, particularly dementia detection. ResNet50 is a leading option in its field because it combines deep architecture with computational efficiency. It can be trained on traditional hardware and achieves excellent performance on benchmarks (He et al., 2016).

To create a performance standard, the baseline models required testing the neuroimagings separately. We used a modified ResNet50 architecture in these trials, with weights that were not pre-trained and tailored for each imaging modality. The top layer of the ResNet50 model was removed, and a dense layer with four outputs representing different stages of dementia was added, along with a Global Average Pooling layer. This modification uses a softmax function in the dense layer for multi-class classification, using ResNet50's capabilities and avoiding the problem of vanishing gradients.

Table 5 offer an insight into the setup and implementation of these baseline studies, covering the configuring the model, and training protocols. In addition to the ResNet, we also experimented with a UNet, a convolutional neural network architecture, specifically designed for semantic segmentation tasks in computer vision. Its architecture includes a contracting path to capture context and a symmetric expanding path for precise localization. This design enables UNet to achieve superior performance in tasks such as medical image segmentation, where pixel-level accuracy is crucial. It's skip connections enable seamless information flow across abstraction levels, addressing the vanishing gradient problem and enabling efficient training with limited data.

4.3. Hyperparameters

For hyperparameters, the following different specifications were tried: Batch sizes 8,16,32: The primary motivation of a lower beginning batch size was GPU memory issues. However, after implementing various memory saving techniques such as using half precision floating points and gradient check-pointing, the memory footprint was reduced and bigger batch sizes were used, since it was hypothesised that larger batch sizes would provide more stable updates. samples sizes: Sample sizes of 3000 and 5000 were tried to see the effect of increased data on the model performance, in particular the transformer based fusion.

Hyperparameter	Value
Batch Size	32 (Train), 16 (Val)
Learning Rate	0.001
Number of Epochs	40
Training-Validation Split	80% Train, 20% Val
Seed for Random Sampling	42
Test Size for Split	0.2
Target Samples for Classes	10000 (Oversampling)
	500 (Undersampling)

Table 8. **Hyperparameters used in the training process**

Transformer Parameters: Different parameters for the transformer encoder were tried, specifically the input size and the number of layers in the transformer. While initially, both were higher (input layer dimension: 1024x3, hidden layers=2048 and transformer layer=2), decreasing them (to input layer dimension: 256x3, hidden layers=64 and transformer layer=1) ameliorated significant memory issues, and was a more prudent choice given the resource limitations.

4.3.1. Encoder Choice

For the UNet, two experiments were done with a trained vs untrained UNet.

4.3.2. Results

Train from scratch: Using a simple CNN (25.4% accuracy) proved unsuitable for feature extraction before transformers. U-Net (69.4% accuracy) excelled, suggesting spatial information is key for dementia classification.

Pretrained: The pretrained model shows good performance across the four phases of dementia. It has an overall accuracy of 81%. With an outstanding recall of 0.98 and a high precision of 0.85 in recognising *moderate_demented* cases, the model performs exceptionally well, demonstrating its ability to accurately identify almost all cases of this class with a low number of false positives. The *mild_demented* category shows excellent precision as well, with a recall of 0.93, yet there is room for improvement in terms of catching all true positives. The *non_demented* and *very_mild_demented* classes have somewhat lower metrics, but they nevertheless have excellent precision and recall, which adds to the model's overall balanced performance as demonstrated by weighted averages of around 0.81, uniform accuracy, and macro. This consistency demonstrates how well the model can classify various stages of dementia, doing especially well in more severe cases.

4.4. Basline vs Proposed Model

Comparing performance measurements between various modalities and model architectures yields distinct patterns in efficacy and efficiency. With an accuracy of 87.5% and the best precision, recall, and F1 scores, CT scans perform the best. This highlights the suitability of CT images for the task at hand, probably because they contain precise structural information. However, compared to CT, MRI and PET modalities display similar, if lower, performance metrics,

due to their complex structure which requires utilizing all slices to capture the intricate details for diagnosis. The basic CNN model that was trained from scratch performs much worse across the board, with a low precision and F1 score and an accuracy of only 25.43%. This indicates that more sophisticated architectures are required for these kinds of image processing tasks. Even when trained from scratch, the U-Net Encoder Model provides a significant enhancement over the basic CNN, suggesting that the U-Net design can more effectively utilise the spatial hierarchies found in medical pictures. Lastly, the pretrained version of U-Net design demonstrates the value of transfer learning by attaining the best performance among models trained from scratch (81.4% accuracy), which, although still less than the CT modality, shows how pretraining has a significant positive influence on improving model performance in all domains. This comparison highlights how crucial it is to select the best model architecture, training plan, and comprehension of modality-specific features in order to maximise performance for medical image processing tasks.

5. Related work

Advances in medical imaging technologies in recent times have made it possible to obtain multimodal images that offer supplementary information about the human anatomy. (Kharfi, 2013) The integration of these multimodal images is essential for improving diagnostic precision and offering a thorough perspective for medical examination. Various methods have been put forth in this regard to deal with the difficulties that come with multimodal medical picture fusion.

The limitations of old approaches were addressed by Li et al. (Li et al., 2021), who presented a deep learning strategy for MRI, CT, and SPECT image fusion, improving time efficiency, detail clarity, and fusion efficacy. Similar to this, Zhang et al. (Zhang, 2023) created a Transformer-Based Conditional GAN for multimodal picture fusion, emphasising knowledge integration and long-distance domain interdependence while concentrating on deep learning for feature extraction. Their approach achieves higher performance in maintaining structural and textural elements, leading to an improvement in global picture perception and convergence.

Furthermore, a multi-objective differential evolution-based deep neural network for multimodal picture fusion was investigated by Kaur and Singh in 2021 (Kaur & Singh, 2021). Using the Xception model for feature extraction and picture decomposition, their approach makes use of the non-subsampled contourlet transform. This method performs better than traditional approaches because it can pick optimal features using multi-objective differential evolution.

However, novel algorithms have been developed for complicated and intelligent systems in the areas of computing efficiency and information retention. Because it uses a method of calculating local mutual information correlation, a study by Peng Guo et al. (Guo et al., 2022) presented a method that is slower than some fusion techniques. This study emphasises the significance of effective threshold determination schemes that take the prior information of source images into account.

Additionally, research published by Wang et al. (Wang et al., 2020) presented a technique for processing multimodal medical images of any size that combines weight map generation and pyramid decomposition. This method emphasises how important it is to convert fully connected layers to convolutional layers in order to create dense prediction maps, which in turn makes it easier to fuse images of any size.

These studies show how multimodal medical picture fusion is moving towards deep learning, emphasising domain expertise and long-distance data linkages through the use of cutting-edge methods like neural networks and GANs. This method improves the diagnostic value and quality of medical images. In particular, the accuracy of diagnosing AD and MCI is increased by combining structural and functional neuroimaging . Our research shows that deep learning models—particularly fusion networks—perform very well at deciphering intricate multimodal data and identifying pertinent patterns in order to create complete models for the diagnosis of neurodegenerative diseases.

6. Conclusions

In this work, we present a novel architecture for the multimodal classification of alzheimers progression. Our model uses MRI, CT and PET scans, encodes them using the downsampling path of a UNet, and attention fuses the resulting features together before making a classification. Analysis of results reveals that the Encoder design is critical, and that UNet outperforms a CNN with attention. While the accuracy of our model does not exceed the accuracy of the baseline uni-modal CNNs, this might be due to an insufficient amount of data, as transformers require more training data due to increased number of parameters. Although due to the limited scope and resources of this project, the open source OASIS dataset was used, the performance of the model could be further increased by access to more datasources like ADNI. In future, our approach should be evaluated on a larger data source to examine it more fairly.

References

AlMansoori, ME, Jemimah, S, Abuhantash, F, and AlShehhi, A. Predicting early alzheimer's with blood biomarkers and clinical features. *Sci Rep*, 14(1):6039, Mar 2024. doi: 10.1038/s41598-024-56489-1.

Aramadaka, S, Mannam, R, Sankara Narayanan, R, Bansal, A, Yanamaladoddi, VR, Sarvepalli, SS, and Vemula, SL. Neuroimaging in alzheimer's disease for early diagnosis: A comprehensive review. *Cureus*, 15(5):e38544, May 2023. doi: 10.7759/cureus.38544.

Bharathi, A and Arunachalam, AS. Pre-processing on alzheimer mri images. *Annals of the Romanian Society for Cell Biology*, pp. 4433–4441, 2021.

Buda, Mateusz, Saha, Ashirbani, and Mazurowski, Maciej A. Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. *Computers in biology and medicine*, 109:218–225, 2019.

Fang, Guian, Liu, Mengsha, Zhong, Yi, Zhang, Zhuolin, Huang, Jiehui, Tang, Zhenchao, and Chen, Calvin Yu-Chian. Multimodal identification of alzheimer's disease: A review. *arXiv preprint arXiv:2311.12842*, 2023.

Gharaibeh, Maha, Elhies, Mwaffaq, Almahmoud, Mothanna, Abualigah, Sayel, and Elayan, Omar. Machine learning for alzheimer's disease detection based on neuroimaging techniques: A review. In *2022 13th International Conference on Information and Communication Systems (ICICS)*, pp. 426–431, 2022. doi: 10.1109/ICICS55353.2022.9811143.

Guo, Peng, Xie, Guoqi, Li, Renfa, and Hu, Hui. Multimodal medical image fusion with convolution sparse representation and mutual information correlation in nsst domain. *Complex & Intelligent Systems*, 9:317–328, 2022. URL https://api.semanticscholar.org/CorpusID:250188442.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kalamkar, Shrida and A., Geetha Mary. Multimodal image fusion: A systematic review. *Decision Analytics Journal*, 9:100327, 2023. ISSN 2772-6622. doi: https://doi.org/10.1016/j.dajour.2023.100327. URL https://www.sciencedirect.com/science/article/pii/S2772662223001674.

Kaur, Manjit and Singh, Dilbag. Multi-modality medical image fusion technique using multi-objective differential evolution based deep neural networks. *Journal of Ambient Intelligence and Humanized Computing*, 12(2): 2483–2493, 2021.

Kharfi, F. *Principles and Applications of Nuclear Medical Imaging: A Survey on Recent Developments*. InTech, Mar. 13 2013. doi: 10.5772/54884.

Kim, J, Jeong, M, Stiles, WR, and Choi, HS. Neuroimaging modalities in alzheimer's disease: Diagnosis and clinical features. *International Journal of Molecular Sciences*, 23(11):6079, May 2022. doi: 10.3390/ijms23116079.

Koenig, Lauren N., Day, Gregory S., Salter, Amber, Keefe, Sarah, Marple, Laura M., Long, Justin, LaMontagne, Pamela, Massoumzadeh, Parinaz, Snider, B. Joy, Kanthamneni, Manasa, Raji, Cyrus A., Ghoshal, Nupur, Gordon, Brian A., Miller-Thomas, Michelle, Morris, John C., Shimony, Joshua S., and Benzinger, Tammie L.S. Select atrophied regions in alzheimer disease (sara): An improved volumetric model for identifying alzheimer disease dementia. *NeuroImage: Clinical*, 26:102248, 2020. ISSN 2213-1582. doi: https://doi.org/10.1016/j.nicl.2020.102248. URL https://www.sciencedirect.com/science/article/pii/S2213158220300851.

Lahat, Dana, Adali, Tülay, and Jutten, Christian. Multimodal data fusion: an overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9):1449–1477, 2015.

LaMontagne, Pamela J., Benzinger, Tammie LS., Morris, John C., Keefe, Sarah, Hornbeck, Russ, Xiong, Chengjie, Grant, Elizabeth, Hassenstab, Jason, Moulder, Krista, Vlassenko, Andrei G., Raichle, Marcus E., Cruchaga, Carlos, and Marcus, Daniel. Oasis-3: Longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. *medRxiv*, 2019. doi: 10.1101/2019.12.13.19014902. URL https://www.medrxiv.org/content/early/2019/12/15/2019.12.13.19014902.

Li, Yi, Zhao, Junli, Lv, Zhihan, and Li, Jinhua. Medical image fusion method by deep learning. *International Journal of Cognitive Computing in Engineering*, 2:21–29, 2021. ISSN 2666-3074. doi: https://doi.org/10.1016/j.ijcce.2020.12.004. URL https://www.sciencedirect.com/science/article/pii/S2666307420300280.

Lin, Tianyang, Wang, Yuxin, Liu, Xiangyang, and Qiu, Xipeng. A survey of transformers. *AI Open*, 3:111–132, 2022. ISSN 2666-6510. doi: https://doi.org/10.1016/j.aiopen.2022.10.001. URL https://www.sciencedirect.com/science/article/pii/S2666651022000146.

Marcus, Daniel S., Wang, Tracy H., Parker, Jamie, Csernansky, John G., Morris, John C., and Buckner, Randy L. Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults. *Journal of Cognitive Neuroscience*, 19(9):1498–1507, 09 2007. ISSN 0898-929X. doi: 10.1162/jocn.2007.19.9.1498. URL https://doi.org/10.1162/jocn.2007.19.9.1498.

Marcus, Daniel S., Fotenos, Anthony F., Csernansky, John G., Morris, John C., and Buckner, Randy L. Open Access Series of Imaging Studies: Longitudinal MRI Data in Nondemented and Demented Older Adults. *Journal of Cognitive Neuroscience*, 22(12):2677–2684, 12 2010. ISSN 0898-929X. doi: 10.1162/jocn.2009.21407. URL https://doi.org/10.1162/jocn.2009.21407.

Palko, Kyle. Diagnosing autism spectrum disorder through brain functional magnetic resonance imaging. 2016. URL https://api.semanticscholar.org/CorpusID:151477197.

Prakash, Aditya, Chitta, Kashyap, and Geiger, Andreas. Multi-modal fusion transformer for end-to-end autonomous driving. *CoRR*, abs/2104.09224, 2021. URL https://arxiv.org/abs/2104.09224.

Ramazanova, Merey, Escorcia, Victor, Heilbron, Fabian, Zhao, Chen, and Ghanem, Bernard. Owl (observe, watch, listen): Localizing actions in egocentric video via audio-visual temporal context, 02 2022.

Ronneberger, Olaf, Fischer, Philipp, and Brox, Thomas. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241. Springer, 2015.

Sharma, Shallu and Mandal, Pravat. A comprehensive report on machine learning-based early detection of alzheimer's disease using multi-modal neuroimaging data. *ACM Computing Surveys*, 55:1–45, 02 2022. doi: 10.1145/3492865.

Wang, Kunpeng, Zheng, Mingyao, Wei, Hongyan, Qi, Guanqiu, and Li, Yuanyuan. Multi-modality medical image fusion using convolutional neural network and contrast pyramid. *Sensors (Basel, Switzerland)*, 20, 2020. URL https://api.semanticscholar.org/CorpusID:215772911.

WHO, World Health Organization:. Dementia, March 2023. URL https://www.who.int/news-room/fact-sheets/detail/dementia.

Ye, Haizhou, Zhu, Qi, Yao, Yuan, Jin, Yichao, and Zhang, Daoqiang. Pairwise feature-based generative adversarial network for incomplete multi-modal alzheimer's disease diagnosis. *The Visual Computer*, 39(6):2235–2244, 2023.

Zhang, [Author Names Redacted for Privacy]. Transformer based conditional gan for multimodal image fusion. *[Journal Information Redacted for Privacy]*, 2023. Manuscript in preparation.

Zhang, Yu-Dong, Dong, Zhengchao, Wang, Shui-Hua, Yu, Xiang, Yao, Xujing, Zhou, Qinghua, Hu, Hua, Li, Min, Jiménez-Mesa, Carmen, Ramirez, Javier, et al. Advances in multimodal data fusion in neuroimaging: Overview, challenges, and novel orientation. *Information Fusion*, 64:149–187, 2020.

A. Methodology

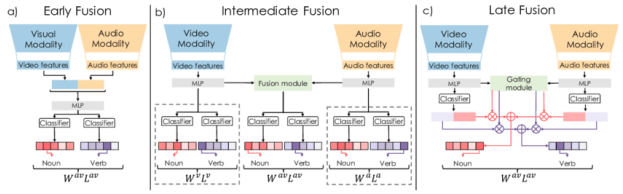

Figure 3. **Multimodal Data Fusion (Ramazanova et al., 2022)**

A.1. Our Proposed Model

Algorithm 2 Training a Multimodal Transformer Model with Multiple Transformer Layer

Input: Multimodal data x_i, labels y_i, batch size b, learning rate lr, epochs E
Initialize Encoders $Encoder_{MRI}$, $Encoder_{CT}$, $Encoder_{PET}$ with respective input channels
Initialize TransformerEncoder $T_{encoder}$ and MLP classifier C
Output: Trained Multimodal Transformer Model
for epoch = 1 **to** E **do**
  Initialize epoch loss and accuracy: $L_{epoch} = 0$, $A_{epoch} = 0$
  for all batches (x_{batch}, y_{batch}) in data **do**
    Split x_{batch} into (MRI, CT, PET) based on modality
    Encode each modality: $e_{MRI} = Encoder_{MRI}(MRI)$, $e_{CT} = Encoder_{CT}(CT)$, $e_{PET} = Encoder_{PET}(PET)$
    Combine encoded features: $e_{combined} = [e_{MRI}, e_{CT}, e_{PET}]$
    Transform combined features: $t_{features} = T_{encoder}(e_{combined})$
    Obtain logits: $logits = C(t_{features})$
    Compute loss: $L = \text{CrossEntropy}(logits, y_{batch})$
    Update model parameters to minimize L
    Accumulate L_{epoch} and compute accuracy A_{epoch}
  end for
  Print epoch summary with L_{epoch} and A_{epoch}
end for

B. Experiments

B.1. Baseline Models

This section in the baseline models' classification report, ROC curve and Precision-Recall curve.

B.1.1. MRI

B.1.2. PET

B.1.3. CT

	PRECISION	RECALL	F1-SCORE	SUPPORT
NON_DEMENTED	0.63	0.62	0.63	4984
VERY_MILD_DEMENTIA	0.55	0.75	0.64	4990
MILD_DEMENTIA	0.89	0.50	0.64	4980
MODERATE_DEMENTIA	0.81	0.89	0.85	4339
ACCURACY			0.69	19293
MACRO AVG	0.72	0.69	0.69	19293
WEIGHTED AVG	0.72	0.69	0.68	19293

Table 9. Classification Report for MRI Baseline Model

Table 10. ROC curve for all the classes of MRI Modalities

	PRECISION	RECALL	F1-SCORE	SUPPORT
NON_DEMENTED	0.54	0.84	0.66	5000
VERY_MILD_DEMENTIA	0.79	0.51	0.62	5000
MILD_DEMENTIA	0.90	0.67	0.77	5000
MODERATE_DEMENTIA	0.52	0.68	0.59	853
ACCURACY			0.67	15853
MACRO AVG	0.69	0.68	0.66	15853
WEIGHTED AVG	0.73	0.67	0.68	15853

Table 11. Classification Report for PET Baseline Model

	PRECISION	RECALL	F1-SCORE	SUPPORT
NON_DEMENTED	0.81	0.86	0.83	5000
VERY_MILD_DEMENTIA	0.90	0.86	0.88	5000
MILD_DEMENTIA	0.97	0.89	0.93	5000
MODERATE_DEMENTIA	0.62	0.97	0.75	504
ACCURACY			0.87	15504
MACRO AVG	0.82	0.90	0.85	15504
WEIGHTED AVG	0.88	0.87	0.88	15504

Table 12. Classification Report for CT Baseline Model