

Analyzing Deep Learning Approaches in Unimodal and Multimodal Frameworks for Alzheimer's Disease Stage Classification

¹Sraddha Sigdel, ¹Stuti Khadka, ²Sujal Prasad Singh, ³Dhiraj Pyakurel, ³Amit Kumar Rauniyar

^{1,2,3}*Electronics and Computer Department, Advanced College of Engineering and Management, Kathmandu, Nepal*

Email: dhiraj@acem.edu.np, amitkumar.rauniyar@acem.edu.np

Abstract

In this paper, we present a detailed study on deep learning approaches for the binary classification of Alzheimer's disease stages using both unimodal and multimodal frameworks. Alzheimer's disease is a progressive neurodegenerative brain disorder and the leading cause of dementia, characterized by cognitive decline, memory loss, and behavioral changes due to amyloid plaques and tau tangles. It typically develops in stages, from mild memory lapses to severe cognitive impairment, and while incurable, treatments can temporarily manage symptoms. We effectively utilised PET and MRI scans, as well as CSF biomarkers, demographics and genetic data, widely used in medical settings, for developing both unimodal and multimodal frameworks that can effectively classify MCI and AD stages of Alzheimer's disease. The traditional diagnostic methods widely use only single modularity for the diagnostic purpose however multimodal effectively utilises all three modality of the data. The unimodal models are implemented using a 3D ResNet-18 architecture, while the multimodal system incorporates modality-specific encoders, cross-modal attention mechanisms, and a gated fusion strategy to combine heterogeneous modality/data sources. We observed that for unimodal setup PET scans have higher AUC of 0.915 than that of MRI 0.84 and performs better for stage classification, It is because PET scans allow for the evaluation of brain metabolism (FDG-PET) and specific neuropathology, which helps distinguish Alzheimer's from other dementias, enabling earlier diagnosis and treatment. MRI scans is also a widely used diagnostic method for Alzheimer's disease which particularly detects brain shrinkage (atrophy), particularly in the hippocampus region. The multimodal architecture combines different modality which complements and compensates for missing information and provides a comprehensive view of disease pathology, and performs the best among all the models with the AUC of 0.940. Integrating multiple modalities can also assist us with differential diagnosis, such as distinguishing Alzheimer's from other forms of dementia or co-existing conditions. Thus, we implemented three different kinds of models, utilising different modalities to provide a comprehensive study for the binary classification of stages of Alzheimer's.

Keywords: *Alzheimer's disease, Cognitive decline, Multimodal, Modalities, Unimodal*

I. INTRODUCTION

Alzheimer's disease is the most common form of dementia and accounts for nearly 60–70% of dementia cases worldwide [1]. More than 55 million people were living with dementia in 2020, and this number is expected to almost double every 20 years, reaching 78 million in 2030 and 139 million by 2050 [2]. In Nepal, although the exact prevalence remains uncertain, estimates suggest that around 78,000 to 135,000 people are affected, with the number also expected to double every two decades [3]. A study conducted among elderly patients (N=210) attending the psychiatry OPD at a tertiary care hospital reported a dementia prevalence of 11.4%, among which Alzheimer's disease accounted for 70.8% of cases, followed by vascular dementia at 25% [4].

Currently, around 60% of people with dementia live in low- and middle-income countries, and this proportion is expected to rise to 71% by 2050. Dementia is a neurocognitive disorder characterized by a decline in cognitive abilities such as memory, reasoning, and thinking, severe enough to affect daily functioning. Alzheimer's disease, the leading cause of dementia, is associated with the accumulation of beta-amyloid plaques outside neurons and neurofibrillary tau tangles inside neurons. Over time, these abnormal protein deposits lead to neuronal death, brain tissue damage, and progressive brain atrophy, which becomes more noticeable in later stages of the disease.

Although Alzheimer's disease currently has no definitive cure, its progression and severity can be assessed using MRI and PET imaging techniques. In addition, cerebrospinal fluid (CSF) biomarkers such as A β 1-42, T-tau, and P-tau, along with clinical assessment scores, can support diagnosis and severity estimation. With advances in artificial intelligence and deep learning, medical imaging data and clinical biomarkers can now be analyzed using unimodal or multimodal approaches for the early detection and classification of Alzheimer's disease [5].

Alzheimer's disease is progressive, and its early stages may remain unnoticed for years. The Alzheimer's Disease Neuroimaging Initiative (ADNI) identifies three major stages of progression: preclinical Alzheimer's disease, Mild Cognitive Impairment (MCI) due to Alzheimer's, and dementia due to Alzheimer's disease. While preclinical Alzheimer's often remains undiagnosed, the later stages are associated with increasing cognitive decline and worsening dementia symptoms. Deep learning models can effectively analyze medical imaging data, clinical records, and biomarkers to assist in accurate disease stage prediction [5]. Multimodal approaches combine different types of data to provide a more comprehensive understanding of disease progression and can improve the reliability of Alzheimer's stage classification.

Recent progress in generative AI and deep learning has expanded applications in health-care, mental well-being, and medical image analysis has created new opportunities for improving Alzheimer's diagnosis, particularly in resource-constrained regions such as Nepal [6, 7, 8]. While multimodal methods

utilize the strengths of multiple data sources, unimodal approaches can still provide strong diagnostic performance when only a single modality is available. Models trained on large datasets can support clinical decision-making and help compensate for limited diagnostic expertise. This project explores both unimodal and multimodal deep learning approaches, with a particular focus on gated multimodal fusion, to evaluate their effectiveness in Alzheimer's disease stage classification. For Nepal, where diagnostic infrastructure and specialist availability may be limited, AI-assisted tools can provide scalable and practical support for clinical settings.

II. LITERATURE REVIEW

Recent research on Alzheimer's disease (AD) diagnosis has progressively moved from traditional machine learning methods to deep learning and multimodal frameworks, leading to significant improvements in performance and feature learning. Early studies using approaches such as PCA, LDA, SVM, and other classical classifiers generally achieved limited performance, particularly in multi-class AD stage classification, due to their inability to capture complex neuroimaging patterns effectively [9].

Recent work has also demonstrated that uncertainty-aware deep learning models integrated with expectation maximization techniques can further improve the robustness of brain tissue analysis and medical image segmentation, providing reliable feature representations for downstream neurodegenerative disease analysis [5]. Deep learning-based methods, especially CNNs and their variants, have shown much stronger results by automatically learning hierarchical representations from MRI data. Both 2D and 3D CNN architectures, along with transfer learning models such as VGG16, ResNet, and Inception-based networks, have been widely used for AD classification and generally outperform traditional approaches in binary classification tasks [10, 11, 12, 13]. However, 2D slice-based methods are often criticized for potential data leakage and loss of spatial consistency, while 3D CNNs, although more robust, require higher computational resources.

To overcome the limitations of unimodal learning, multimodal approaches integrate multiple data sources such as MRI, PET, CSF biomarkers, and clinical/EHR data. These approaches consistently demonstrate improved diagnostic accuracy compared to single-modality models by capturing complementary disease-related information [14, 15]. Advanced fusion strategies such as kernel-based fusion and nonlinear graph fusion further improve performance, achieving strong results in both binary and multi-class classification settings [16]. Despite these gains, many fusion strategies still rely on simple concatenation or linear combinations, which limits their ability to model complex inter-modality relationships.

More recent deep learning frameworks attempt to address this through hybrid architectures and feature-level integration, but challenges remain in achieving stable

performance for fine-grained AD stage classification, where accuracy is still significantly lower compared to binary tasks [9, 15, 13].

Overall, existing methods demonstrate that multimodal learning is more effective than unimodal approaches, but most current fusion strategies fail to dynamically adjust the contribution of each modality. To address this gap, our work proposes a ****gated multimodal fusion framework****, which learns adaptive gating weights to selectively emphasize the most informative modalities for each subject. This allows more robust feature integration and improved discrimination across Alzheimer's disease stages.

III. METHODOLOGY

The data was acquired from ADNI database [17] and NACC database [18]. ADNI data repository contains imaging, clinical, and genetic data for patients that span four studies (ADNI1, ADNI2, ADNI GO, and ADNI3). The imaging data was taken from two of the studies ADNI1, 2 and GO. Similarly, the NACC database also contains the MRI, PET as well as Clinical score and the CSF biomarker. The NACC database is organized into several sub-datasets: Uniform Data Set, Minimum Data Set, Neuropathology (NP) Data Set, Brief Data Set. The Uniform dataset consists of there are primary clinical datasets, collected longitudinally since 2005. The Neuropathology Data Set consists of autopsy data on UDS/MDS participants. The Brief Data Set consists of a brief clinical capture form initiated around 2021 for neuropathology submissions.

A multimodal dataset was constructed by identifying subjects from both the ADNI and NACC repositories who had complete data across all four modalities MRI, PET, CSF biomarkers, and clinical scores recorded at the same study visit. This alignment ensured that each subject's modality data corresponded to a single, consistent diagnosis of either MCI or AD. Subjects who had progressed from MCI to AD were retained in the dataset, with their AD-stage visit data included. The statistics of resulting dataset is summarized in Table 1.

Table 1: Number of Subjects per Class by Data Source

Source	MCI	AD	Total
ADNI	199	241	440
NACC	351	309	660
Total	550	550	1100

Multimodal data analysis

MRI scan

MRI is a medical imaging technique used in radiology to generate pictures of the anatomy and the physiological processes inside the body. MRI scanners use strong magnetic fields, magnetic field gradients, and radio waves to form images of the organs in the body. It is used in hospitals and clinics for medical diagnosis, staging and follow-up of disease. The MRI preprocessing pipeline consisted of seven steps to

transform raw scanner output into clean, standardized brain image. Raw MRI scans acquired directly from the scanner contained noise, intensity inconsistencies, and non-brain tissues, making them unsuitable for direct analysis. To address positional variability, each scan was first reoriented to the LPS (Left, Posterior, Superior) coordinate system, followed by rigid registration to the MNI152 standard brain template (MNI152 T1 1mm.nii) using ANTs, which aligned all scans to a common spatial reference frame through rotation and translation transformations while preserving brain morphology. Skull stripping was performed using FSL's Brain Extraction Tool (BET) with a fractional intensity threshold of 0.45 to remove the skull, scalp, and other non-brain tissues, leaving only the brain region relevant to AD-specific changes such as the hippocampus. This threshold was chosen by testing several values: lower thresholds left behind bits of skull or scalp tissue, while higher thresholds removed parts of the actual brain. A threshold of 0.45 gave the best balance between these two problems. Subsequently, bias field correction was applied using FSL's FAST tool to eliminate smooth intensity gradients caused by magnetic field inhomogeneities, ensuring uniform intensity distribution across the brain volume. Finally, min-max normalization was applied to rescale all non-zero brain voxel intensities to a standardized range of 0 to 1, excluding background voxels to ensure that scans acquired across different patients and scanner settings were treated on an equal scale during model training.

PET Scan

A positron emission tomography (PET) scan is an imaging test that produces images of your organs and tissues at work. While other imaging studies like CT and MRI produce still images of organs and body structures. PET scans use a radioactive tracer to show how an organ is functioning in real time. PET scan images can detect cellular changes in organs and tissues earlier than CT and MRI scans.

The PET preprocessing pipeline consisted of seven steps to convert raw scanner images into clean, standardized brain scans suitable for machine learning. Raw PET scans acquired directly from the scanner were noisy, low in resolution, and lacked anatomical detail, making them unsuitable for direct analysis. The scans were first converted from DICOM format, the standard scanner output format, into NIfTI format (.nii or .nii.gz), which is the standard format required by all neuroimaging tools and analysis pipelines. Since PET images lack clear anatomical boundaries on their own brain mask derived from the corresponding, a MRI scan was then applied to remove non-brain tissues such as the skull and scalp. The brain-masked PET image was then co-registered to the patient's own T1-weighted MRI scan, mapping the metabolic activity captured by PET onto the anatomical structures visible in the MRI, ensuring that key AD-affected regions such as the hippocampus and entorhinal cortex were accurately identified. The co-registered PET image was then aligned to the MNI152 standard brain template using rigid registration, which repositioned the brain through

rotation and translation without changing its shape or size, ensuring that the same brain regions were in the same spatial location across all patient scans. Finally, SUVR normalization was applied by dividing each voxel's intensity by the average tracer uptake in a reference region unaffected by AD, such as the cerebellum, correcting for patient-to-patient variability in tracer uptake and producing intensity values that could be fairly compared across all patients and scanning sessions.

Clinical Score

CSF (Cerebro-Spinal Fluid) Biomarker is a measurable indicator within the body that can help diagnose, monitor, or predict the disease's progression or response to treatment. CSF is a clear, colorless transcellular body fluid found within the meningeal tissue that surrounds the vertebrate, brain and spinal cord, and in the ventricles of the brain. The tabular data preprocessing pipeline consisted of three steps to clean and prepare the dataset for machine learning. First, missing values were filled in by identifying other patients with similar characteristics such as APOE4 status, sex, age, and class, so that the imputed values were clinically realistic rather than simply replacing missing entries with a global average, and any values falling outside a valid range were clamped to the minimum or maximum threshold to remove extreme entries. Next, categorical columns such as gender and class, which store information as text labels rather than numbers, were converted into integers using label encoding, since machine learning models cannot process categorical values directly. Finally, normalization was applied to bring all features onto a comparable scale by calculating the mean and standard deviation of each feature, ensuring that the model did not unfairly favor features with larger numerical ranges during training.

Problem Formulation

Let the input data be derived from multiple modalities, including magnetic resonance imaging (MRI), positron emission tomography (PET), and cerebrospinal fluid (CSF) biomarkers. In the unimodal setting, the input is defined as: X^{MRI} , X^{PET} . In the multimodal setting, the input is represented as a combination of all available modalities: $X = \{X^{MRI}, X^{PET}, X^{tab}\}$. The objective is to learn a mapping function $f: X \rightarrow y$ where the output label is defined as $y \in \{MCI, AD\}$.

This formulation corresponds to a binary classification problem, where the classes represent mild cognitive impairment (MCI), and Alzheimer's disease (AD). To focus learning on harder misclassified samples, the model is optimized by minimizing focal loss [19], defined as:

$$L_{FL} = -\frac{1}{N} \sum_{i=1}^N \alpha_t (1 - \hat{y}_i)^{\gamma} \left[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right]$$

where,

- N is the number of samples, $y_i \in \{0, 1\}$ is the ground truth label,

- $\hat{y}_i$ is the predicted probability, $\alpha_t \in [0, 1]$ is a weighting factor that balances the contribution of positive and negative classes, and
- $\gamma \geq 0$ is the focusing parameter that reduces the loss contribution of easy examples.

When $\gamma = 0$, focal loss reduces to standard binary cross-entropy.

The optimization was performed using the AdamW[20] optimizer. It is a modification of the Adam[21] optimizer that improves generalization by decoupling weight decay from the gradient-based update. Instead of modifying the gradient, weight decay is applied directly to the parameters during the update step. It was introduced to address a key limitation of Adam, where weight decay is incorrectly implemented as L2 regularization. This results in two separate operations; gradient-based update using Adam, and explicit weight decay applied independently. The parameter update in AdamW is given by

$$\theta_{t+1} = \theta_t - \alpha \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} - \alpha \lambda \theta_t \quad (2)$$

where, λ is the weight decay coefficient, and the last term represents decoupled weight decay. The performance of the resulting models are evaluated using AUC (Area under the Receiver Operating Characteristic Curve). This metric summarizes the model's performance across all classification thresholds. AUC can be interpreted as the probability that the model ranks a randomly chosen positive instance higher than a randomly chosen negative instance. Here, AUC=1 is a perfect classifier, AUC=0.5 is random guessing and AUC < 0.5 is poor performance (worse than random).

Unimodal Learning Framework To evaluate the individual contribution of each modality, separate unimodal architectures were developed for MRI and PET data, with each model trained independently. A unimodal architecture is formally defined as:

$$\mathcal{W} = \{f_e, f_c\} \quad (3)$$

where,

- $f_e: X \rightarrow Z$ is a modality-specific encoder that maps the input volume $x \in X$ to a latent representation $z \in Z$, and
- $f_c: Z \rightarrow Y$ is a classifier that maps the learned representation to a predicted class label $\hat{y} \in Y$.

The predicted output is thus given by:

$$\hat{y} = f_c(f_e(\mathbf{x})) \quad (4)$$

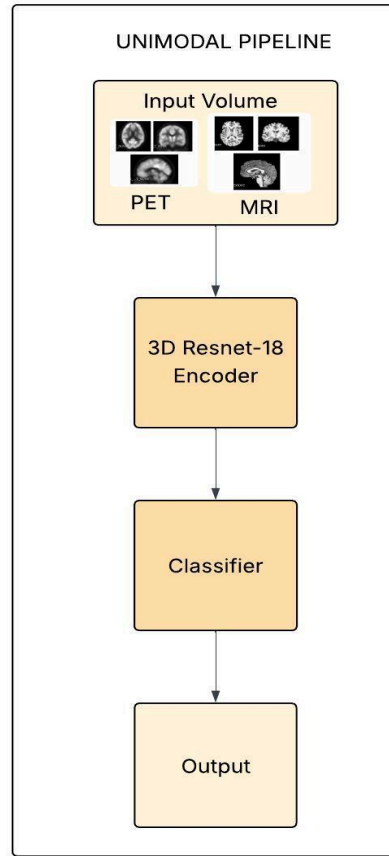


Fig. 1: Unimodal System Architecture

Architectural Details:

Since both MRI and PET are three-dimensional volumetric inputs, the same encoder architecture f_e is shared across both unimodal models — specifically, a 3D variant of ResNet-18 implemented using the MONAI framework [22] — while the weights are learned independently for each modality. The original fully connected classification head was replaced by a custom classifier f_c , that consists of a dropout layer followed by a fully connected layer. Formally:

$$f_c(\mathbf{z}) = \mathbf{W} \text{Dropout}(\mathbf{z}; p = 0.5) + \mathbf{b} \quad (5)$$

where,

- $\mathbf{z} \in R^{512}$ is the latent representation produced by f_e ,
- $\text{Dropout}(\cdot; p = 0.5)$ randomly zeroes elements with probability 0.5 during training to mitigate overfitting, and
- $\mathbf{W} \in R^{2 \times 512}$, $\mathbf{b} \in R^2$ are the learnable weight matrix and bias of the linear projection onto the two output classes (MCI and AD).

The entire unimodal learning framework is described in Algorithm 1.

Algorithm 1 Unimodal Alzheimer's Classification Pipeline (MRI / PET)

-
- 1: **Input:** Volume $\mathbf{x}_m \in \mathbb{R}^{1 \times D \times H \times W}$, where $m \in \{\text{MRI}, \text{PET}\}$
 - 2: **Output:** Predicted class label $\hat{y} \in \{\text{MCI}, \text{AD}\}$
 - 3: **Step 1: Feature Extraction via 3D ResNet-18**
 - 4: $\mathbf{z}_m \leftarrow f_e^m(\mathbf{x}_m)$ ($\in \mathbb{R}^{512}$)
 - 5: **Step 2: Regularization**
 - 6: $\mathbf{z}_{\text{drop}} \leftarrow \text{Dropout}(\mathbf{z}_m, p = 0.5)$ ($\in \mathbb{R}^{512}$)
 - 7: **Step 3: Classification**
 - 8: $\hat{y} \leftarrow \arg \max(\mathbf{W}_c \mathbf{z}_{\text{drop}} + \mathbf{b}_c)$
 - 9: **return** $\hat{y}$
-

Encoder Architecture: 3D ResNet-18

Residual Networks (ResNet), introduced by He et al. (2015)[23], were designed to solve the vanishing gradient and degradation problems in very deep convolutional neural

networks (CNNs). It introduces the concept of skip connections that enables the training

of very deep models (e.g., 152+ layers) without performance degradation. In residual learning, instead of approximating $H(x)$ directly, ResNet learns

$$H(x) = F(x) + x \quad (6)$$

where,

- x is the input to a block, $F(x)$ is the residual function (learned by stacked layers), and
- $H(x)$ is the desired output.

3D Convolution: Standard convolutional neural networks operate on two-dimensional inputs, applying filters across the height and width of an image. However, MRI and PET scans are inherently volumetric, consisting of a stack of slices that together encode three-dimensional spatial structure. Processing these volumes slice-by-slice with 2D convolutions discards inter-slice spatial relationships, which are critical for capturing anatomical context such as hippocampal atrophy or regional hypometabolism. To address this, 3D convolutions extend the standard 2D operation by applying filters across three spatial dimensions — height H , width W , and depth D . Formally, for an input volume $x \in \mathbb{R}^{C \times D \times H \times W}$ with C input channels, the 3D convolution operation is defined as:

$$\mathbf{z}(d, h, w) = \sum_{c=1}^C \sum_{k_d} \sum_{k_h} \sum_{k_w} \mathbf{w}(c, k_d, k_h, k_w) \cdot \mathbf{x}(c, d + k_d, h + k_h, w + k_w) \quad (7)$$

where,

- w is the learnable filter kernel, and

- (k_d, k_h, k_w) index over the kernel dimensions.

This allows the model to capture local volumetric patterns jointly across all three spatial axes.

The ResNet framework was extended to three-dimensional convolutional networks by Tran et al. (Tran et al. 2017), originally for spatiotemporal video modeling. In the context of volumetric medical imaging, this architecture is repurposed to operate over three spatial dimensions rather than two spatial and one temporal dimension. The 3D ResNet-18 encoder f_e processes inputs of the form:

$$\mathbf{x} \in \mathbb{R}^{C \times D \times H \times W} \quad (8)$$

where,

- C is the number of input channels, and
- D, H, W denote the depth, height, and width of the volumetric scan respectively.

Each residual block within the network consists of two 3D convolutional layers (Conv3D), 3D batch normalization, ReLU activations, and a skip connection that adds the block's input directly to its output. These skip connections allow gradients to flow unimpeded through the network, mitigating the vanishing gradient problem and enabling stable training of deeper architectures. The overall architecture of ResNet-3D consists of; initial 3D convolution and 3D max pooling layer, multiple stages of stacked 3D residual blocks, increasing feature depth while reducing spatial and temporal dimensions, global average pooling and fully connected layer for classification.

Multimodal Learning Framework

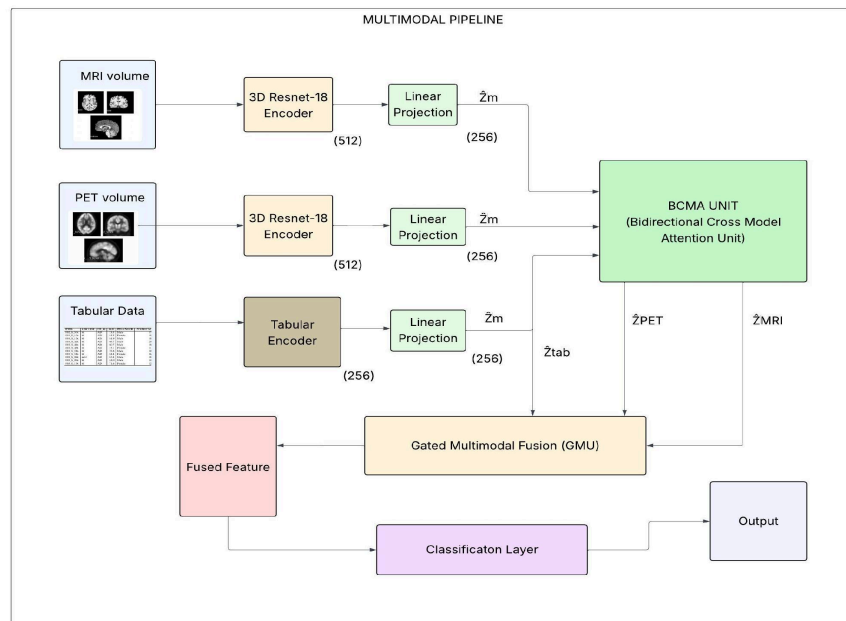


Fig. 2: Multimodal System Architecture

The multimodal framework extends the unimodal architecture by jointly incorporating three complementary modalities: 3D MRI, 3D PET, and tabular clinical data. The model is formally defined as:

$$\mathcal{W} = \{f_e^{\text{MRI}}, f_e^{\text{PET}}, f_e^{\text{tab}}, f_{\text{proj}}, f_{\text{fuse}}, f_c\} \quad (9)$$

where,

- $f_e^{\text{MRI}}: X_{\text{MRI}} \rightarrow Z_{\text{MRI}}$ $f_e^{\text{PET}}: X_{\text{PET}} \rightarrow Z_{\text{PET}}$ are the imaging encoders — identical in architecture to those used in the unimodal setting, and
- $f_e^{\text{tab}}: X_{\text{tab}} \rightarrow Z_{\text{tab}}$ is an MLP-based encoder for the tabular modality, which includes demographic variables, APOE4 genetic risk status, CSF biomarkers ($A\beta_{42}$, T-tau, P-tau), and derived biomarker ratios.

Each encoder independently maps its respective input to a modality-specific latent representation:

$$\mathbf{z}_m = f_e^m(\mathbf{x}_m), \quad m \in \{\text{MRI, PET, tab}\} \quad (10)$$

Prior to fusion, each latent representation is passed through a modality-specific linear projection $f_{\text{proj}}^m: Z_m \rightarrow \mathbb{R}^{256}$, mapping all modalities into a common 256-dimensional space:

$$\tilde{\mathbf{z}}_m = f_{\text{proj}}^m(\mathbf{z}_m) \in \mathbb{R}^{256}, \quad m \in \{\text{MRI, PET, tab}\} \quad (11)$$

The projected representations are then combined by the fusion module $f_{\text{fuse}}: \mathbb{R}^{256} \times \mathbb{R}^{256} \times \mathbb{R}^{256} \rightarrow \mathbb{R}^{512}$, which produces a unified joint representation:

$$\mathbf{z}_{\text{fused}} = f_{\text{fuse}}(\tilde{\mathbf{z}}_{\text{MRI}}, \tilde{\mathbf{z}}_{\text{PET}}, \tilde{\mathbf{z}}_{\text{tab}}) \in \mathbb{R}^{512} \quad (12)$$

The final prediction is obtained by passing $\mathbf{z}_{\text{fused}}$ through the classifier f_c :

$$\hat{y} = f_c(\mathbf{z}_{\text{fused}}) \quad (13)$$

The entire multimodal learning framework is described in Algorithm 2.

Algorithm 2 Multimodal Alzheimer's Classification Pipeline

```

1: Input: MRI volume  $\mathbf{x}_{\text{MRI}} \in \mathbb{R}^{1 \times D \times H \times W}$ , PET volume  $\mathbf{x}_{\text{PET}} \in \mathbb{R}^{1 \times D \times H \times W}$ , Tabular
   vector  $\mathbf{x}_{\text{tab}} \in \mathbb{R}^d$ 
2: Output: Predicted class label  $\hat{y} \in \{\text{MCI}, \text{AD}\}$ 
3: Step 1: Modality-Specific Encoding
4:  $\mathbf{z}_{\text{MRI}} \leftarrow f_e^{\text{MRI}}(\mathbf{x}_{\text{MRI}})$  ( $\in \mathbb{R}^{512}$ )
5:  $\mathbf{z}_{\text{PET}} \leftarrow f_e^{\text{PET}}(\mathbf{x}_{\text{PET}})$  ( $\in \mathbb{R}^{512}$ )
6:  $\mathbf{z}_{\text{tab}} \leftarrow f_e^{\text{tab}}(\mathbf{x}_{\text{tab}})$  ( $\in \mathbb{R}^{256}$ )
7: Step 2: Linear Projection to Common Dimension
8: for each  $m \in \{\text{MRI}, \text{PET}, \text{tab}\}$  do
9:    $\tilde{\mathbf{z}}_m \leftarrow f_{\text{proj}}^m(\mathbf{z}_m)$  ( $\in \mathbb{R}^{256}$ )
10: end for
11: Step 3: Bidirectional Cross-Modal Attention
12:  $\tilde{\mathbf{z}}'_{\text{MRI}} \leftarrow \text{CMA}(\tilde{\mathbf{z}}_{\text{MRI}}, \tilde{\mathbf{z}}_{\text{PET}})$ 
13:  $\tilde{\mathbf{z}}'_{\text{PET}} \leftarrow \text{CMA}(\tilde{\mathbf{z}}_{\text{PET}}, \tilde{\mathbf{z}}_{\text{MRI}})$ 
14:  $\tilde{\mathbf{z}}''_{\text{MRI}} \leftarrow \text{CMA}(\tilde{\mathbf{z}}'_{\text{MRI}}, \tilde{\mathbf{z}}_{\text{tab}})$ 
15:  $\tilde{\mathbf{z}}''_{\text{PET}} \leftarrow \text{CMA}(\tilde{\mathbf{z}}'_{\text{PET}}, \tilde{\mathbf{z}}_{\text{tab}})$ 
16: Step 4: Gated Multimodal Fusion
17:  $\mathbf{z}_{\text{fused}} \leftarrow \text{GMU}(\tilde{\mathbf{z}}''_{\text{MRI}}, \tilde{\mathbf{z}}''_{\text{PET}}, \tilde{\mathbf{z}}_{\text{tab}})$  ( $\in \mathbb{R}^{512}$ )
18: Step 5: Fusion Projection
19:  $\mathbf{z}_{\text{out}} \leftarrow \text{LN}(\text{GELU}(\mathbf{W}_{\text{fuse}}\mathbf{z}_{\text{fused}} + \mathbf{b}_{\text{fuse}}))$  ( $\in \mathbb{R}^{512}$ )
20: Step 6: Classification
21:  $\mathbf{h}_1 \leftarrow \text{GELU}(\mathbf{W}_1\mathbf{z}_{\text{out}} + \mathbf{b}_1)$ 
22:  $\mathbf{h}'_1 \leftarrow \text{Dropout}(\mathbf{h}_1, p = 0.5)$ 
23:  $\mathbf{h}_2 \leftarrow \text{GELU}(\mathbf{W}_2\mathbf{h}'_1 + \mathbf{b}_2)$ 
24:  $\mathbf{h}'_2 \leftarrow \text{Dropout}(\mathbf{h}_2, p = 0.5)$ 
25:  $\hat{y} \leftarrow \arg \max(\mathbf{W}_3\mathbf{h}'_2 + \mathbf{b}_3)$ 
26: return  $\hat{y}$ 

```

Architectural Details**Modality Encoders**

The encoders for MRI and PET modalities, f_e^{MRI} and f_e^{PET} , follow the same architecture as described in the unimodal setting (Section [sec:unimodal_arch]), each producing a latent representation $\mathbf{z}_m \in \mathbb{R}^{512}$. As in the unimodal case, the architecture is shared but the weights are learned independently for each modality.

Tabular Encoder: Unlike imaging data, tabular features lack spatial structure and are better handled by fully connected networks. Hence, the tabular encoder $f_e^{\text{tab}}: X_{\text{tab}} \rightarrow Z_{\text{tab}}$ is implemented as a Multi-Layer Perceptron (MLP) with residual connections, designed to handle heterogeneous clinical and biomarker features. It

maps the input feature vector $x_{tab} \in R^d$, comprising demographic variables, APOE4 status, CSF biomarkers ($A\beta_{42}$, T-tau, P-tau), and derived biomarker ratios to a latent representation $z_{tab} \in R^d$. The encoder takes an input vector $x_{tab} \in R^d$ and produces a latent representation $z_{tab} \in R^{256}$. The network consists of two residual blocks followed by a linear projection. Each residual block applies the following transformation:

$$\mathbf{h} = \text{Dropout}(\text{GELU}(\text{LN}(\mathbf{W}_2 \cdot \text{GELU}(\text{LN}(\mathbf{W}_1 \cdot \mathbf{x} + \mathbf{b}_1)) + \mathbf{b}_2))) + f_{sc}(\mathbf{x}) \quad (14)$$

where,

- $\text{LN}(\cdot)$ denotes Layer Normalization,
- $\text{GELU}(\cdot)$ is the activation function, and
- $f_{sc}(x)$ is a shortcut connection — a linear projection when the input and output dimensions differ, or an identity mapping otherwise.

The GELU (Gaussian Error Linear Unit)[25] activation function is defined as

$$\text{GELU}(x) = x \cdot \Phi(x) \quad (15)$$

where, $\Phi(x)$ is the cumulative distribution function (CDF) of the standard normal distribution. It is a modified form of ReLU allows small negative outputs which deals with the the dying neuron problem by maintaining non-zero gradients for negative inputs. The first residual block projects the input from R^d to a hidden dimension R^{128} , while the second block operates entirely within R^{128} . The output of the second block is then linearly projected to the final embedding dimension:

$$\mathbf{z}_{tab} = \mathbf{W}_{proj} \cdot \mathbf{h}_2 + \mathbf{b}_{proj} \in \mathbb{R}^{256} \quad (16)$$

Layer normalization is particularly suitable for tabular data because it operates independently of batch statistics, ensuring stable behavior even with small batch sizes. The use of GELU, a smooth and probabilistic activation function, further improves learning for continuous-valued. When the dimensionality changes between layers, a linear projection aligns the input with the output before addition. This design enables the model to capture complex relationships among clinical variables such as interactions between age, genetic risk, and biomarker levels without losing important baseline information. Dropout ($p = 0.3$) further enhances by regularizing training, thereby, improving generalization.

Fusion Module

The fusion module f_{fuse} implements an intermediate fusion strategy that combines the three projected modality representations $\tilde{z}_{MRI}, \tilde{z}_{PET}, \tilde{z}_{tab} \in R^{256}$ through three sequential steps: bidirectional cross-modal attention, gated fusion, and a final

projection. The intermediate fusion is the one that is responsible for learning how these modalities interact, complement each other, and jointly contribute to Alzheimer's prediction. Instead of simply concatenating features, it uses attention mechanisms and adaptive gating to model cross-modal relationships in a biologically meaningful way.

Bidirectional Cross-modal attention

Cross-modal attention (CMA) allows each modality to selectively attend to relevant information from another modality. Given a query modality m_q and a key-value modality m_{kv} , the cross-modal attention operation is defined as:

$$\text{CMA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax} \left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}} \right) \mathbf{V} \quad (17)$$

where,

- $\mathbf{Q} = W_Q \tilde{\mathbf{z}}_{m_q}$, $\mathbf{K} = W_K \tilde{\mathbf{z}}_{m_{kv}}$, $\mathbf{V} = W_V \tilde{\mathbf{z}}_{m_{kv}}$ are learned linear projections, and
- d_k is the key dimension.

The attention is applied in two rounds. In the first round, MRI and PET mutually attend to each other:

$$\tilde{\mathbf{z}}'_{\text{MRI}} = \text{CMA}(\tilde{\mathbf{z}}_{\text{MRI}}, \tilde{\mathbf{z}}_{\text{PET}}), \quad \tilde{\mathbf{z}}'_{\text{PET}} = \text{CMA}(\tilde{\mathbf{z}}_{\text{PET}}, \tilde{\mathbf{z}}_{\text{MRI}}) \quad (18)$$

It allows for the interaction between modalities i.e. MRI asks: "Which PET features are relevant to my structure?" and PET asks: "Which MRI regions provide context to my metabolism?"

In the second round, the updated imaging representations attend to the tabular biomarker context:

$$\tilde{\mathbf{z}}''_{\text{MRI}} = \text{CMA}(\tilde{\mathbf{z}}'_{\text{MRI}}, \tilde{\mathbf{z}}_{\text{tab}}), \quad \tilde{\mathbf{z}}''_{\text{PET}} = \text{CMA}(\tilde{\mathbf{z}}'_{\text{PET}}, \tilde{\mathbf{z}}_{\text{tab}}) \quad (19)$$

This two-round design allows imaging features to first be refined by inter-modal correspondence and then grounded in biochemical context provided by the CSF biomarkers and clinical variables.

Gated Multimodal Unit

The attended representations $\tilde{\mathbf{z}}''_{\text{MRI}}$, $\tilde{\mathbf{z}}''_{\text{PET}}$, $\tilde{\mathbf{z}}_{\text{tab}}$ are combined using a Gated Multimodal Unit (GMU) (Arevalo et al. 2020), which learns soft gates to control the contribution of each modality to the fused representation:

$$\mathbf{z}_{\text{fused}} = \text{GMU}(\tilde{\mathbf{z}}''_{\text{MRI}}, \tilde{\mathbf{z}}''_{\text{PET}}, \tilde{\mathbf{z}}_{\text{tab}}) \in \mathbb{R}^{512} \quad (20)$$

It serves as the decision-making component of the multimodal system. Instead of naively combining modalities, it enables the model to dynamically learn how much to trust each modality. The Gated Multimodal Unit (GMU) function is described in Algorithm 3.

Algorithm 3 Gated Multimodal Fusion (GMU)

Require: Modalities $\{x_1, x_2, \dots, x_M\}$, weights $\{W_m, U_m\}_{m=1}^M$

Ensure: Fused representation $\mathbf{y}$

```

1: Initialization:  $\mathcal{Z} \leftarrow [], \mathcal{H} \leftarrow []$ 
2: for each modality  $m = 1$  to  $M$  do
3:   Compute gate signal:  $z_m \leftarrow \tanh(W_m \cdot x_m)$ 
4:   Compute transformed feature:  $h_m \leftarrow \tanh(U_m \cdot x_m)$ 
5:   Append  $z_m$  to  $\mathcal{Z}$  and  $h_m$  to  $\mathcal{H}$ 
6: end for
7: Compute gating weights:  $\mathbf{g} \leftarrow \text{softmax}(\mathcal{Z})$ 
8: Weighted fusion:  $\mathbf{y} \leftarrow \sum_{m=1}^M g_m \cdot h_m$ 
9: Normalization (optional):  $\mathbf{y} \leftarrow \text{LayerNorm}(\mathbf{y})$ 
10: return  $\mathbf{y}$ 
  
```

GMU is effective because different modalities carry varying importance depending on the clinical context which is shown in Table 2.

Table 2: Modality Importance by Clinical Case

Case	Important Modality
Early Alzheimer's Disease	PET (amyloid)
Late Alzheimer's Disease	MRI (atrophy)
Risk Prediction	Tabular (APOE, age)

Thus, GMU dynamically adapts modality importance as per the given sample.

Projection

The fused representation is passed through a linear layer followed by Layer Normalization and GELU activation to produce the final input to the classifier:

$$\mathbf{z}_{\text{out}} = \text{LN}(\text{GELU}(\mathbf{W}_{\text{fuse}} \mathbf{z}_{\text{fused}} + \mathbf{b}_{\text{fuse}})) \in \mathbb{R}^{512} \quad (21)$$

This trimodal fusion composed of 3 steps integrates cross-modal attention for interaction, tabular conditioning for global context, gated fusion for adaptive weighting and projection layers for feature refinement. This design enables the model to dynamically combine structural, functional, and clinical information for robust multimodal learning.

Classifier

The custom classifier f_c consists of stacked fully connected layers with GELU activations and dropout regularization. Its purpose is to take the fused feature vector and transform it to the predicted logits.

Formally, the computation is defined as:

$$\mathbf{h}_1 = \text{GELU}(\mathbf{W}_1 \mathbf{z} + \mathbf{b}_1) \quad (22)$$

$$\mathbf{h}'_1 = \text{Dropout}(\mathbf{h}_1, p = 0.5) \quad (23)$$

$$\mathbf{h}_2 = \text{GELU}(\mathbf{W}_2 \mathbf{h}'_1 + \mathbf{b}_2) \quad (24)$$

$$\mathbf{h}'_2 = \text{Dropout}(\mathbf{h}_2, p = 0.5) \quad (25)$$

$$f_c(\mathbf{z}) = \mathbf{W}_3 \mathbf{h}'_2 + \mathbf{b}_3 \quad (26)$$

IV. RESULTS AND DISCUSSION

MRI Unimodal System

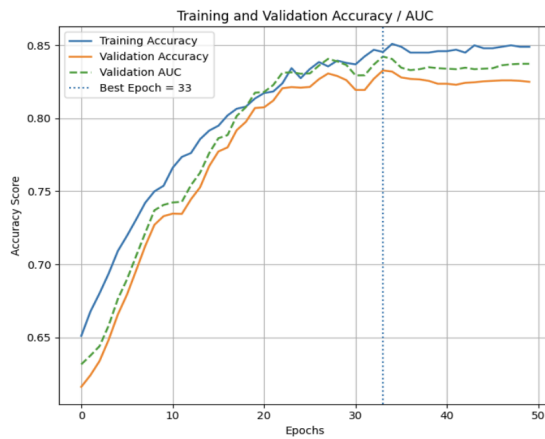


Fig. 3: Accuracy/AUC Graph of Proposed Unimodal Model for MRI Scan

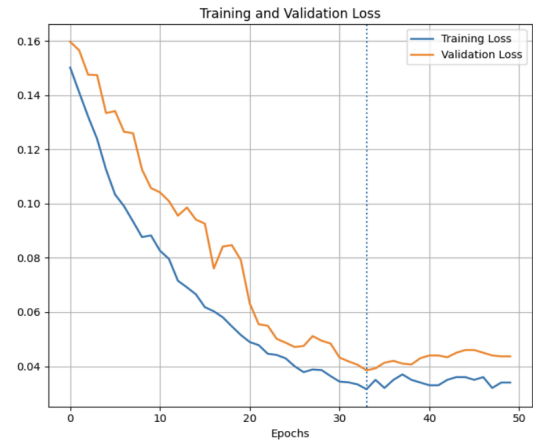


Fig. 4: Loss Graph of Proposed Unimodal Model for MRI Scan

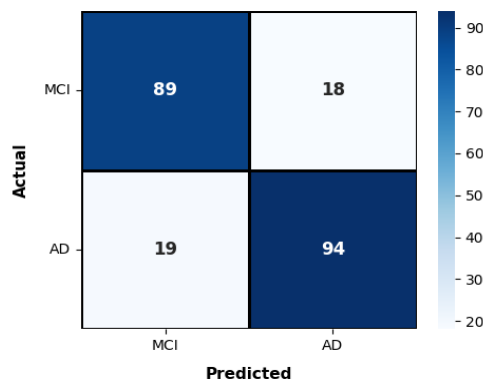


Fig. 5: Confusion Matrix of Proposed Unimodal Model for MRI Scan

The graph of MRI training and validation accuracy and AUC shows that the model converged gradually on the beginning, having few fluctuations and spikes on the middle, while finally converging in the end. The best epoch was at 33 with validation accuracy 0.838, training accuracy 0.846, and AUC 0.840. The graph for loss shows that the validation loss keeps fluctuating when dealt with a hard sample, but eventually converges. The precision and sensitivity is 0.8240 and 0.8317 respectively.

PET Unimodal System

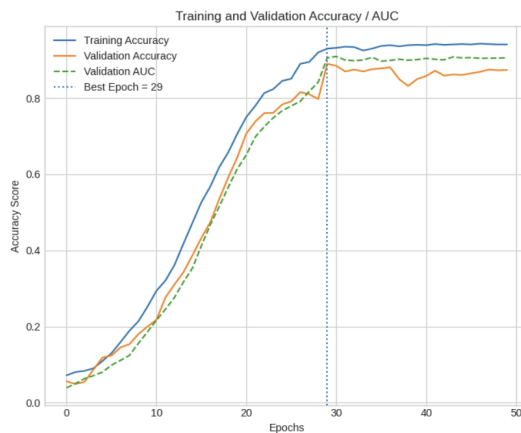


Fig. 6: Accuracy/AUC Graph of Proposed Unimodal Model for PET Scan

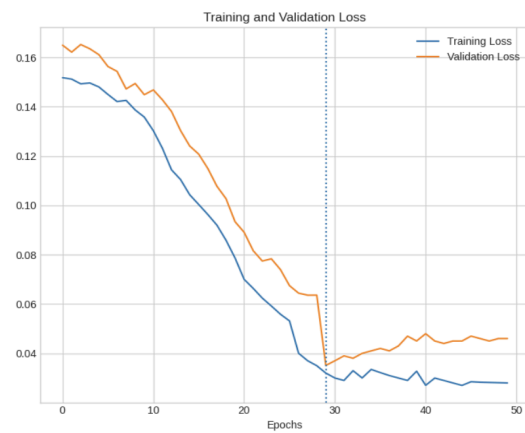


Fig. 7: Loss Graph of Proposed Unimodal Model for PET Scan

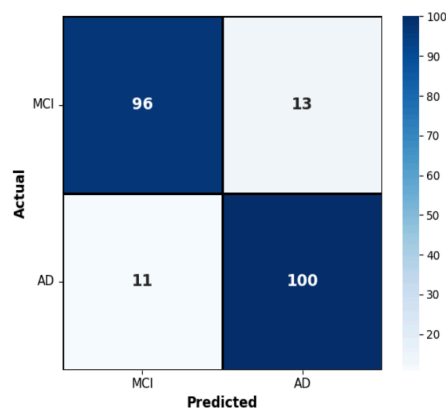


Fig. 8: Confusion Matrix of Proposed Unimodal Model for PET Scan

The accuracy graph of PET scan shows that it had lesser fluctuation than MRI, and had a smoother rising graph. The best epoch is at 29 where the value fluctuates around it but eventually converges. The validation accuracy at best epoch is 0.899, training accuracy is 0.931 and AUC is 0.915. The loss graph shows a sudden dip at best epoch which later recovers and converges. The training loss and validation loss at best epoch are 0.030 and 0.039 respectively.

Multimodal System

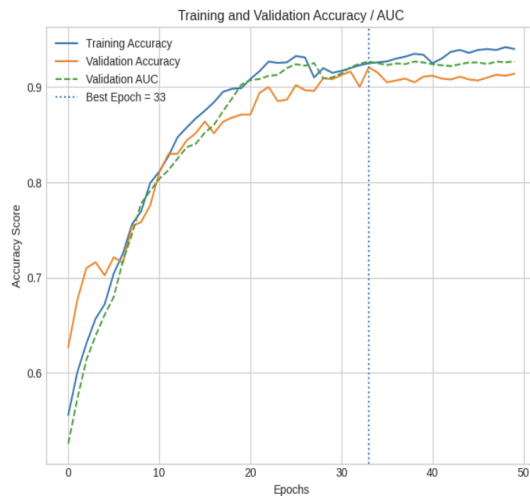


Fig. 9: Accuracy/AUC Graph of Proposed Multimodal Model

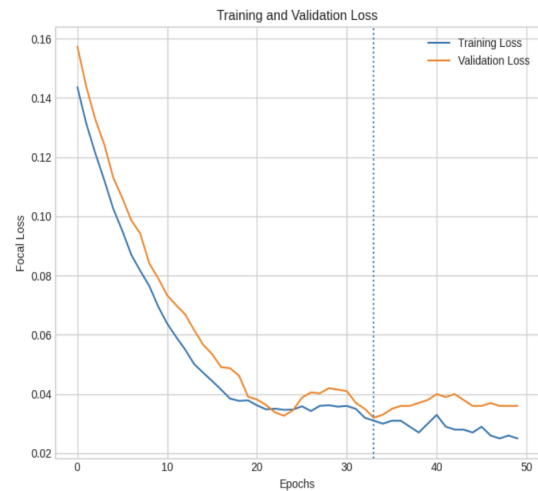


Fig. 10: Loss Graph of Proposed Multimodal Model

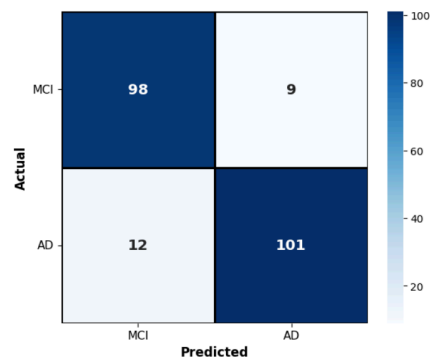


Fig. 11: Confusion Matrix of the Proposed Multimodal Model

The accuracy graph of Multimodal accuracy shows that it had high fluctuation but also had highest accuracy among all three models. The best epoch is at 33 where the values of the validation accuracy is 0.930, training accuracy is 0.951 and AUC is 0.940. The training loss and validation loss at best epoch is 0.034 and 0.035 respectively.

Comparative Analysis

Table 3 shows the comparative analysis between different training modality we have used.

Table 3: Comparison Between Different Model Architecture

Model	AUC	Accuracy	
		Training Accuracy	Validation Accuracy
Unimodal MRI Model	0.840	0.846	0.838
Unimodal PET Model	0.915	0.931	0.899
Multimodal Model	0.94	0.951	0.930

As we can see from the graphs and table above, multimodal system performed better in comparison to unimodal systems. Multimodal has the highest AUC of 0.94, with unimodal PET having AUC of 0.915, and unimodal MRI model having the lowest AUC of 0.840 compared at their best epoch respectively. The loss graph of MRI showed greater fluctuations in the earlier epoch, loss of PET showed a great dip near the best epoch, but the losses of all three models converged eventually.

Hence, we can conclude that multimodal approach provided a better AUC, than unimodal approach, and is the best method for alzheimers disease stage classification.

V. CONCLUSION AND LIMITATION

This study successfully developed and evaluated unimodal (MRI and PET) and multimodal deep learning models for Alzheimer's disease stage classification using standard datasets. The results clearly show that all models were able to learn meaningful patterns from neuroimaging data, with PET performing better than MRI, and the multimodal system achieving the best overall performance. The multimodal model obtained the highest validation accuracy (0.930) and AUC (0.940), demonstrating that combining complementary information from different modalities leads to more reliable and discriminative predictions compared to using a single data source.

Overall, the main objective of the study—to investigate whether multimodal learning improves Alzheimer's disease classification compared to unimodal approaches—was achieved. The results consistently confirm that integrating MRI, PET, and clinical data enhances model performance and stability, particularly in distinguishing disease stages such as MCI and AD. The use of fusion strategies such as gated multimodal units and cross-modal attention further strengthened the model by enabling more meaningful interaction between different data types.

These findings highlight the importance of multimodal learning in medical diagnosis, especially for complex diseases like Alzheimer's, where no single data source is sufficient to capture all pathological changes. In practical terms, the proposed system demonstrates potential to support clinicians by providing faster, more consistent, and

data-driven decision support, particularly in resource-limited settings where specialist availability is low.

However, there is still room for improvement. Future work should focus on expanding the system to multi-class classification, including earlier preclinical stages, and improving data integration strategies to handle missing modalities more effectively. Additionally, optimizing the model for real-time or clinical deployment would further enhance its practical usability and impact.

REFERENCES

- [1] World Health Organization, "Dementia." WHO Fact Sheet, Apr. 2025. Accessed: 2025-06-09.
- [2] Alzheimer's Disease International, "Dementia statistics," 2025. Accessed: 2025-06-09.
- [3] The Kathmandu Post, "Dementia is a growing health epidemic, but no one is paying attention," The Kathmandu Post. "More than 78,000 Nepalis suffer from dementia . . . and the number is expected to double every 20 years."
- [4] Wikipedia contributors, "Health in nepal — wikipedia, the free encyclopedia," 2024. Accessed: 2025-07-03.
- [5] L. P. Bhatt and S. Shakya, "Deep learning-driven uncertainty estimation for brain tissue segmentation by integrating expectation maximization algorithms," in 2025 4th International Conference on Sentiment Analysis and Deep Learning (ICSADL), pp. 914–920, 2025.
- [6] S. Sitoula, L. P. Bhatt, A. Pokhrel, and A. Neupane, "Generative AI for mental well-being and insights from high and low resource languages," in Enhancing Edge Devices With Generative Intelligence: Architecture, Challenges, and Applications (S. Shakya, T. Shahi, E. Ceh-Varela, and M. Cowling, eds.), pp. 391–424, IGI Global Scientific Publishing, 2026.
- [7] S. Lamichhane, L. P. Bhatt, and S. Shakya, "Deeprust: A hybrid transformer-cnn model for deepfake detection with zero-knowledge-based blockchain authentication," Journal of Advanced College of Engineering and Management, vol. 12, p. 57–71, May 2026.
- [8] A. Pokhrel, L. Mahara, M. Upadhyaya, S. Shrestha, and B. R. Lamichhane, "Drowsy driver detection with crash alert mechanism using arduino and image processing," Journal of Soft Computing Paradigm, vol. 5, p. 194–217, Jul. 2023.
- [9] W. Lin, Q. Gao, M. Du, W. Chen, and T. Tong, "Multiclass diagnosis of stages of alzheimer's disease using linear discriminant analysis scoring for multimodal data," Computers in Biology and Medicine, vol. 134, p. 104478, 2021.
- [10] H. Helaly, M. Badawy, and A. Haikal, "Deep learning approach for early detection of alzheimer's disease," Cognitive Computation, vol. 14, pp. 1711–1727, 2022.
- [11] M. Maqsood, F. Nazir, U. Khan, F. Aadil, H. Jamal, I. Mehmood, and O.-y. Song, "Transfer learning assisted classification and detection of alzheimer's disease stages using 3d mri scans," Sensors, vol. 19, no. 11, p. 2645, 2019.

- [12] A. B. Tufail, Y.-K. Ma, and Q.-N. Zhang, "Binary classification of alzheimer's disease using smri imaging modality and deep learning," *Journal of Digital Imaging*, vol. 33, no. 5, pp. 1073–1090, 2020.
- [13] R. Baral, S. Satyal, and A. Pokhrel, "Cnn-transformer based speech emotion detection," *Journal of Advanced College of Engineering and Management*, vol. 10, p. 135–145, Mar. 2025.
- [14] D. Zhang, Y. Wang, L. Zhou, H. Yuan, D. Shen, and the Alzheimer's Disease Neuroimaging Initiative, "Multimodal classification of alzheimer's disease and mild cognitive impairment," *NeuroImage*, vol. 55, no. 3, pp. 856–867, 2011.
- [15] J. Venugopalan, L. Tong, H. R. Hassanzadeh, and M. D. Wang, "Multimodal deep learning models for early detection of alzheimer's disease stage," *Scientific Reports*, vol. 11, p. 3254, 2021.
- [16] T. Tong, K. Gray, Q. Gao, L. Chen, D. Rueckert, and A. D. N. Initiative, "Multimodal classification of alzheimer's disease using nonlinear graph fusion," *Pattern Recognition*, vol. 63, pp. 171–181, 2017.
- [17] Alzheimer's Disease Neuroimaging Initiative, "Alzheimer's disease neuroimaging initiative (adni)." <http://adni.loni.usc.edu>. Accessed: 2025-06-09.
- [18] National Alzheimer's Coordinating Center (NACC), "National alzheimer's coordinating center," 2025.
- [19] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Doll'ar, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, 2020.
- [20] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.
- [21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [22] M. J. Cardoso, W. Li, R. Brown, N. Ma, E. Kerfoot, Y. Wang, B. Murrey, A. Myronenko, C. Zhao, D. Yang, V. Nath, Y. He, Z. Xu, A. Hatamizadeh, A. Myronenko, W. Zhu, Y. Liu, M. Zheng, Y. Tang, I. Yang, M. Zephyr, B. Hashemian, S. Alle, M. Z. Darestani, C. Budd, M. Modat, T. Vercauteren, G. Wang, Y. Li, Y. Hu, Y. Fu, B. Gorman, H. Johnson, B. Genereaux, B. S. Erdal, V. Gupta, A. Diaz-Pinto, A. Dourson, L. Maier-Hein, P. F. Jaeger, M. Baumgartner, J. Kalpathy-Cramer, M. Flores, J. Kirby, L. A. D. Cooper, H. R. Roth, D. Xu, D. Bericat, R. Floca, S. K. Zhou, H. Shuaib, K. Farahani, K. H. Maier-Hein, S. Aylward, P. Dogra, S. Ourselin, and A. Feng, "Monai: An open-source framework for deep learning in healthcare," 2022.
- [23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.
- [24] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," *arXiv preprint arXiv:1711.11248*, 2017.