Hybrid Multi-Modal Deep Learning for Alzheimer's Disease Detection Using MRI Images and Clinical Metadata

Aniket Ghosh
Department of Computer Science and Engineering
National Institute of Technology Calicut
Email: aniket$_m$240303$cs@nitc.ac.in$

***Abstract*—Alzheimer's disease (AD) is a progressive neurodegenerative disorder requiring early and accurate diagnosis for effective management. This paper presents a hybrid multi-modal deep learning framework that combines structural MRI brain scans with clinical patient metadata from the OASIS dataset. A convolutional neural network (CNN) processes resized MRI slices to capture brain atrophy patterns, while a multi-layer perceptron (MLP) processes seven normalized clinical features (Age, Educ, SES, MMSE, eTIV, nWBV, ASF). Features from both branches are concatenated and passed through a final dense layer for four-class classification: NonDemented, VeryMildDemented, MildDemented, and ModerateDemented. The model achieves 83% overall accuracy on the test set, with particularly strong performance on the majority NonDemented class (F1-score 0.92) and promising recall (0.80) for the clinically important MildDemented class, demonstrating the value of multi-modal fusion for automated AD screening.**

***Index Terms*—Alzheimer's disease, multi-modal deep learning, convolutional neural network, clinical metadata, MRI, OASIS dataset**

I. Introduction

Alzheimer's disease is the most common cause of dementia, affecting millions worldwide. Early detection is critical yet challenging due to subtle structural changes in the brain and overlapping cognitive symptoms. Traditional diagnosis relies on neuroimaging (MRI, PET) combined with clinical assessments such as the Mini-Mental State Examination (MMSE) and Clinical Dementia Rating (CDR).

Recent deep learning approaches have shown success in analyzing MRI scans, but most treat images in isolation. Clinical practice, however, integrates both imaging and patient metadata. This work addresses that gap by proposing a hybrid dual-branch architecture that simultaneously processes MRI slices and tabular clinical data, mimicking real-world diagnostic reasoning.

The contributions are: (1) an end-to-end multi-modal pipeline built on the public OASIS dataset, (2) a lightweight hybrid CNN+MLP model suitable for deployment, and (3) empirical evidence that fusing imaging and clinical features improves classification, especially for early-stage dementia.

II. Dataset Overview

The dataset is the publicly available "OASIS Alzheimer's Detection Multi-Class Dataset" [1]. It consists of 2D axial MRI slices extracted from 3D NIfTI volumes of patients from the Open Access Series of Imaging Studies (OASIS).

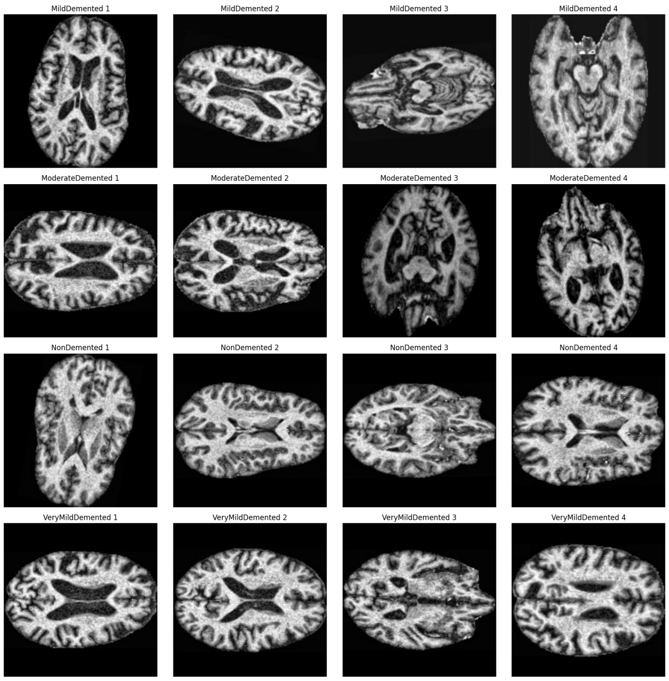

Fig. 1: Class distribution visualization highlighting severe imbalance across the four diagnostic categories in the combined train/test sets.

TABLE I: Class distribution in the provided train/test splits (image-level counts)

Class	Train Images	Test Images
NonDemented	$\approx$ 53,000	44,596
VeryMildDemented	$\approx$ 5,300	4,505
MildDemented	$\approx$ 800	651
ModerateDemented	$\approx$ 200	168
Total	$\approx$ 59,304	49,920

Each image is paired with patient-level metadata (Age, Education (Educ), Socioeconomic Status (SES), MMSE, estimated Total Intracranial Volume (eTIV), normalized Whole Brain Volume (nWBV), Atlas Scaling Factor (ASF)). The metadata

CSV files provide one row per patient scan session; image filenames encode the patient ID, enabling perfect merging without missing values.

Severe class imbalance is evident (Fig. 1 and Table I), with NonDemented dominating and ModerateDemented being extremely rare.

A. Sample Images

To illustrate the visual characteristics of the data, Figure 2 shows representative 128×128 axial MRI slices from each diagnostic class. Progressive structural changes, such as ventricular enlargement and cortical thinning, become more pronounced with increasing dementia severity.

III. Problem Statement

Given a 128×128×3 MRI slice and its corresponding seven clinical features, classify the slice into one of four Alzheimer's progression stages: NonDemented, VeryMildDemented, MildDemented, or ModerateDemented. The objective is to maximize overall accuracy while maintaining acceptable recall on minority dementia classes for potential clinical screening utility.

IV. Methodology

The proposed hybrid multi-modal architecture (Fig. 3) consists of two parallel branches whose features are fused through late concatenation.

Let $I \in \mathbb{R}^{128\times128\times3}$ be the input MRI slice and $\mathbf{x}_{tab} \in \mathbb{R}^7$ the normalized clinical feature vector.

The CNN branch extracts visual features:

$$\mathbf{f}_{CNN} = \text{GAP}\left(\text{ConvBlock}_3\left(\cdots \text{ConvBlock}_1(I)\cdots\right)\right) \in \mathbb{R}^{128} \tag{1}$$

where each ConvBlock consists of Conv2D → BatchNorm → ReLU → MaxPooling.

The tabular branch processes clinical data:

$$\mathbf{f}_{MLP} = W_2 \cdot \sigma\left(W_1\mathbf{x}_{tab} + \mathbf{b}_1\right) + \mathbf{b}_2 \in \mathbb{R}^{16} \tag{2}$$

with $W_1 \in \mathbb{R}^{64\times7}$, $W_2 \in \mathbb{R}^{16\times64}$, and σ the ReLU activation.

Late fusion concatenates both representations:

$$\mathbf{z} = [\mathbf{f}_{CNN}; \mathbf{f}_{MLP}] \in \mathbb{R}^{144} \tag{3}$$

The final prediction is obtained via:

$$\mathbf{h} = \sigma\left(W_h\mathbf{z} + \mathbf{b}_h\right) \tag{4}$$

$$\hat{\mathbf{y}} = \text{softmax}\left(W_o \cdot \text{Dropout}(\mathbf{h}) + \mathbf{b}_o\right) \in \mathbb{R}^4 \tag{5}$$

where $W_h \in \mathbb{R}^{64\times144}$ and $W_o \in \mathbb{R}^{4\times64}$.

The model is trained by minimizing categorical cross-entropy:

$$\mathcal{L} = -\sum_{c=1}^{4} y_c \log(\hat{y}_c) \tag{6}$$

Total trainable parameters: approximately 103,572.

TABLE II: Classification report on test set

Class	Precision	Recall	F1-Score	Support
MildDemented	0.23	0.80	0.36	651
ModerateDemented	0.00	0.00	0.00	168
NonDemented	0.98	0.86	0.92	44,596
VeryMildDemented	0.31	0.57	0.40	4,505
Accuracy			0.83	49,920
Macro avg	0.38	0.56	0.42	
Weighted avg	0.91	0.83	0.86	

V. Implementation

Data preparation steps:

1) Load train/test metadata CSVs and extract patient IDs from image filenames.
2) Merge image paths with metadata on ID, yielding complete multi-modal records.
3) Standardize the seven clinical features (zero mean, unit variance).
4) Implement a custom Keras generator yielding batches of (images, tabular) pairs and corresponding labels.
5) Resize all images to 128×128 and normalize pixel values to [0,1].

The model is built using the Keras Functional API. Training employs Adam optimizer (initial learning rate 0.001), EarlyStopping (patience=5), and ReduceLROnPlateau (factor=0.2).

VI. Training

Training was performed on a single NVIDIA Tesla T4 GPU. Key observations:

- Epoch 1 achieved training accuracy ≈79% and validation accuracy 83.3%.
- Rapid convergence: by Epoch 4, training accuracy exceeded 99.9%.
- Early stopping triggered after Epoch 6, restoring weights from the epoch with highest validation accuracy (Fig. 4).

Total training time was approximately one hour.

VII. Experiments and Results

The final model (best checkpoint) was evaluated on the held-out test set (49,920 images).

Table II shows strong performance on the majority NonDemented class and clinically useful recall (0.80) for MildDemented. The ModerateDemented class remains challenging due to extreme under-representation.

The confusion matrix (Fig. 5) confirms that most errors occur within adjacent dementia stages, which is acceptable for screening purposes.

VIII. Conclusion

The proposed hybrid multi-modal model successfully integrates MRI imagery and clinical metadata, achieving 83% accuracy on a highly imbalanced four-class Alzheimer's classification task. The architecture demonstrates that even a lightweight CNN+MLP fusion can capture complementary

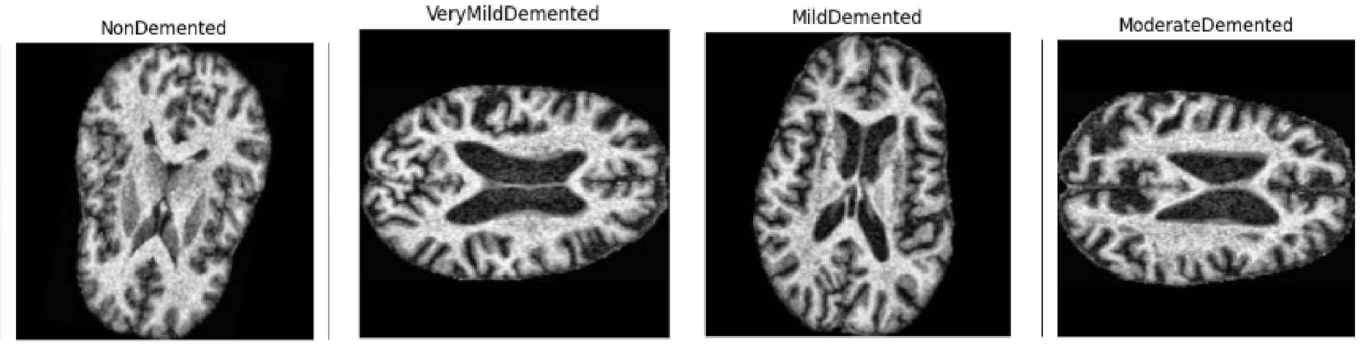

Fig. 2: Representative axial MRI slices from the dataset (left to right): NonDemented, VeryMildDemented, MildDemented, ModerateDemented. The images demonstrate increasing severity of brain atrophy characteristic of Alzheimer's disease progression.

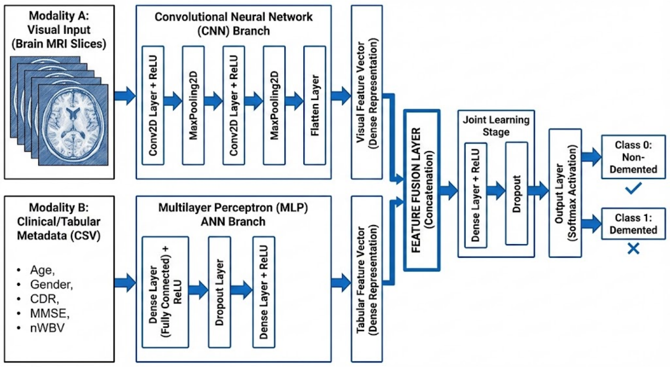
Modality A: Visual Input (Brain MRI Slices)
Convolutional Neural Network (CNN) Branch
Conv2D Layer + ReLU
MaxPooling2D
Conv2D Layer + ReLU
MaxPooling2D
Flatten Layer
Visual Feature Vector (Dense Representation)
Modality B: Clinical/Tabular Metadata (CSV)
• Age,
• Gender,
• CDR,
• MMSE,
• nWBV
Multilayer Perceptron (MLP) ANN Branch
Dense Layer (Fully Connected) + ReLU
Dropout Layer
Dense Layer + ReLU
Tabular Feature Vector (Dense Representation)
FEATURE FUSION LAYER (Concatenation)
Joint Learning Stage
Dense Layer + ReLU
Dropout
Output Layer (Softmax Activation)
Class 0: Non-Demented
✓
Class 1: Demented
✕

Fig. 3: Hybrid multi-modal architecture: CNN branch for MRI images and MLP branch for clinical metadata, fused via concatenation.

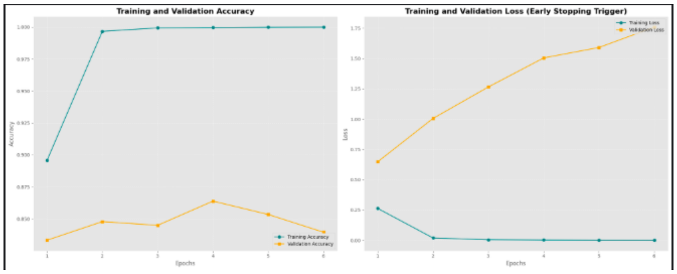

Fig. 4: Training and validation accuracy/loss curves over epochs. Rapid convergence is observed, with early stopping triggered after the peak validation performance.

diagnostic signals, yielding high recall for early-to-moderate dementia stages. Future improvements include synthetic oversampling (SMOTE, GANs) for the ModerateDemented class and attention-based late fusion mechanisms.

REFERENCES

[1] S. Mohanty, "OASIS Alzheimer's Detection Multi-Class Dataset," Kaggle, 2024. [Online]. Available: https://www.kaggle.com/datasets/shreyanmohanty/oasis-alzheimers-detection-multi-class-dataset

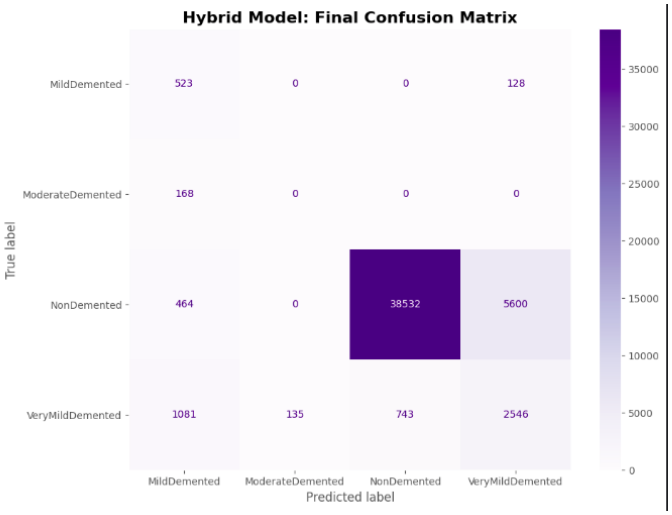

Fig. 5: Confusion matrix on test set (normalized by true class).

[2] D. S. Marcus *et al.*, "Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults," *J. Cogn. Neurosci.*, vol. 19, no. 9, pp. 1498–1507, 2007.