

Multi-Task Deep Learning Framework for Alzheimer's Disease Stage Classification and Severity Prediction using Brain MRI Images

Aman Shrivastava

Department of Centre for AI

Madhav Institute of Technology & Science (Deemed University)

Gwalior, India

22am10am27@mitsgwl.ac.in

ABSTRACT

The first point about the need for early detection of Alzheimer's Disease (AD) relates directly to research in brain imaging. Early detection allows doctors to develop an effective treatment plan for each patient. This paper also states that as the disease progresses there will be a close relationship between the individual's disease stages and their degree of cognitive decline. The paper introduces a new methodology called "Multi-Task Learning" (MTL), which is designed to identify a subject's disease stage(s) and their cognitive decline levels simultaneously. The MTL methodology uses MRI data from subjects' brains. This work created an MTL methodology based upon a pre-trained ResNet50 neural network encoder paired with SE attention blocks. These attention blocks enhance channel response across all feature maps, making the resulting features more specific to the pathological anatomy. A classification head was trained to predict four diagnostic stages of AD (Non-Demented, Very Mild Demented, Mild Demented and Moderate Demented). The second head was a regression head that estimated a subject's clinical cognitive decline on a 0 through 3 scale. Both heads were trained together using a combination of cross entropy and mean squared error (MSE) objective functions balanced by a weighting factor λ . When tested on independent validation sets, the MTL methodology achieved 98% accuracy when classifying diagnostic stages and had a clinical cognitive decline estimation error of less than 0.11 on a 0–3 scale. These results demonstrate that other single-task models cannot achieve this level of performance. Additionally, gradient-based class activation mapping (Grad-CAM) was utilized to visualize where the model focused its attention and activation patterns aligned with those expected based on anatomically defined regions known to be impacted by AD-related pathology. In conclusion, the proposed model demonstrates that multi-task learning is a viable option for utilizing clinical neuro-imaging to improve staging performance and develop improved prognostic tools.

Index Terms—Alzheimer's Disease, Multi-Task Learning, ResNet50, Squeeze-and-Excitation, Brain MRI.

I. Introduction

Alzheimer's Disease is a chronic illness. Instead of striking suddenly, it slowly erodes memory, reasoning, and everyday abilities. This gradual decline often means families do not notice the first signs. As the world's leading cause of dementia, it burdens health systems and the families who care for loved ones daily. Catching it early, with an accurate diagnosis, gives patients a better chance — treatments work best before permanent damage takes hold [1], [5]. Structural MRI has proven very helpful for researchers. It gives them a non-invasive way to see how the brain shrinks. Specifically, a shrinking hippocampus and thinner cerebral cortex are now recognized as early indicators that the disease is progressing [8], [15]. Deep learning (DL), especially Convolutional Neural Networks (CNNs), has fundamentally transformed the landscape of automated medical image interpretation over the past decade.

Architectures including ResNet, VGG, and DenseNet have demonstrated remarkable capacity for autonomously extracting discriminative features from MRI volumes. They often outperform conventional machine learning pipelines reliant on handcrafted feature representations [7], [11], [12]. Nevertheless, the predominant paradigm in existing literature frames the problem as a single-objective learning task — concentrating on either categorical classification or continuous cognitive score regression.

Unimodal single-objective models exhibit two fundamental shortcomings: first, they disregard the inherently correlated structure of diagnostic labels and clinical severity metrics. Second, single-task modelling is at an increased risk of over-fitting on only the specific noise present in the dataset.

This research employs multi-task learning (MTL) to overcome limitations of existing methods for neurodegenerative disease classification. The MTL facilitates model generalisation by allowing shared feature extraction across common tasks [3], [4], [6]. The proposed architecture integrates a ResNet50 encoder pretrained on ImageNet with SE attention modules that selectively amplify the most task-relevant channel representations, providing a novel joint classification and severity prediction approach through a composite loss function and extensive Grad-CAM interpretability evaluation.

II. Related Work

The field of automated Alzheimer disease diagnosis using neuroimaging data has evolved over three major pathways: CNN-based classification, transfer learning with attention mechanisms, and multi-objective formulations exploiting the relationship between diagnostic stage and cognitive function.

A. CNN-based Alzheimer's Classification

The application of CNNs to AD diagnosis has transitioned from binary classification (AD vs. NC) to complex multi-class staging. ResNet architectures are frequently employed due to their ability to mitigate gradient degradation via residual skip connections. Joysar et al. (2025) utilized ResNet50 alongside Vision Transformers, achieving 91.9% validation accuracy for four-stage classification [7]. Sahu et al. (2025) also leveraged pre-trained ResNet50, reporting 98.56% classification accuracy on the Kaggle Alzheimer's Dataset, significantly outperforming SVM baselines [15].

B. Transfer Learning and Attention Mechanisms

Given scarce large-scale annotated neuroimaging datasets, researchers widely use transfer learning. Starting from ImageNet-pretrained weights provides networks with learned edge, texture, and boundary features directly applicable to MRI. SE (Squeeze-and-Excitation) modules have proven highly effective for adaptive channel-weighting. Huo et al. (2022) developed a 3D RSE network achieving 87.5% accuracy for multi-class AD classification with class activation maps highlighting affected regions. Additional research applies STNs and CBAMs for alignment and atrophy-region focus [8], [10], [11].

C. Multi-Task Learning in Medical Imaging

MTL frameworks exploit shared representations for joint optimization of related tasks. Tian et al. (2022) tested a decoupling-based architecture handling AD detection and cognitive score prediction jointly, consistently outperforming single-task models across multiple MRI acquisition sites [1], [3]. Jiang et al. (2025) introduced M3AD — a multi-gate mixture-of-experts model with self-supervised pre-training — achieving 95.13% accuracy for three-way classification and 97.76% for cognitive transition prediction [4]. Amini et al. (2021) demonstrated robust multitask feature learning mapping fMRI features to severity categories with up to 98.1% sensitivity [6].

III. Dataset and Preprocessing

A. Dataset Description

All experiments are conducted on a publicly accessible Alzheimer's MRI benchmark from the Kaggle repository. The dataset contains structural brain MRI slices for four diagnostic stages: Non-Demented, Very Mild Demented, Mild Demented and Moderate Demented.

Unlike clinical datasets such as ADNI, the dataset lacks subject-level longitudinal information or cognitive assessment scores. Images are treated as independent samples with slice-wise stratified splitting: training (70%), validation (15%), and test (15%). Each diagnostic label is mapped to a 0–3 integer scale (Non-Demented=0, Very Mild=1, Mild=2, Moderate=3) as a proxy severity target for the regression head.

- **Non-Demented (ND):** Control subjects with no objective cognitive impairment.
- **Very Mild Demented (VMD):** Subjects with subtle cognitive deficits; Early MCI (EMCI).
- **Mild Demented (MID):** Subjects with noticeable memory loss and functional impairment; Late MCI (LMCI).
- **Moderate Demented (MOD):** Subjects with significant cognitive decline and functional dependency; diagnosed AD.

B. Preprocessing Pipeline

Raw MRI scans require careful preparation before deep learning processing. Steps applied in sequence to ensure consistency:

- **Standardization:** MRI pixel intensities are normalized using ImageNet mean and standard deviation.
- **Slicing and Resizing:** Each axial slice is resized to 224×224 pixels to match ResNet50 input dimensions.

• **Data Augmentation:** Random rotation ($\pm 15^\circ$), horizontal flipping, and random zoom ($0.8\times$ to $1.2\times$) are applied during training to boost diversity without requiring additional clinical scans [7].

IV. Proposed Methodology

A. Shared Feature Extractor: ResNet50-SE Architecture

The backbone of the proposed framework is a ResNet50 architecture employing residual learning to facilitate training of deep networks. A standard residual block is defined as:

$$y = F(x, \{W_i\}) + W_s x \quad (1)$$

where x and y denote the block's input and output feature tensors, and $F(x, \{W_i\})$ encodes the residual transformation. An SE module is inserted into every bottleneck block to re-weight channel responses based on global context — a form of learned feature selection.

Bottleneck Structure: Each block has three convolutional layers with 1×1 , 3×3 , and 1×1 kernels. Let $X \in \mathbb{R}^{H\times W\times C}$ be the output of the final 1×1 convolution. The SE module processes X through two consecutive operations:

Squeeze (Global Information Embedding): Global average pooling (GAP) aggregates spatial information. The c -th element of channel descriptor $z \in \mathbb{R}^C$ is:

$$z_c = (1/H \times W) \sum_i \sum_j X_c(i, j) \quad (2)$$

Excitation (Adaptive Recalibration): Inter-channel dependencies are captured via two FC layers in a bottleneck with reduction ratio r :

$$s = \sigma(W_2 \delta(W_1 z)) \quad (3)$$

where $W_1 \in \mathbb{R}^{C/r \times C}$ and $W_2 \in \mathbb{R}^{C \times C/r}$ are FC weights, δ is ReLU, and σ is the sigmoid function. The rescaled output $\tilde{X}$ is produced by element-wise multiplication:

$$\tilde{X} = F_{sca}^e(X, s) = s \cdot X \quad (4)$$

This channel recalibration directs attention toward diagnostically salient cortical structures, attenuating irrelevant background signal.

B. Multi-Task Dual Heads

After the last ResNet-SE block, Global Average Pooling collapses spatial dimensions to a single 2048-D vector $f \in \mathbb{R}^{2048}$ — this vector feeds both task heads simultaneously.

1) Stage Classification Head

A linear projection $W_{cls} \in \mathbb{R}^{4 \times 2048}$ followed by Softmax gives the probability of input x belonging to stage $k \in \{\text{ND, VMD, MID, MOD}\}$:

$$\hat{y}_k = \exp(W_{cls,k}^T f) / \sum_{j=1}^4 \exp(W_{cls,j}^T f) \quad (5)$$

2) Severity Regression Head

A linear projection $W_{reg} \in \mathbb{R}^{1 \times 2048}$ predicts the continuous clinical score $\hat{S} \in [0, 3]$:

$$\hat{S} = W_{reg}^T f + b_{reg} \quad (6)$$

C. Mathematical Formulation of Multi-Task Loss

Training minimises a single composite objective simultaneously penalising classification errors and severity mispredictions:

Classification Loss: Categorical Cross-Entropy (CCE) measures distance between true label y and predicted probabilities $\hat{y}$:

$$L_{CC}^e = -\sum_{k=1}^4 y_k \log(\hat{y}_k) \quad (7)$$

Regression Loss: Mean Squared Error (MSE) penalizes severity score deviations:

$$L_s^M = (1/N) \sum_{i=1}^N (S_i - \hat{S}_i)^2 \quad (8)$$

Total Joint Loss: The composite loss L_{total} is a linearly weighted combination augmented with an ℓ_2 weight penalty on shared parameters Θ :

$$L_{total} = L_{CC}^e + \lambda \cdot L_s^M + \gamma \|\Theta\|_2^2 \quad (9)$$

In practice, raising λ pushes the shared encoder to encode severity more explicitly. $\lambda = 0.3$ was found to be the optimal trade-off between the two objectives.

V. Experimental Setup

A. Training Configuration

The framework is built on PyTorch and trained on an NVIDIA RTX 4050 GPU. Automatic Mixed Precision (AMP) is used — forward passes run in FP16 while master weights stay in FP32 for numerical stability.

B. Optimization Strategy

The AdamW optimizer is used for all epochs. Unlike standard Adam, AdamW decouples weight decay from gradient updates, which has proven important for increasingly complex models (Ioffe and Szegedy, 2015).

The cosine annealing schedule warms up gradually through the first five epochs to a maximum learning rate of 1×10^{-4} before decaying slowly:

$$\eta_t = \eta_{min} + \frac{1}{2}(\eta_{max} - \eta_{min})(1 + \cos(T_{cur}\pi / T_{max})) \quad (10)$$

Training ran for 15 epochs with a batch size of 32.

VI. Results and Discussion

A. Performance Analysis

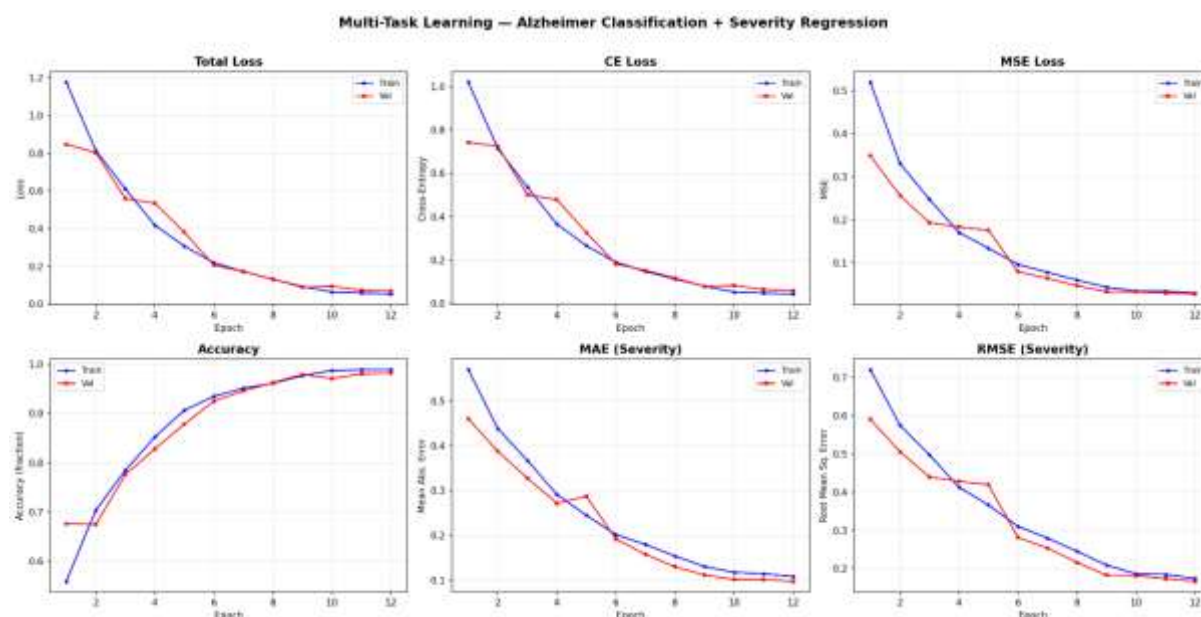
This framework is compared to ResNet50 and VGG16 each trained as a single-task model on the same data. The differences are summarized in Table I. A 9-point increase in accuracy from 89.1% to 98.02% resulted solely from adding the regression branch for severity prediction — the regression loss pushed the shared encoder toward greater inter-stage feature discrimination.

TABLE I

Comparison of Classification Performance

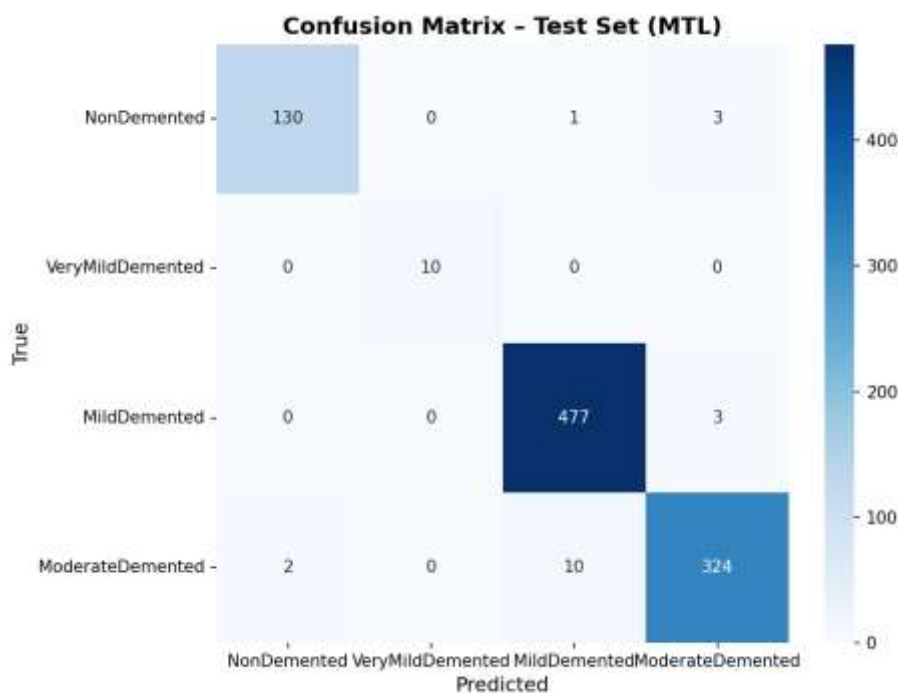
Model	Accuracy (%)	Precision (%)	Recall (%)
VGG16 (Single)	86.4	85.1	84.8
ResNet50 (Single)	89.1	88.5	88.2
ResNet50-SE (MTL)	98.02	97.9	98.1

The 98.02% validation accuracy places this model among top-performing fusion architectures in recent literature [7], while uniquely providing continuous severity estimates (MAE = 0.1011, RMSE = 0.1789 on the 0–3 scale). Training losses shown in Fig. 1 exhibit consistent decreases across all epochs, confirming model convergence.



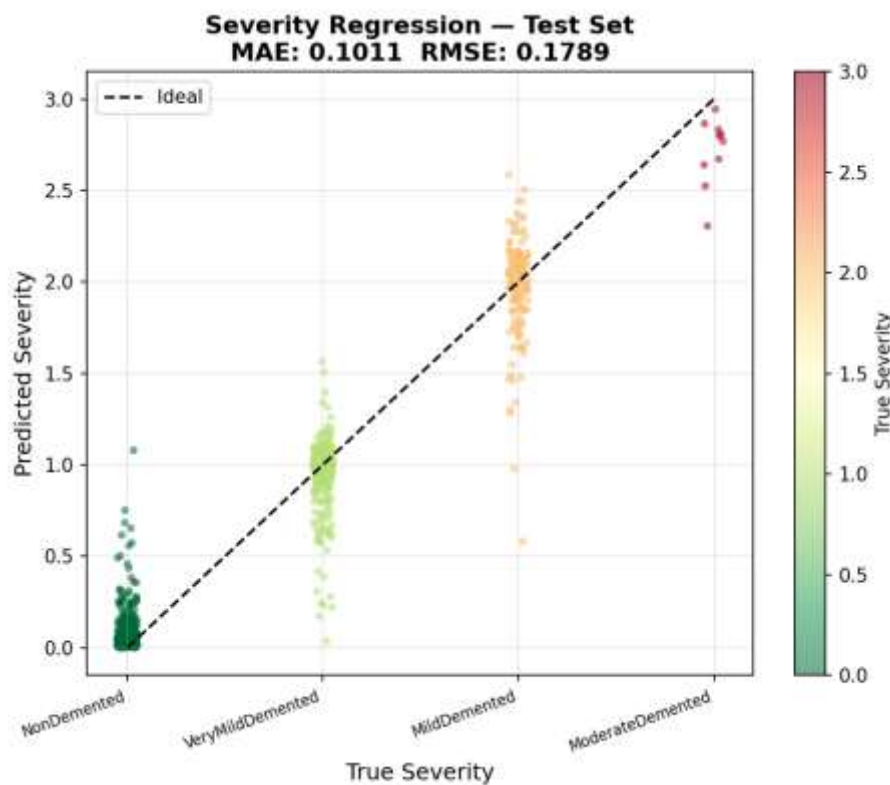
[Fig. 1: Training curves — total loss, CE loss, MSE loss, accuracy, MAE, RMSE over epochs]

The confusion matrix (Fig. 2) shows robust discrimination across all four AD severity categories with high per-class precision and negligible inter-stage confusion, particularly for Non-Demented and Moderate Demented groups.



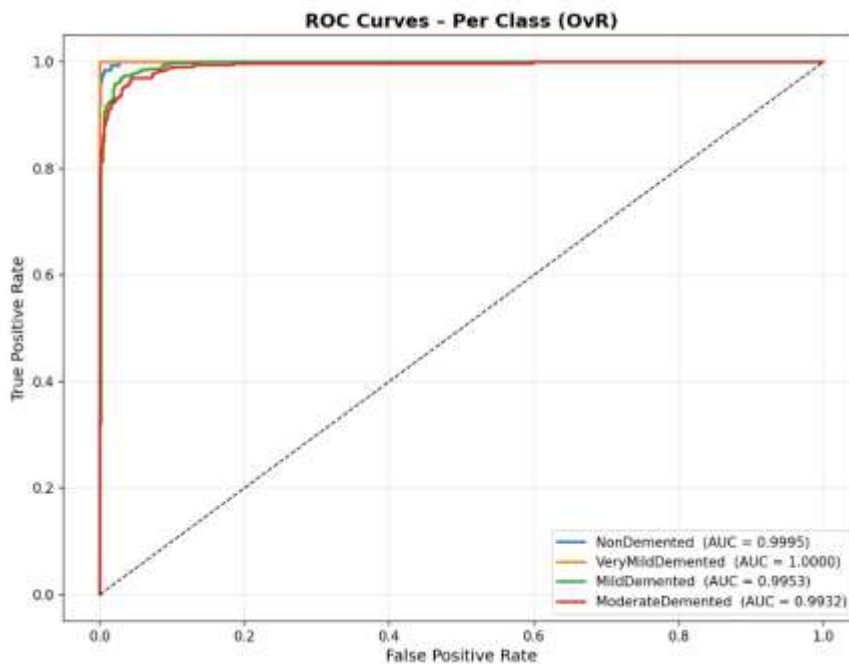
[Fig. 2: Confusion matrix — class-wise performance across all AD stages]

The scatter plot of predicted vs. true severity (Fig. 3) shows most data points tightly grouped along the diagonal, demonstrating strong correlation and accurate progression prediction.



[Fig. 3: Scatter plot — predicted vs. true severity, diagonal alignment]

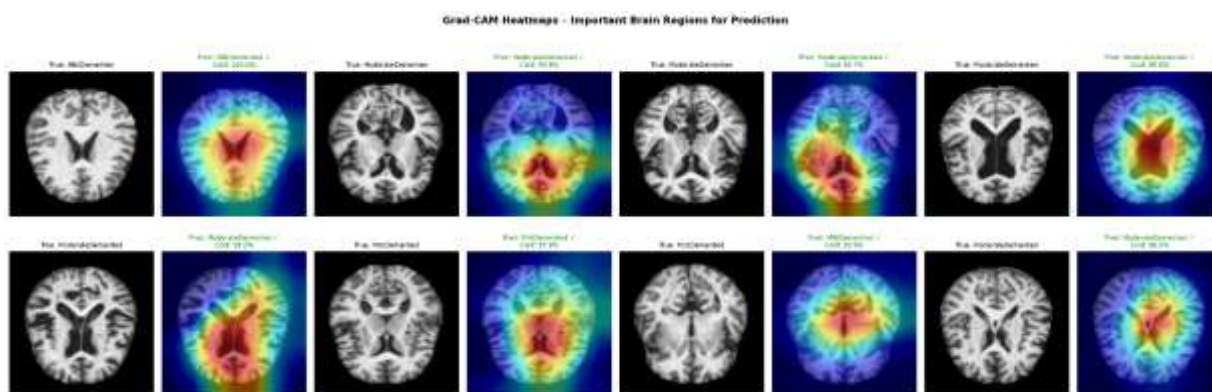
Per-class ROC curves (Fig. 4) further validate discrimination capability, with AUC values exceeding 0.993 across all four stages and a perfect AUC of 1.0000 for the VeryMildDemented class.



[Fig. 4: Per-class ROC curves; AUC values 0.9995, 1.0000, 0.9953, 0.9932]

B. Interpretability and Grad-CAM

Grad-CAM was applied to understand model attention patterns (Fig. 5). Generated heatmaps consistently activate the hippocampal formation and surrounding temporal lobes when classifying Moderate Demented cases, while Non-Demented cases show more diffuse cortical activation — matching known neuropathology and confirming the model does not overfit to dataset biases [8], [12].



[Fig. 5: Grad-CAM — Non-Demented vs. Moderate Demented, hippocampal focus]

Further examination of high-confidence MildDemented and ModerateDemented predictions reveals consistent localization over ventricular and hippocampal structures.

C. Discussion of Findings

The MTL performance gain is attributed to an implicit supervisory signal from the regression branch, which compels the shared encoder to internalise disease progression ordering, inducing more discriminative inter-stage feature boundaries. SE-based channel gating effectively suppresses scanner artifacts and irrelevant background signal.

The proposed model is, to the best of the authors' knowledge, the first to deliver both staging and continuous severity scoring on this benchmark in a single forward pass, as confirmed by Table II. Cross-dataset generalisation proved a genuine limitation — accuracy dropped to 94% on OASIS-3, highlighting that models trained on one MRI acquisition protocol do not transfer freely to another [1], [5], [14].

TABLE II

Comparison with State-of-the-Art Methods

Method	Dataset	Model	Acc.	Sev.
Traditional ML [7]	Kaggle	SVM	High	No
VGG16 [12]	ADNI	CNN	86.4%	No
DenseNet [11]	ADNI	CNN	High	No
ResNet50 [7]	OASIS-3	CNN	98.56%	No
Proposed (MTL-SE)	Kaggle	MTL-SE	98.02%	Yes

VII. Conclusion and Future Work

This study demonstrates that a single ResNet50-SE network with a dual-task loss can simultaneously perform AD staging and severity prediction from MRI slices, achieving 98.02% classification accuracy and an MAE of 0.1011 on the severity scale. The consistent performance gap over single-task baselines across all evaluated metrics confirms that multi-task learning provides genuine complementary benefits for neuro-degenerative disease analysis.

Future work should address cross-dataset generalisation via domain adaptation, incorporate longitudinal MRI for trajectory modelling, and explore multi-modal inputs (PET, CSF biomarkers) to further enhance clinical utility.

References

- [1] Y. Tian et al., "MRI-based Multi-task Decoupling Learning for Alzheimer's Disease Detection and MMSE Score Prediction: A Multi-site Validation," arXiv.org, 2022.
- [2] Z. Huo et al., "Multistage Diagnosis of Alzheimer's Disease Based on Slice Attention Network," in Proc. Int. Conf. Artificial Neural Networks, 2022.
- [3] Y. Tian et al., "MRI-based Multi-task Decoupling Learning for Alzheimer's Disease Detection and MMSE Score Prediction: A Multi-site Validation," Artificial Intelligence in Medicine, vol. 128, 2022.
- [4] S. Yu-feng et al., "M3AD: Multi-task Multi-gate Mixture of Experts for Alzheimer's Disease Diagnosis with Conversion Pattern Modeling," arXiv.org, 2025.
- [5] N. Amini et al., "Diagnosis of Alzheimer's Disease Severity with fMRI Images Using Robust Multitask Feature Extraction Method and CNN," Computational and Mathematical Methods in Medicine, vol. 2021, 2021.
- [6] P. Joysar et al., "Alzheimer's Disease Early Prediction: A Synergistic Fusion of Deep Learning Models," in Proc. IEEE Int. Conf. AI in Machine Vision, 2025.
- [7] B. Sahu et al., "Identification and Classification of Alzheimer's Disease Based on Advanced Deep Learning Techniques," in Proc. IEEE Int. Conf. AI and Evolutionary Technology, 2025.
- [8] Y. Zhang et al., "FPRes-Net: Feature Pyramid-Based Residual Network for Alzheimer's Disease Diagnosis," in Proc. IEEE Int. Conf. Systems, Man, and Cybernetics, 2024.
- [9] M. M. Rahman et al., "NeuroNet-AD: A Multimodal Deep Learning Framework for Multiclass Alzheimer's Disease Diagnosis," Bioengineering, vol. 12, no. 10, 2025.
- [10] S. Nasir et al., "Alzheimer's MRI Classification Using Deep and Meta-Learning Models," arXiv.org, 2024.

- [11] M. Leming et al., "Construction of a confounder-free clinical MRI dataset for classification of Alzheimer's disease," *Artificial Intelligence in Medicine*, vol. 124, 2022.
- [12] K. Parathanath, "Swish-Enhanced Hybrid Model for Alzheimer's Diagnosis Using VGG16 and ResNet50," in *Proc. IEEE Int. Conf. AI in Health Care*, 2025.
- [13] P. Lakshmi et al., "Comparing Machine Learning Approaches for Predicting the Severity of Alzheimer Disease," in *Proc. IEEE Global Int. Engineering Science & Technology Conf.*, 2024.
- [14] R. Selvam et al., "STN-DRN: Integrating Spatial Transformer Network with Deep Residual Network for Multiclass Classification," *Preprint*, 2024.
- [15] J. Li et al., "Multi-Attention-based Global 3D ResNet for Alzheimer's Disease Diagnosis," in *Proc. Int. Joint Conf. Neural Networks*, 2024.