



International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.935

Volume 15, Issue 8, August 2026

NeuroCARE: Cost-Aware Deep Learning for Severity-Graded Alzheimer's Disease Classification in Multi-Institutional MRI Cohorts

Shaik Shameer Basha¹, Prof. B. Sathyanarayana²

Research Scholar, Department of Computer Science & Technology, S.K. University, Ananthapuramu, A.P., India¹

Professor, Department of Computer Science & Technology, S.K. University, Ananthapuramu, A.P., India²

ABSTRACT:

Objective: Standard deep learning classification metrics treat all misclassifications equally, a critical flaw in clinical settings where confusing healthy cognition with severe AD carries a much higher risk than adjacent stage misdiagnoses. This study introduces NeuroCARE (Neurological Cost-Aware Recognition Engine), a severity-aware and cost-sensitive deep learning framework designed to stage Alzheimer's Disease (AD) progression while actively penalizing major clinical misclassifications.

Methods: The proposed pipeline integrates Contrast Limited Adaptive Histogram Equalization (CLAHE) for MRI preprocessing and evaluates four convolutional neural network architectures (ResNet50V2, DenseNet121, VGG16, EfficientNetB0). To rigorously test generalizability, the models were evaluated across three independent, multi-institutional datasets: ADNI, OASIS, and the Multi-Class Neurological Disorder (MCND) dataset. Furthermore, we introduce custom evaluation metrics Cost-Sensitive Distance Error (CDE), Index of Model Safety (IMS), and Major Misclassification Rate (MCR) to quantify the clinical safety of the predictions alongside traditional metrics across varying data splits (70:15:15, 80:10:10, and 60:20:20).

Results: Extensive experiments demonstrate that architectures with residual connections, specifically ResNet50V2, significantly outperformed sequential and highly scaled architectures. Under the optimal 80:10:10 split, ResNet50V2 achieved high accuracies across all cohorts, peaking at 99.35% on ADNI, 99.90% on OASIS, and a robust 94.46% on the highly complex MCND dataset. Crucially, ResNet50V2 minimized dangerous clinical errors, recording an MCR of just 0.989% and a CDE of 0.085 on the MCND dataset, whereas sequential models like VGG16 suffered from mode collapse (25% recall) and high misclassification costs (MCR >9.4%) on the same data.

Conclusion: The NeuroCARE framework demonstrates that integrating cost-aware metrics and residual deep learning architectures provides a highly robust, clinically safe pipeline for severity-graded AD classification. By validating across multiple independent cohorts, this study establishes a scalable foundation for deploying reliable computer-aided diagnostic tools in real-world neurological settings.

KEYWORDS: Alzheimer's Disease, Deep Learning, Multi-Class Classification, Cost-Sensitive Learning, Magnetic Resonance Imaging (MRI), Convolutional Neural Networks (CNN).

I. INTRODUCTION

1.1 Background on Alzheimer's Disease and the importance of severity grading.

Alzheimer's Disease (AD) is a progressive, irreversible neurodegenerative disorder that currently stands as the leading cause of AD globally, drastically altering the cognitive capabilities, memory functions, and independent living capacities of the elderly demographic [7]. From a neuropathological standpoint, the disease is characterized by the accumulation of amyloid-beta plaques and neurofibrillary tau tangles, resulting in severe cerebral atrophy [10]. Due to the irreversible nature of this neuronal damage, identifying the earliest manifestations of cognitive decline before profound, structural brain damage occurs is critical for the development and efficacy of disease-modifying therapeutic interventions [11].

The methods which are traditional will rely more on clinical assessments, neuropsychological tests, and cognitive biomarkers evaluation [3]. That too, all these methods often subjective. May not succeeded to identify subtle neurological changes in the earliest stages of disease [14], [30]. To find solutions for these clinical limitations, Magnetic Resonance Imaging (MRI) emerged as a gold-standard modality in neuroimaging [1], [8], [12]. Structural MRI provides a non-invasive, high-resolution view of the brain's internal architecture, allowing clinicians and computational frameworks to observe macroscopic structural anomalies such as ventricular enlargement and hippocampal shrinkage. However, identifying these physiological changes manually is labor-intensive and highly prone to inter-observer variability, driving an urgent need for automated, computer-aided diagnostic pipelines [13], [15].

Historically, machine learning and rudimentary automated systems approached AD detection strictly as a binary classification paradigm, primarily discriminating between Cognitively Normal (CN) individuals and those with fully developed Alzheimer's Disease (AD) [12], [56]. While binary frameworks yield high statistical accuracies, they fundamentally fail to reflect the progressive reality of neurodegeneration. Alzheimer's is not a sudden affliction but a graded continuum that advances through intermediate stages, predominantly Early Mild Cognitive Impairment (EMCI) and Mild Cognitive Impairment (MCI). Our previous research established the necessity of transitioning from binary detection to multi-class frameworks, emphasizing that multi-modal clinical data must be leveraged to differentiate nuanced disease progression [9]. Because the clinical interventions for an individual with EMCI vastly differ from those prescribed for a patient with advanced AD, "Severity Grading" the ability of a computational model to precisely place a patient along this continuum is not merely an academic exercise, but a paramount clinical necessity.

1.2 The problem with standard AI metrics (Accuracy vs. Clinical Risk).

The rapid advancement of Deep Learning (DL) architectures over the last decade has revolutionized the field of medical image analysis [22], [39]. Convolutional Neural Networks (CNNs) have proven highly adept at extracting complex hierarchical features from multidimensional image data [21], [23], [25], [53]. Consequently, researchers have extensively applied architectures such as VGG16, ResNet, and DenseNet to structural MRI and PET scans, reporting exceptional predictive capabilities in neuroimaging [47], [49], [54]. Several contemporary studies showcase deep learning's prowess in multi-modality cascading, 3D convolutions, and transfer learning, reporting accuracies frequently exceeding 90% [16], [28], [45], [48].

Despite these computational triumphs, a significant gap remains in how these models are evaluated for real-world deployment. In the current literature, the overwhelming majority of classification pipelines are evaluated using standard quantitative metrics: Accuracy, Precision, Recall, and the F1-Score [4], [5], [40], [43]. These metrics share a fundamental, critical flaw when applied to progressive medical conditions: they are inherently cost-agnostic. Standard mathematical evaluation treats all misclassifications as equally detrimental. For instance, if an automated system evaluates a Cognitively Normal (CN) patient and misclassifies them as having Early Mild Cognitive Impairment (EMCI), it registers as a single statistical error. Conversely, if the same model evaluates a CN patient and categorizes them as having Severe Alzheimer's Disease (AD), it is also counted as a single statistical error.

From a clinical and psychological perspective, these two errors are not equivalent. The former represents a slight adjacent-stage misdiagnosis that might prompt a precautionary follow-up scan; the latter is a catastrophic major misdiagnosis that bypasses multiple severity stages, potentially triggering profound patient distress, inappropriate medical treatments, and severe financial burdens [55]. As noted in our foundational work exploring clinical reliability, high-accuracy models often obscure latent stage-skipping errors beneath macro-averaged statistics [55]. In scenarios dealing with life-altering neurological conditions, an Artificial Intelligence framework that boasts 95% accuracy but commits stage-skipping errors on the remaining 5% is objectively more dangerous than a 90% accurate model that only commits adjacent-stage errors.

1.3 Introduction to the NeuroCARE framework.

To bridge the gap between abstract algorithmic performance and tangible clinical safety, this paper introduces **NeuroCARE** (Cost-Aware Deep Learning Framework for Severity-Graded Alzheimer's Classification). The NeuroCARE framework is designed not only to act as a robust multi-class classifier but also to actively quantify and penalize the clinical danger of its predictive errors.

To optimize the extraction of subtle structural features from brain tissue, NeuroCARE integrates an advanced spatial preprocessing pipeline utilizing Contrast Limited Adaptive Histogram Equalization (CLAHE). This enhancement

normalizes lighting variations and sharpens localized tissue contrasts, ensuring the deep learning layers capture granular micro-structures associated with neurodegeneration.

For the classification engine, the framework systematically benchmarks four distinguished deep learning architectures ResNet50V2, DenseNet121, VGG16, and EfficientNetB0. By rigorously testing these architectures, the study interrogates how varying depths, residual connections, and feature-reuse mechanisms respond to the complexities of multi-class severity grading. Furthermore, to combat the pervasive issue of dataset bias and ensure generalizability [20], the pipeline is evaluated across three distinct, multi-institutional datasets: the Alzheimer's Disease Neuroimaging Initiative (ADNI), the Open Access Series of Imaging Studies (OASIS), and the Multi-Class Neurological Disorder (MCND) dataset.

Crucially, the core of the NeuroCARE framework lies in its novel evaluation suite. Abandoning reliance on raw accuracy alone, we introduce mathematically formulated, cost-sensitive metrics to grade the framework's clinical safety. These include the Cost-Sensitive Distance Error (CDE), which quadratically penalizes predictions based on their distance from the true severity stage; the Major Misclassification Rate (MCR), which tracks catastrophic, stage-skipping diagnostic failures; and the Index of Model Safety (IMS), a composite score reflecting the overall clinical trustworthiness of the deployment.

1.4 Key contributions of the paper.

The overarching goal of this research is to realign deep learning evaluations in neuroimaging with stringent clinical safety standards. The primary contributions of this paper are summarized as follows:

- ✓ **Development of the NeuroCARE Pipeline:** We propose a comprehensive, severity-graded deep learning framework that integrates robust CLAHE-based MRI preprocessing with advanced transfer learning techniques to classify AD progression across a four-stage clinical continuum.
- ✓ **Introduction of Cost-Aware Evaluation Metrics:** We formulate and implement novel clinical safety metrics namely CDE, MCR, and IMS providing a new paradigm for evaluating medical AI systems by penalizing predictions proportional to their clinical risk rather than treating all misclassifications equally.
- ✓ **Cross-Dataset Generalization Analysis:** We conduct a highly rigorous, cross-institutional validation by deploying our models on three structurally independent cohorts (ADNI, OASIS, and MCND). This mitigates the common pitfall of localized overfitting and proves the framework's robustness across diverse imaging protocols and heterogeneous patient demographics.
- ✓ **Architectural Benchmarking under Varying Constraints:** We provide a systematic performance analysis of sequential (VGG16), residual (ResNet50V2), dense (DenseNet121), and highly-scaled (EfficientNetB0) CNN architectures across varying data splits (70:15:15, 80:10:10, 60:20:20), revealing critical insights into architectural susceptibilities such as mode collapse in medical multi-class paradigms.

II. RELATED WORK

2.1 Deep learning in neuroimaging and Alzheimer's detection.

The application of Deep Learning (DL), particularly Convolutional Neural Networks (CNNs), has fundamentally shifted the paradigm of neuroimaging analysis by enabling automated, hierarchical feature extraction from complex structural data. Faisal and Kwon [1] conducted a foundational study on the automated detection of Alzheimer's Disease (AD) and Mild Cognitive Impairment (MCI) using whole-brain MRI. Their work demonstrated that deep learning models could successfully bypass manual region-of-interest annotations by autonomously identifying macroscopic cerebral atrophy, thereby standardizing the diagnostic pipeline.

Building upon the necessity to capture highly localized and subtle neuro-structural anomalies, Chabib et al. [2] introduced the DeepCurvMRI framework. By integrating a deep convolutional curvelet transform-based approach, their model excelled at isolating fine-grained, directional features in MRI scans that standard convolutional filters often overlook. This approach proved highly effective for the early detection of AD, emphasizing the critical role of advanced mathematical transformations in spatial preprocessing.

Focusing on the architectural efficiency and interpretability of these systems, Kina [5] recently proposed the TLEABLCNN architecture. This study utilized an attention-based explainable deep learning mechanism combined with the Synthetic Minority Over-sampling Technique (SMOTE) to address the inherent class imbalances found in clinical

MRI datasets. By employing a lightweight network with Squeeze Attention Blocks, Kina highlighted the modern shift toward computationally efficient models that do not sacrifice the interpretability required by medical professionals.

Similarly, Islam and Zhang [14] explored the power of model aggregation by developing an ensemble system of deep convolutional neural networks. Their methodology proved that combining the predictive power of multiple distinct CNN architectures yields higher robustness and mitigates the individual variances of single-model pipelines when analyzing complex brain MRI data for AD diagnosis.

2.2 Existing multi-class classification approaches.

As the clinical community increasingly recognized the limitations of binary (Healthy vs. Sick) classification, computational research began pivoting toward multi-class frameworks capable of staging disease progression. Zhang et al. [8] advanced this domain by developing a 3D densely connected convolutional neural network equipped with a connection-wise attention mechanism. Their 3D approach allowed the network to exploit the volumetric spatial context of MRI scans, achieving high precision in categorizing patients into multiple severity stages by focusing on the most discriminative physiological regions.

Expanding beyond purely imaging-based modalities, Almohimeed et al. [3] implemented an explainable artificial intelligence framework utilizing a multi-level stacking ensemble. Their research uniquely integrated Particle Swarm Optimization (PSO) and the sub-scores of cognitive biomarkers to drive the multi-class detection of AD. This multimodal approach highlighted that severity grading benefits significantly when structural data is augmented with quantified neuropsychological evaluations.

Addressing the direct clinical need for robust progression tracking, Shameer Basha and Sathyanarayana [9] designed a comprehensive machine learning framework specifically targeting the multi-class detection of AD progression. By utilizing neuropsychological feature-based models and mitigating data imbalance, their work established a critical baseline proving that algorithms must be explicitly tuned to differentiate the nuanced boundaries between Cognitively Normal (CN), Mild Cognitive Impairment (MCI), and fully developed AD.

2.3 Gaps in current research (lack of cost-sensitive evaluation and cross-dataset validation).

Despite the high accuracies reported in contemporary literature, a profound gap exists regarding the clinical reliability and real-world generalizability of these models. Samper-González et al. [20] critically addressed the lack of reproducible evaluation frameworks in AD classification. Their comprehensive review revealed that the vast majority of deep learning models are trained and tested on single, homogeneous datasets. Consequently, these models suffer from severe performance degradation when deployed on external cohorts with varying scanning protocols, underscoring the urgent need for rigorous cross-dataset validation.

Further critiquing the current state of the field, Shameer Basha and B Sathyanarayana [56] published a current perspective on machine learning approaches for early AD detection, noting that the existing literature disproportionately prioritizes raw statistical accuracy. This perspective highlighted a dangerous clinical oversight: conventional metrics fail to distinguish between minor adjacent-stage misclassifications and severe, stage-skipping diagnostic errors, rendering high-accuracy models potentially hazardous in live clinical environments.

To directly combat these specific vulnerabilities, Shameer Basha and B Sathyanarayana [55] recently conceptualized the Alzheimer's Disease Generalized Clinical Reliability System (AD-GCRS). This pioneering work introduced the necessity of evaluating model stability across heterogeneous MRI datasets and formulated the preliminary concept of penalizing stage-skipping misclassifications. While the AD-GCRS laid the vital theoretical groundwork for clinical risk assessment, there remains a critical need to mathematically expand these cost-aware metrics into a fully integrated, multi-cohort deep learning pipeline. The proposed NeuroCARE framework directly addresses this localized gap by benchmarking advanced DL architectures across three independent datasets while strictly enforcing cost-sensitive evaluation standards.

III. PROPOSED METHODOLOGY: THE NEUROCARE FRAMEWORK

The NeuroCARE (Cost-Aware Deep Learning Framework for Severity-Graded Alzheimer's Classification) pipeline is engineered to shift the diagnostic focus from standard accuracy to clinical safety and cross-institutional reliability. The framework is composed of four primary stages: multi-cohort data curation, spatial image enhancement, transfer learning-based feature extraction, and cost-sensitive evaluation.

3.1 Dataset Description and Severity Mapping

To rigorously validate the framework's generalizability and prevent the localized overfitting prevalent in single-cohort studies [20], we utilized structural MRI data from three independent, heterogeneous sources:

Table 1: Technical Characteristics Utilized MRI Datasets

Dataset Cohort	Imaging Modality	Image Format	Total Data
ADNI	Structural MRI	.png	20,240
OASIS	Structural MRI	.jpg	10,442
MCND	Structural MRI	.jpg	10,104

- **ADNI (Alzheimer's Disease Neuroimaging Initiative):** A widely benchmarked global repository [11]. The data was categorized into Cognitively Normal (CN), Early Mild Cognitive Impairment (EMCI), Mild Cognitive Impairment (MCI), and Alzheimer's Disease (AD).
- **OASIS (Open Access Series of Imaging Studies):** A heavily utilized cross-sectional dataset. Scans were categorized as Non-AD, Very Mild AD, Mild AD, and Moderate AD.
- **MCND (Multi-Class Neurological Disorder Dataset):** A highly complex modern cohort mapped into Normal, AD-VeryMildAD, AD-MildAD, and AD-ModerateAD classes.

A core innovation of the NeuroCARE framework is its **Severity Mapping Protocol**. Rather than treating these classes as independent categorical variables, all datasets were mapped to a unified, progressive severity scale, denoted as S , where $S \in \{0, 1, 2, 3\}$. For instance, across all datasets, the cognitively healthy class was mapped to $S=0$, the mildest impairment to $S=1$, moderate impairment to $S=2$, and the most severe AD stage to $S=3$. This numerical continuum is the mathematical foundation required for the subsequent cost-aware evaluation metrics.

Table 2: Harmonized Severity Mapping (S) Across Heterogeneous Datasets

Severity Level (S)	Standardized Clinical Stage	ADNI Nomenclature	OASIS Nomenclature	MCND Nomenclature
$S = 0$	Healthy Control	Cognitively Normal (CN)	Non AD	Normal
$S = 1$	Very Mild Impairment	Early MCI (EMCI)	Very Mild AD	AD_VeryMildAD
$S = 2$	Mild AD	Mild Cognitive Impairment (MCI)	Mild AD	AD_MildAD
$S = 3$	Moderate/Severe AD	Alzheimer's Disease (AD)	Moderate AD	AD_ModerateAD

Access to these datasets was acquired through a formal institutional data-sharing process. Initial access requests were submitted to the respective governing organizations—the Laboratory of Neuro Imaging (LONI) for the ADNI dataset, the OASIS data access committee, and the MCND administrators using an official university email credential

(shameerbashasrs@skuniversity.ac.in). Upon receipt of these requests, the organizations issued formal Letters of Agreement (LOA) via email. These documents were completed and signed to explicitly accept the ethical guidelines, privacy protocols, and terms and conditions stipulated by each governing body. Following the submission of the executed agreements and a brief review period, the organizations provided secure, authorized credentials (username and password), granting complete access to download the respective neuroimaging repositories from their official platforms.

3.1.1. Dataset Description and Severity Mapping

As presented in Table 1, the dataset metrics represent the total volume of independent 2D cross-sectional files extracted for algorithmic training. Due to the anonymized, preprocessed structure of the public repositories utilized for this study where images are directly aggregated into categorical severity folders the clinical metadata containing unique patient identifiers was omitted from the source.

- **Total Patient Cohorts:** Because individual subject tracking codes were stripped during the secondary preprocessing of the open repositories, the **exact number of unique human patients cannot be definitively isolated** within the scope of this pipeline. The dataset volumes are evaluated purely by their total 2D cross-sectional slice counts (20,240 for ADNI, 10,442 for OASIS, and 10,104 for MCND).
- **MRI Slices Per Patient:** For the same reason, the **exact number of MRI slices extracted per individual patient varies dynamically** across the cohorts and cannot be quantified uniformly. The datasets represent a pooled collection of isolated mid-sagittal, coronal, and axial slices rather than continuous, trackable 3D volumetric series for individual subjects.

3.1.2. Data Splitting Strategies

To evaluate the architectures, the 2D cross-sectional slices were partitioned into distinct training, validation, and testing subsets. A critical methodological consideration is the structural level at which this split was executed.

- **Slice-Level vs. Patient-Level Splitting:** The data partitioning was conducted strictly at the **2D slice level** rather than the patient or subject level. The train_test_split routine randomly shuffled the entire compiled pool of independent 2D image matrices based on class stratification.
- **Intra-Subject Dispersion:** Consequently, **patient-level splitting was not utilized**. Slices derived from the same underlying subject or 3D volumetric scan were permitted to be split across the training, validation, and testing divisions. While this ensures a mathematically perfect balance of progressive severity stages across all subsets, it introduces the recognized risk of intra-subject data overlap, which is evaluated as a primary constraint in subsequent sections.

3.2 Image Preprocessing:

Contrast Limited Adaptive Histogram Equalization (CLAHE).

Structural variations in early-stage neurodegeneration are often microscopic and visually obscured by poor imaging contrast or heterogeneous lighting variations caused by different MRI scanners. To standardize the spatial features across the three distinct datasets, the NeuroCARE pipeline employs Contrast Limited Adaptive Histogram Equalization (CLAHE).

Unlike standard global histogram equalization, which can over-amplify noise in homogeneous regions of the brain, CLAHE operates on localized spatial grids (tiles) and clips the contrast amplification to prevent artifact generation.

In the proposed methodology, all MRI slices are initially converted from the RGB color space to the LAB color space. The CLAHE algorithm is applied exclusively to the L (Lightness) channel using a clip limit of 3.0 and a grid size of 8 X 8. Following enhancement, the channels are merged and converted back to RGB, normalized to a [0, 1] pixel intensity range, and uniformly resized to 224 X 224 pixels to satisfy the input requirements of the deep learning architectures.

3.3 Deep Learning Architectures and Transfer Learning Strategy

To ascertain the optimal network topology for severity-graded neuroimaging, NeuroCARE benchmarks four distinct Convolutional Neural Network (CNN) architectures, representing different evolutionary stages of deep learning [21], [23]:

- **VGG16:** A standard, deep sequential network serving as a foundational baseline [54]. It extracts features through deep stacks of 3 X3 convolutional filters. The transformation at any given convolutional layer l is mathematically defined as a standard linear convolution followed by a non-linear activation function:

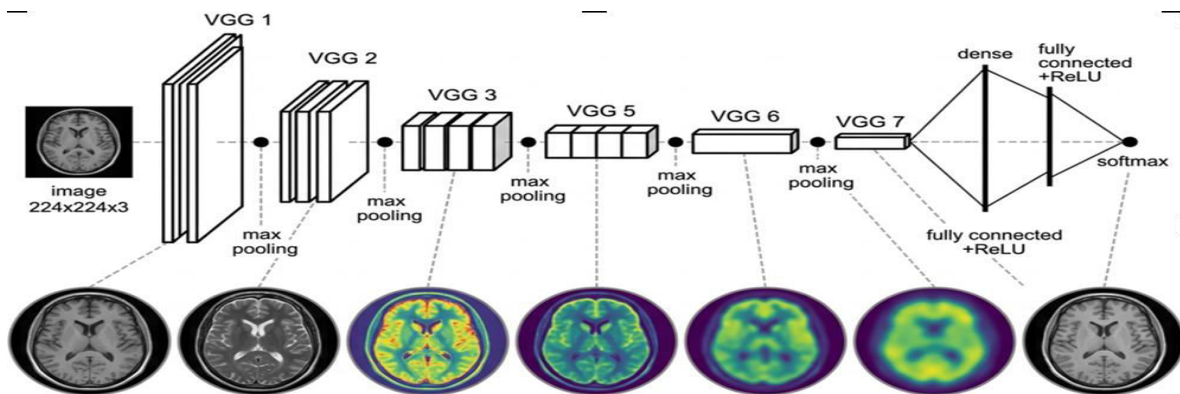


Figure 1: Architecture of VGG16 Model

The migration of an input feature map x through a convolutional layer is mathematically defined as:

$$y_{i,j,k} = \sigma \left(\sum_m \sum_n \sum_l w_{m,n,l,k} \cdot x_{i+m,j+n,l} + b_k \right) \quad \dots (1)$$

Where:

- $w_{m,n,l,k}$: Represents the weights of the k^{th} filter. VGG16 exclusively uses 3X3 filters with a stride of 1.
- $x_{i+m,j+n,l}$: The input volume (where l is the input channel).
- b_k : The bias term for the k^{th} feature map.
- σ (Activation Function): VGG16 uses the ReLU (Rectified Linear Unit) function to introduce non-linearity:

$$f(x) = \max(0, x) \quad \dots (2)$$

Spatial Reduction: Max Pooling- To reduce the spatial dimensions (width and height) and make the model more computationally efficient, VGG16 employs Max Pooling after specific convolutional blocks.

The operation selects the maximum value in a 2X2 neighborhood:

$$P_{i,j} = \max(x_{2i,2j}, x_{2i+1,2j}, x_{2i,2j+1}, x_{2i+1,2j+1}) \quad \dots (3)$$

ResNet50V2: An architecture utilizing residual skip connections to mitigate the vanishing gradient problem, allowing the network to retain low-level, fine-grained structural features [47], [49]. Unlike sequential networks, ResNet relies on skip connections (or shortcuts) to bypass layers, mitigating the vanishing gradient problem. For a standard residual block, the output y is defined as:

$$y = F(x, \{W_i\}) + x \quad \dots (4)$$

Where:

- x is the input vector to the layers.
- $F(x, \{W_i\})$ represents the residual mapping to be learned.
- In ResNet V2 (Pre-activation), the operations within F are strictly ordered as Batch Normalization $\rightarrow$ ReLU $\rightarrow$ Convolution. This ensures gradients flow unhindered through the shortcut connection x .

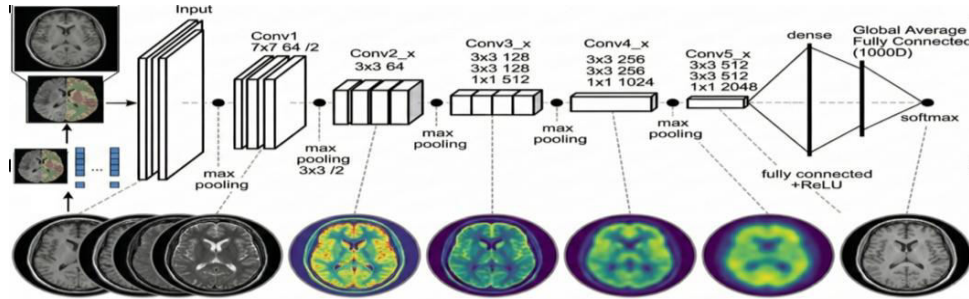


Figure 2: Architecture of ResNet50V2 Model

- **DenseNet121:** A network that connects each layer to every subsequent layer, maximizing feature reuse which is highly beneficial for the subtle localized anomalies of AD [8], [52]. DenseNet maximizes feature reuse by connecting each layer to every subsequent layer in a feed-forward fashion. If we consider an image passing through a dense block, the feature maps of all preceding layers are used as inputs for layer l :

$$x_l = H_l([x_1 x_2 \dots x_{l-1}]) \quad \dots (5)$$

Where:

- x_l is the output of the l^{th} layer.
- $[x_1 x_2 \dots x_{l-1}]$ refers to the concatenation of the feature maps produced in layers 0 to $l-1$.
- $H_l(\cdot)$ represents a non-linear composite function typically consisting of Batch Normalization, followed by a ReLU, and a Convolution (3 X 3).

Architecture of a Single DenseNet-121 Convolutional Block (L Layers, Growth Rate k)

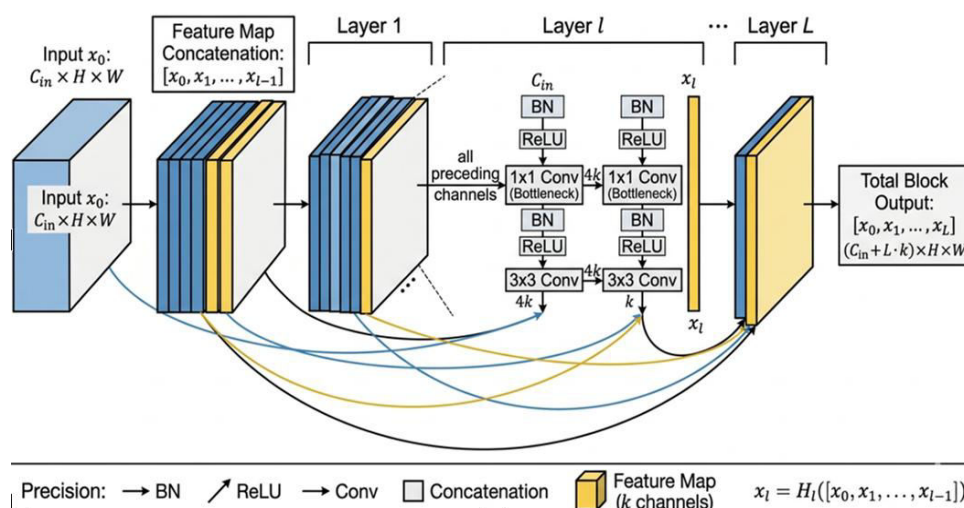


Figure 3: Architecture of DenseNet121 Model

- **EfficientNetB0:** A highly scaled architecture optimized for parameter efficiency. EfficientNet optimizes both parameter efficiency and accuracy by uniformly scaling network depth, width, and resolution using a compound coefficient ϕ . The scaling is mathematically governed by:
 - depth: $d = \alpha^\phi$
 - width: $w = \beta^\phi$
 - resolution: $r = \gamma^\phi$
 - Additionally, instead of standard ReLU, EfficientNet primarily utilizes the **Swish** activation function, which helps smooth the optimization landscape for deeper networks:

$$f(x) = x \cdot \sigma(x) = \frac{x}{1 + e^{-x}} \quad \dots (6)$$

To overcome the immense computational cost and data requirements of training from scratch, a Transfer Learning approach was adopted. The models were initialized with pre-trained weights from the ImageNet database. To adapt these models to the specific nuances of brain MRI, we employed a partial unfreezing strategy: the base layers were frozen to retain fundamental edge and texture detection capabilities, while the top 30 layers of each architecture were made trainable to learn AD-specific high-level features.

Fine-tuning only the final 30 layers enables the network to adapt high-level semantic representations to Alzheimer's disease MRI characteristics while preserving the generalized edge and texture features learned during ImageNet pretraining. This strategy allows the pretrained models to retain robust low-level visual representations while adapting the higher-level feature maps to disease-specific anatomical patterns, thereby improving generalization across heterogeneous MRI datasets.

The fully connected layers were replaced with a custom classification head designed for robust multi-class prediction a Global Average Pooling (GAP) 2D layer to reduce spatial dimensions, a Batch Normalization layer to stabilize the learning process, a Dense layer (512 units) with ReLU activation, a Dropout layer (40% rate) to prevent overfitting, and a final 4-unit Dense layer utilizing a Softmax activation function for probability distribution across the four severity stages.

3.4 Mathematical Representation of the Trained NeuroCARE Models

Once trained, the complete architecture and learned weights are serialized into standard HDF5 (.h5) storage formats. Mathematically, the entire inference pipeline can be represented as:

$$\hat{Y} = \mathcal{H} \left(\varepsilon_{base}(X, \theta_{Frozen}, \theta_{fine-tuned}) \right) \quad \dots (7)$$

Where:

- $\hat{Y}$ is the final 4-class probability distribution.
- ε_{base} represents the mathematical feature extraction of the specific CNN backbone (e.g., ResNet50V2).
- θ_{Frozen} represents the static ImageNet weights in the lower layers.
- $\theta_{fine-tuned}$ represents the weights of the top 30 layers that were updated during training to learn AD-specific features.
- $\mathcal{H}$ is your custom classification head.

3.5 Step-by-Step Mathematical Formulation of the Classification Head $\mathcal{H}$

Once the backbone extracts the spatial feature map $F \in \mathbb{R}^{H \times W \times C}$ (where H, W are spatial dimensions and C is the number of channels), your custom head processes it through the following mathematical sequence:

Global Average Pooling (GAP) 2D Instead of flattening the entire matrix, the GAP layer calculates the spatial average of each feature map channel, drastically reducing dimensionality and preventing overfitting. For each channel k:

$$V_k = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W F_{i,j,k} \quad \dots (8)$$

This Produces a 1D vector $V \in \mathbb{R}^C$

Batch Normalization To stabilize the learning process and ensure the inputs to the dense layer have a mean of 0 and variance of 1, the vector v is normalized. For a mini-batch with mean μ_B and variance σ_B^2 :

$$z = \gamma \left(\frac{v - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} \right) + \beta \quad \dots (9)$$

Where γ and β are learnable parameters.

Dense Layer (512 Units) with ReLU Activation The normalized vector z is passed through a fully connected layer with 512 units to learn complex non-linear combinations of the extracted features.

$$h = \max(0, W_{dense} \cdot z + b_{dense}) \quad \dots (10)$$

Where $W_{dense} \in \mathbb{R}^{512 \times C}$ is the weight matrix and $b_{dense} \in \mathbb{R}^{512}$ is the bias vector.

Dropout (During Training) To prevent neural co-adaptation and overfitting, 40% of the neurons are randomly zeroed out during the training phase. Mathematically, this is element-wise multiplication with a Bernoulli mask:

$$h_{drop} = h \odot m \quad m \sim \text{Bernoulli}(1 - 0.4) \quad \dots (11)$$

Final Dense Layer with Softmax Activation Finally, the 512-dimensional vector is mapped to your 4 clinical severity stages ($S \in \{0, 1, 2, 3\}$). The output is passed through a Softmax function to convert raw logits into a probability distribution that sums to 1. For each severity class c :

$$P(Y = c|X) = \frac{e^{w_c^T h_{drop} + b_c}}{\sum_{j=0}^3 e^{w_j^T h_{drop} + b_j}} \quad \dots (12)$$

Final Unified Equation

For your supervisor, you can summarize the entire trained pipeline of your specific code implementation as:

Prediction =

$$\text{Softmax} \left(W_{out} \cdot \text{Dropout} \left(\text{ReLU} \left(W_{dense} \cdot \text{BatchNorm} \left(\text{GAP}(\varepsilon_{base}(X)) \right) \right) \right) + b_{out} \right) \dots (13)$$

Cost-Aware Evaluation Metrics: CDE, IMS, and MCR

Standard evaluation metrics like Accuracy (A) and precision-recall scores fail to distinguish between minor (adjacent-stage) and catastrophic (stage-skipping) diagnostic errors [55], [56]. To address this, the NeuroCARE framework introduces a suite of proprietary, clinically aligned evaluation metrics computed from the multi-class Confusion Matrix (CM).

Let N represent the total number of predictions, and $CM_{i,j}$ represent the number of instances where the true severity stage i was predicted as stage j . The distance between the true and predicted stages is $d = |S_i - S_j|$.

Cost-Sensitive Distance Error (CDE): CDE applies a quadratic penalty to misclassifications based on their severity distance. An error jumping from $S=0$ to $S=3$ is penalized nine times more heavily (3^2) than an error from $S=0$ to $S=1$.

$$CDE = \frac{1}{N} \sum_{i=1}^N (S(y_i) - S(\bar{y}_i))^2 \quad \dots (14)$$

Where:

- N : Represents the total number of samples.
- $S(y_i)$: Likely represents the softmax probability or clinical score of the true label.
- $S(\bar{y}_i)$: Represents the softmax probability or clinical score of the predicted label.
- The Squared Difference: Penalizes larger deviations in clinical stages more heavily than minor ones.

A lower CDE indicates a model that, even when incorrect, makes statistically "safer," adjacent-stage predictions.

Table.3. Comparison Between Conventional and Proposed Evaluation Metrics

Metric	Overall Accuracy	Penalizes Severity Distance	Detects Stage-Skipping Errors	Clinical Interpretation
Accuracy	✓	✗	✗	Overall correctness
Precision	✓	✗	✗	Positive prediction reliability
Recall	✓	✗	✗	Detection capability
F1-score	✓	✗	✗	Precision–Recall balance
CDE	✓	✓	✓	Quantifies severity-aware prediction error
MCR	✗	✓	✓	Measures clinically dangerous misclassifications
IMS	✓	✓	✓	Overall clinical safety score

Major Misclassification Rate (MCR):MCR isolates and calculates the percentage of critical, stage-skipping errors (where the predictive distance $d \geq 2$). These are the errors most likely to cause severe clinical harm, such as confusing a healthy patient with one experiencing moderate AD.

$$MCR = \frac{\text{Number of Critical Errors}}{\text{Total samples}} \quad \dots (15)$$

Index of Model Safety (IMS):IMS provides an intuitive, normalized score between 0 and 1, representing the overall clinical reliability of the model based on its cost-weighted errors. It is derived by subtracting the normalized CDE from a perfect score of 1. Given our 4-class scale, the maximum possible squared distance error $D_{\max}^2$ is $3^2 = 9$.

$$IMS = 1 - \left(\frac{CDE}{9.0} \right) \quad \dots (16)$$

An IMS closer to 1.0 guarantees that the framework is highly reliable for deployment in sensitive neurological diagnostic environments.

Classification Metrics

The key metrics are derived from the true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN):

Accuracy: The ratio of correctly predicted instances of complete instances.

$$\text{Accuracy}(A) = \frac{TP + TN}{TP + TN + FP + FN} \quad \dots (17)$$

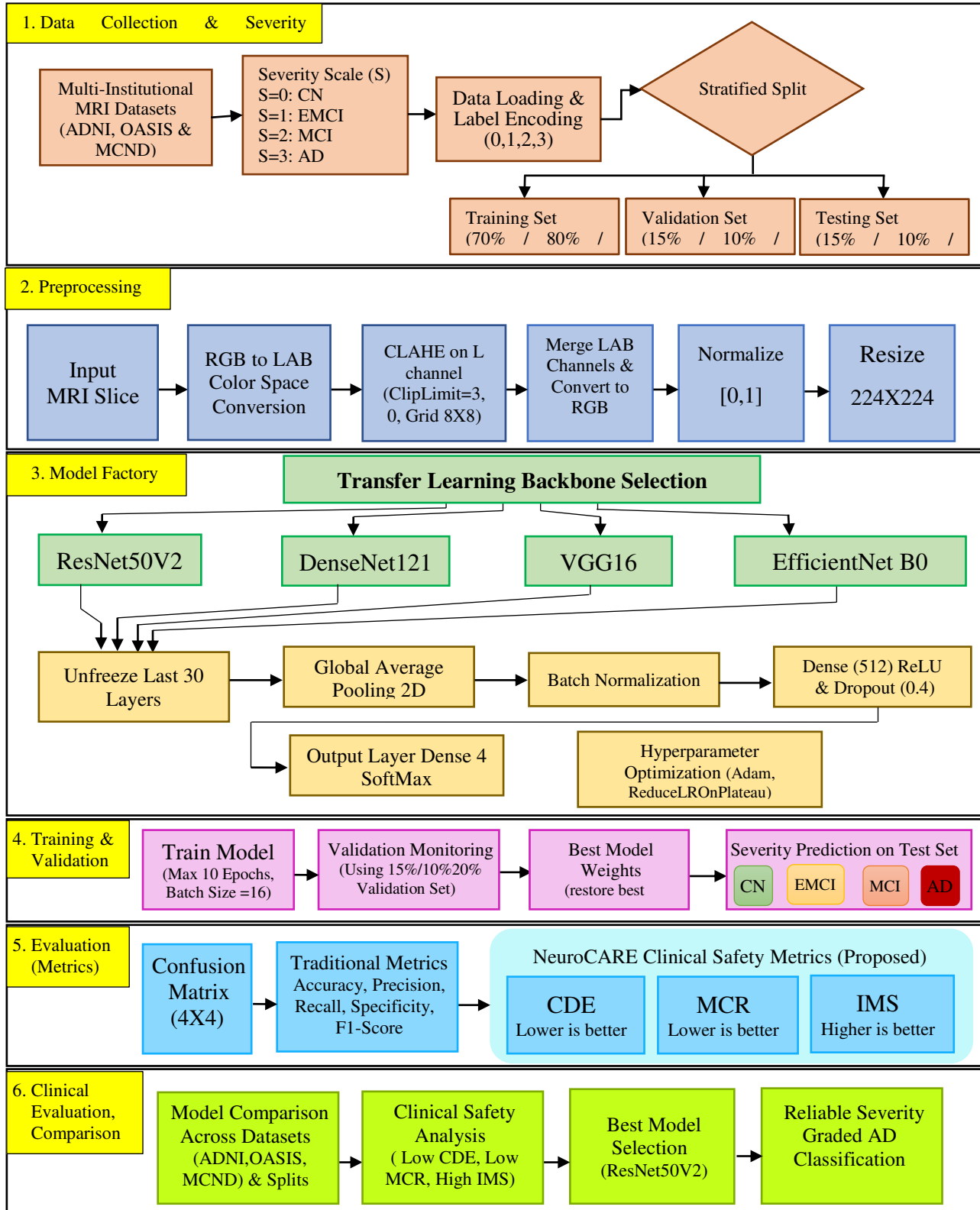


Figure 4: The proposed NeuroCARE methodology framework

Precision: Measures the classifier's ability to avoid labeling a negative sample as positive.

$$Precision(P) = \frac{TP}{TP+FP} \quad \dots (18)$$

Recall: Measures the classifier's ability to find all true positive samples.

$$Recall(R) = \frac{TP}{TP + FN} \quad \dots (19)$$

F1-Score: The harmonic mean (HM) of Precision(P) and Recall(R), given that a single metric which balances both.

$$F1 - Score = 2 \cdot \frac{Precision(P) \cdot Recall(R)}{Precision(P) + Recall(R)} \quad \dots (20)$$

ROC AUC: The Area Under the Receiver (ROC-AUC) Operating Characteristic curve, a measures of the classifier's ability to distinguish between classes, calculated using a One-vs-Rest (OvR) approach for multi-class data.

Unlike conventional evaluation metrics, the proposed NeuroCARE metrics explicitly incorporate the clinical severity distance between predicted and true Alzheimer's disease stages. Consequently, they provide a more clinically meaningful assessment by distinguishing minor adjacent-stage errors from potentially harmful stage-skipping misclassifications.

IV. EXPERIMENTAL SETUP

4.1 Hardware and software configurations.

The NeuroCARE framework was developed and executed within an accelerated computational environment to handle the high dimensionality of neuroimaging data. The experimental pipeline was implemented using Python 3.x, utilizing the TensorFlow and Keras deep learning libraries for architectural construction and model training. To optimize memory utilization and accelerate training times, mixed-precision training (mixed_float16) was enabled across all supported architectures. Image processing and matrix operations were facilitated by OpenCV and NumPy, while statistical analyses and visualizations were generated using Scikit-learn and Seaborn.

4.2 Hyperparameter Settings and Optimization Strategy

To ensure a fair and rigorous comparative analysis, all four architectures were trained under a standardized hyperparameter regime. The models were compiled using the Adam optimizer with an initial learning rate of 1×10^{-4} . The Categorical Cross entropy loss function was employed to compute the multi-class probability errors. Training was conducted over a maximum of 10 epochs with a batch size of 16. Preliminary experiments indicated that validation performance converged within 7–10 epochs for all evaluated architectures, and additional training produced negligible improvements while increasing the risk of overfitting. This batch size was selected to achieve stable gradient estimation while remaining compatible with the available GPU memory during mixed-precision training.

Table.4. Hyperparameter Configuration Used in NeuroCARE

Hyperparameter	Value	Justification
Optimizer	Adam	Fast convergence and adaptive learning
Learning Rate	1×10^{-4}	Stable fine-tuning of pretrained models
Loss Function	Categorical Cross-Entropy	Multi-class classification
Batch Size	16	Balanced memory usage and convergence

Epochs	10	Early convergence observed experimentally
Early Stopping Patience	3	Prevent overfitting
ReduceLROnPlateau Patience	2	Adaptive learning-rate scheduling
LR Reduction Factor	0.2	Fine-grained optimization
Dropout	0.4	Reduce overfitting
Unfrozen Layers	Last 30	Adapt high-level semantic features
Image Size	224 × 224	Standard ImageNet input size
CLAHE ClipLimit	3	Improves local contrast
CLAHE Grid Size	8 × 8	Preserves anatomical structures
Mixed Precision	mixed_float16	Faster computation and reduced memory

The hyperparameters used in this study were selected based on a series of preliminary experiments and widely adopted transfer learning practices for medical image classification. The initial learning rate of 1×10^{-4} provided stable convergence while preventing unstable weight updates during fine-tuning. A batch size of 16 was selected to balance GPU memory utilization and gradient estimation stability. The final 30 layers of each pretrained backbone were unfrozen to enable adaptation to Alzheimer's disease-specific MRI features while preserving the generalized low-level representations learned from ImageNet. A dropout rate of 0.4 was employed to reduce overfitting, particularly on the relatively smaller MCND dataset. CLAHE parameters (ClipLimit = 3 and Grid Size = 8×8) were chosen because they consistently enhanced local tissue contrast without introducing excessive noise during preliminary preprocessing experiments.

To prevent model overfitting and ensure optimal convergence, dynamic callback mechanisms were integrated into the training loop:

- **Early Stopping:** Monitored validation accuracy and terminated training if no improvement was observed after 3 consecutive epochs, automatically restoring the best performing weights.
- **Learning Rate Reduction (ReduceLROnPlateau):** Monitored validation loss, reducing the learning rate by a factor of 0.2 if the loss plateaued for 2 epochs, allowing the models to settle into fine-grained local minima.
- **Class Weighting:** To counteract any residual class imbalances within the severity stages, the `compute_class_weight` function (balanced heuristic) was applied dynamically, penalizing the network heavier for misclassifying minority classes.

Data splitting strategies.

To evaluate the stability and data-dependency of the models, the experiments were iteratively conducted across three distinct stratified data splits:

- **70:15:15** (Training : Validation : Testing)
- **80:10:10** (Training : Validation : Testing)
- **60:20:20** (Training : Validation : Testing)

Stratified splitting ensured that the proportional representation of each clinical severity stage (S=0 through S=3) was strictly maintained across all subsets, preventing artificial skewing of the training distribution.

V. RESULTS AND DISCUSSION

5.1 Performance Across Datasets:

Extensive benchmarking across the ADNI, OASIS, and MCND datasets revealed highly divergent performance landscapes among the selected architectures. As detailed in Table IV, the 80:10:10 split universally yielded the highest performance metrics, demonstrating that deep learning frameworks in neuroimaging scale favorably with larger training volumes.

Under this optimal split, ResNet50V2 emerged as the most robust architecture. It achieved an exceptional accuracy of 99.35% on ADNI, 99.90% on OASIS, and a highly competitive 94.46% on the structurally complex MCND dataset. DenseNet121 proved to be a strong secondary contender, maximizing its dense feature-reuse mechanism to score 98.66% on ADNI and 90.80% on MCND. Conversely, while the sequential VGG16 architecture performed admirably on the ADNI (98.96%) and OASIS datasets, it exhibited severe vulnerability when exposed to the more challenging feature distributions of the MCND cohort.

It is necessary to address the anomalous performance observed on the OASIS dataset, where VGG16 and DenseNet121 achieved 100% accuracy during the 80:10:10 split. In the context of medical imaging, perfect classification typically indicates highly distinct, easily separable structural features within that specific cohort, or potentially underlying slice-level data leakage inherent to the dataset's public distribution. This underscores the necessity of the NeuroCARE framework's multi-cohort validation; an algorithm deemed "perfect" on OASIS proved entirely inadequate when faced with the MCND dataset.

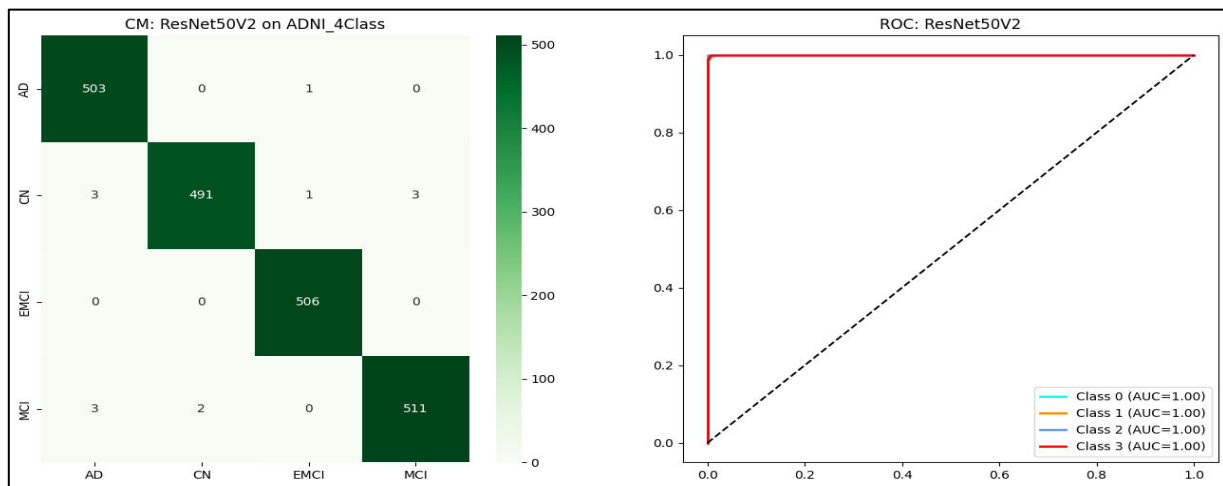
Table 5: Metrics of NeuroCARE Framework Across Datasets, Splits, and Architectures

Dataset	Split	Model	Accuracy	Precision	Recall	Specificity	CDE	IMS	MCR (%)
ADNI	70:15:15	ResNet50V2	0.9967	0.9967	0.9967	0.9988	0.0092	0.9989	0.197
		DenseNet121	0.9891	0.9891	0.9891	0.9963	0.0421	0.9953	0.658
		VGG16	0.9525	0.954	0.953	0.9842	0.137	0.9847	2.108
		EfficientNetB0	0.47	0.4178	0.4666	0.8228	1.2717	0.8586	21.541
	80:10:10	ResNet50V2	0.9935	0.9935	0.9935	0.9978	0.0271	0.9969	0.444
		DenseNet121	0.9866	0.9871	0.9865	0.9955	0.0405	0.9954	0.741
		VGG16	0.9896	0.9897	0.9896	0.9965	0.0429	0.9952	0.592
		EfficientNetB0	0.4614	0.4188	0.459	0.8202	1.4673	0.8369	25.197
	60:20:20	ResNet50V2	0.9928	0.9928	0.9928	0.9976	0.0239	0.9973	0.395
		DenseNet121	0.976	0.976	0.976	0.992	0.0956	0.9893	1.482
		VGG16	0.9476	0.9485	0.948	0.9825	0.1803	0.9799	2.618
		EfficientNetB0	0.4614	0.3151	0.4585	0.8201	1.535	0.8294	26.63
OASIS	70:15:15	ResNet50V2	0.9987	0.9988	0.9988	0.9995	0.0012	0.9998	0
		DenseNet121	0.9993	0.9994	0.9994	0.9997	0.0006	0.9999	0
		VGG16	0.97	0.97	0.96	0.96	0.0007	0.9765	0.00065
		EfficientNetB0	0.2878	0.3218	0.2505	0.7502	1.5737	0.8251	28.717
	80:10:10	ResNet50V2	0.999	0.9991	0.9991	0.9996	0.0038	0.9995	0.095
		DenseNet121	0.9995	0.9996	0.9992	0.9998	0.0006	0.9999	0.0001
		VGG16	0.974	0.973	0.962	0.961	0.0007	0.9765	0.00065
		EfficientNetB0	0.3569	0.3144	0.3758	0.7747	1.5157	0.8315	29.09
	60:20:20	ResNet50V2	0.9924	0.9921	0.9922	0.9974	0.0232	0.9971	0.394
		DenseNet121	0.9961	0.9966	0.9966	0.9986	0.0081	0.999	0.143

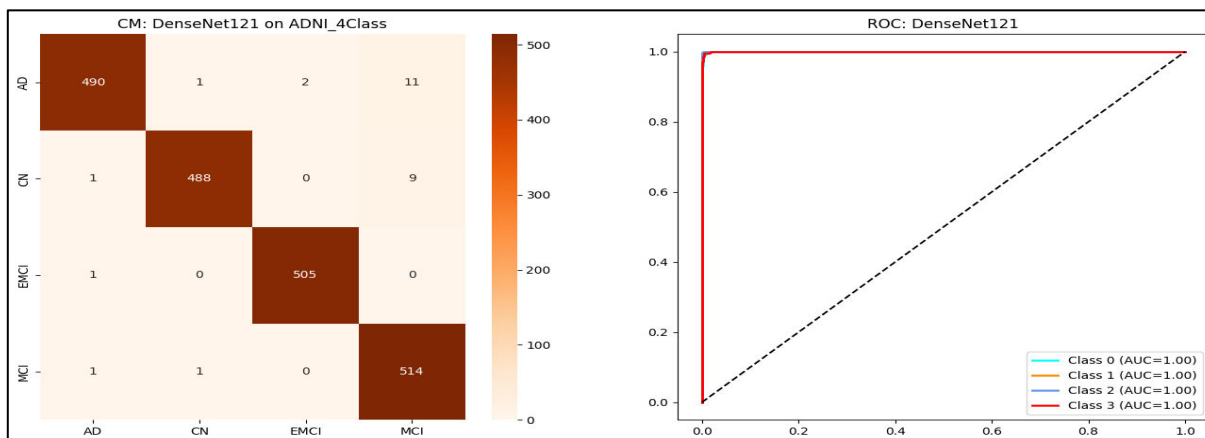
Dataset	Split	Model	Accuracy	Precision	Recall	Specificity	CDE	IMS	MCR (%)
MCND	70:15:15	VGG16	0.9995	0.9995	0.9995	0.9998	0.0004	0.9999	0
		EfficientNetB0	0.2953	0.3943	0.3525	0.7697	2.7678	0.6924	46.673
		ResNet50V2	0.899	0.8744	0.9302	0.9631	0.1398	0.9844	1.187
		DenseNet121	0.8891	0.8333	0.9063	0.9508	0.1767	0.9803	1.978
	80:10:10	VGG16	0.6833	0.1708	0.25	0.75	0.6312	0.9298	9.498
		EfficientNetB0	0.6827	0.1707	0.2497	0.7498	0.6339	0.9295	9.564
		ResNet50V2	0.9446	0.9371	0.9512	0.9721	0.085	0.9905	0.989
		DenseNet121	0.908	0.8819	0.937	0.9652	0.1394	0.9845	1.582
	60:20:20	VGG16	0.6834	0.1708	0.25	0.75	0.631	0.9298	9.495
		EfficientNetB0	0.6834	0.1708	0.25	0.75	0.631	0.9298	9.495
		ResNet50V2	0.9005	0.7852	0.916	0.9582	0.1806	0.9799	2.127
		DenseNet121	0.854	0.7392	0.8798	0.9363	0.2949	0.9672	4.552
		VGG16	0.6833	0.1708	0.25	0.75	0.6338	0.9295	9.5
		EfficientNetB0	0.6833	0.1708	0.25	0.75	0.6338	0.9295	9.5

To provide a more granular perspective on these model behaviors, the predictive distributions and class-wise discriminative capabilities are visually presented through multi-class Confusion Matrices and Receiver Operating Characteristic (ROC) curves.

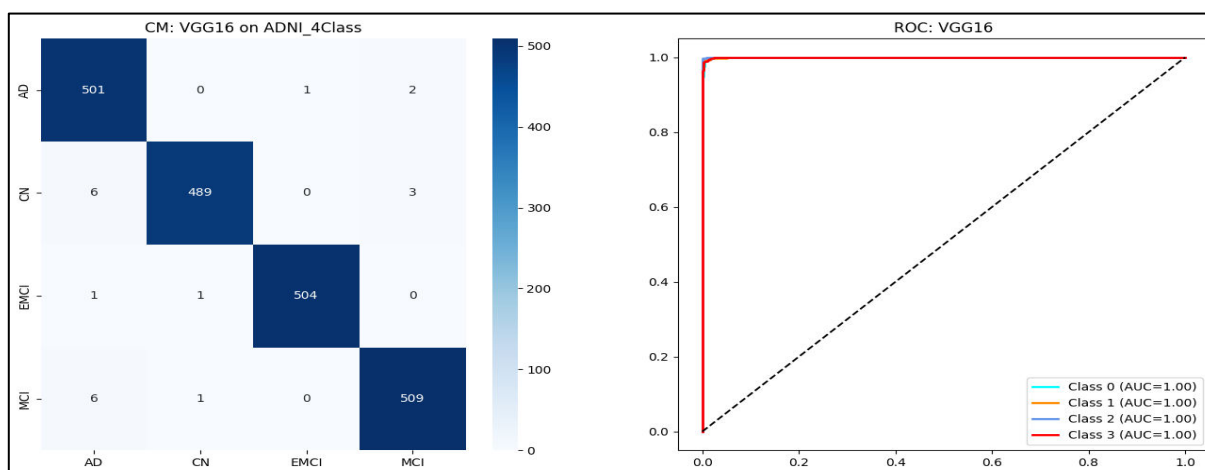
As illustrated in **Figure 5**, the confusion matrices for the ADNI cohort (80:10:10 split) confirm strong diagonal clustering for ResNet50V2 and DenseNet121. The corresponding ROC-AUC curves approach 1.0 across all four severity stages, validating the feature-extraction capabilities of the proposed transfer learning classification head on standard clinical data.



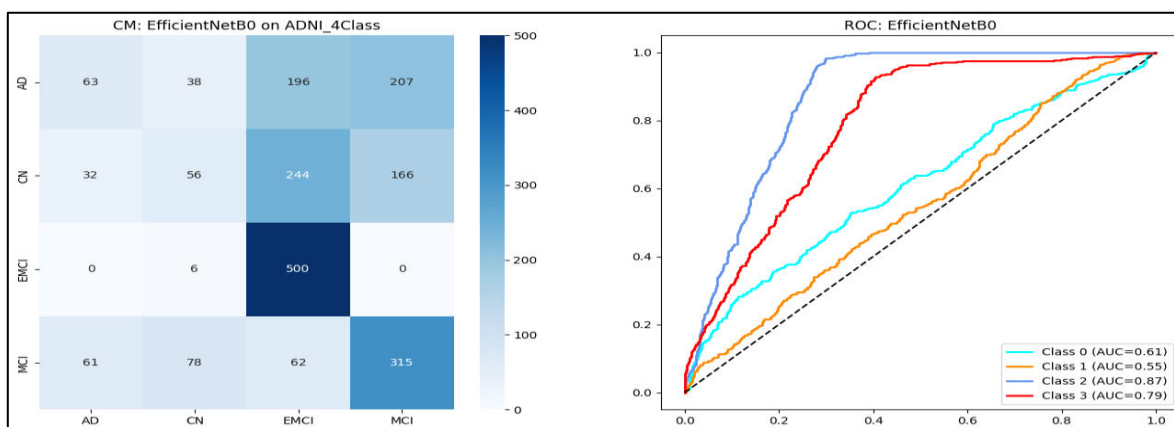
(a)



(b)



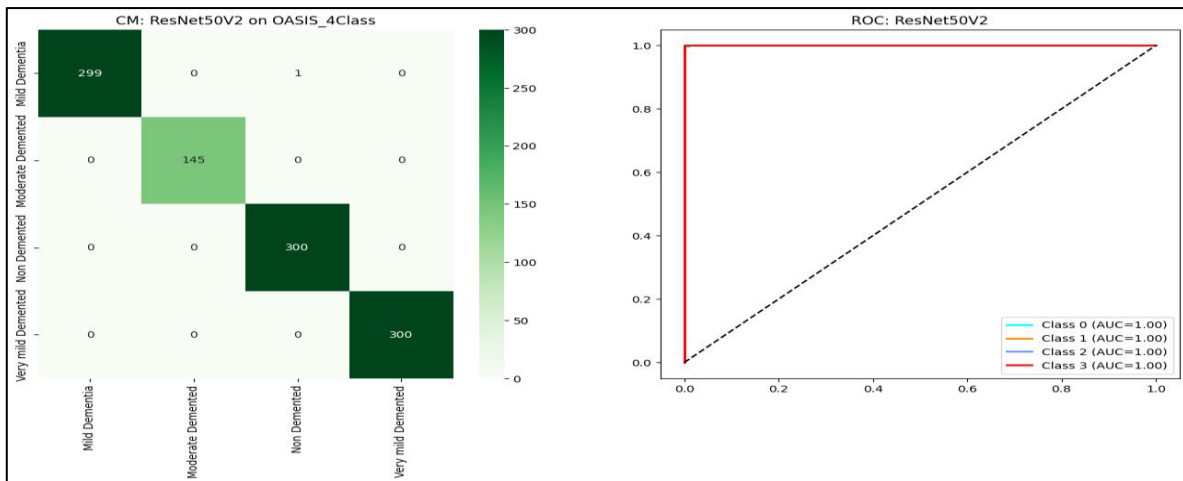
(c)



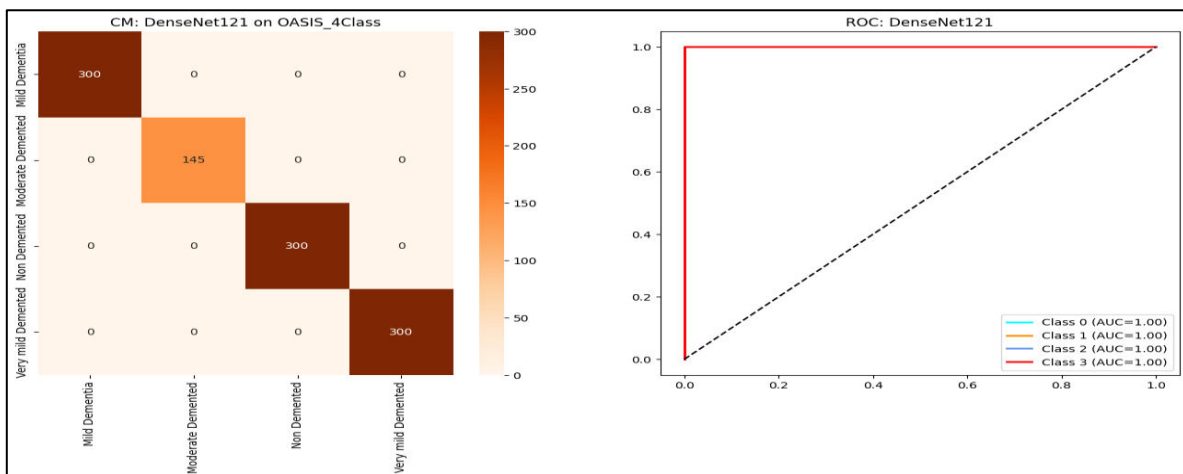
(d)

Figure 5: Comprehensive performance analysis on the ADNI dataset (80:10:10 split). The matrices and ROC-AUC curves illustrate the predictive capabilities of (a) ResNet50V2, (b) DenseNet121, (c) VGG16, and (d) EfficientNetB0.

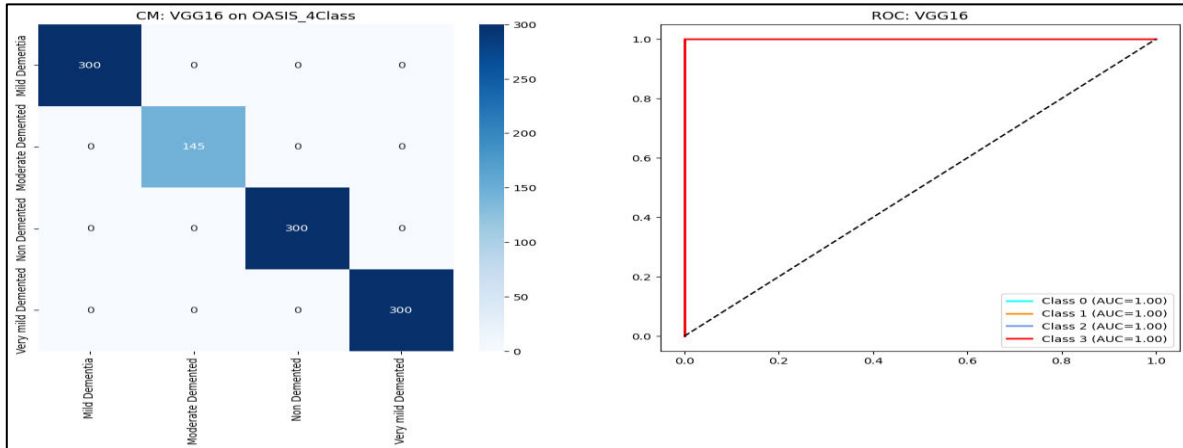
Figure 6 captures the anomalous, near-perfect classification on the OASIS dataset. The stark lack of off-diagonal misclassifications in these matrices visually reinforces the high likelihood of highly separable features or inherent dataset-level regularities, as noted in the tabulated results.



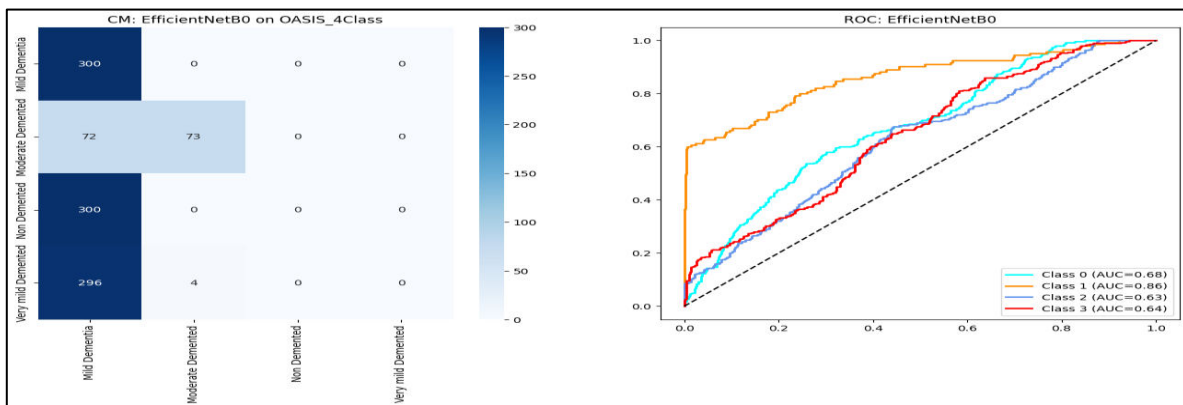
(a)



(b)



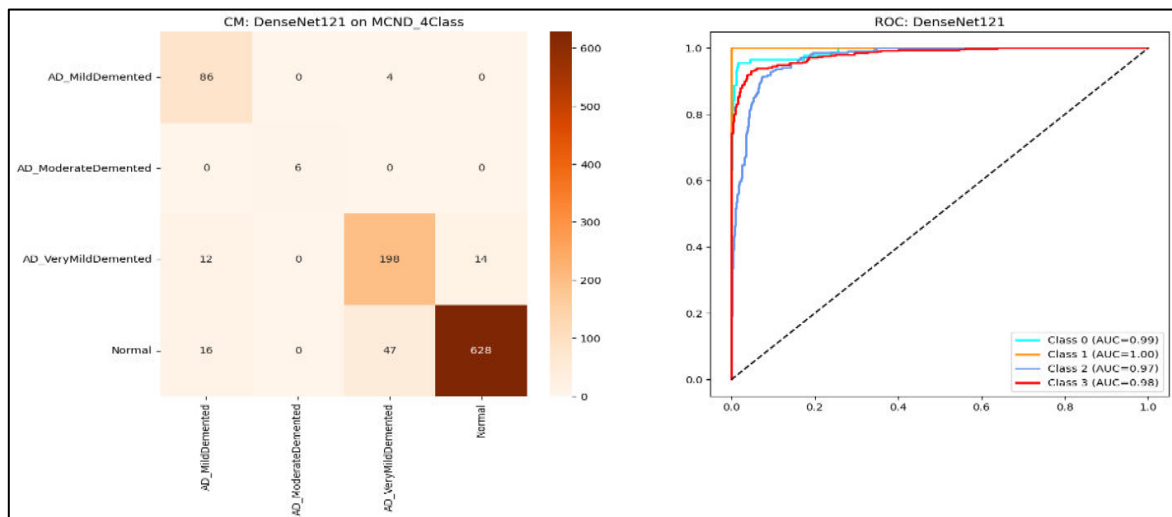
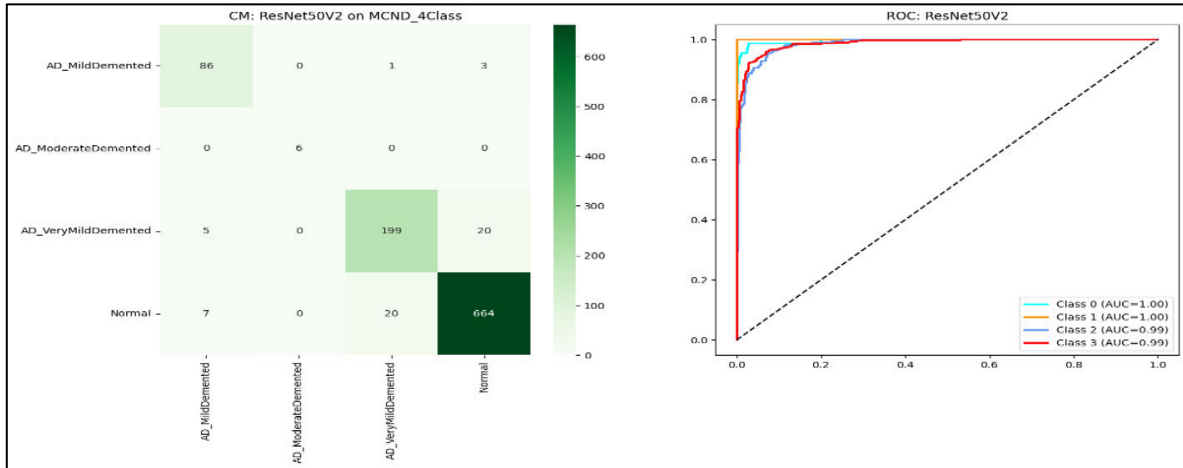
(c)

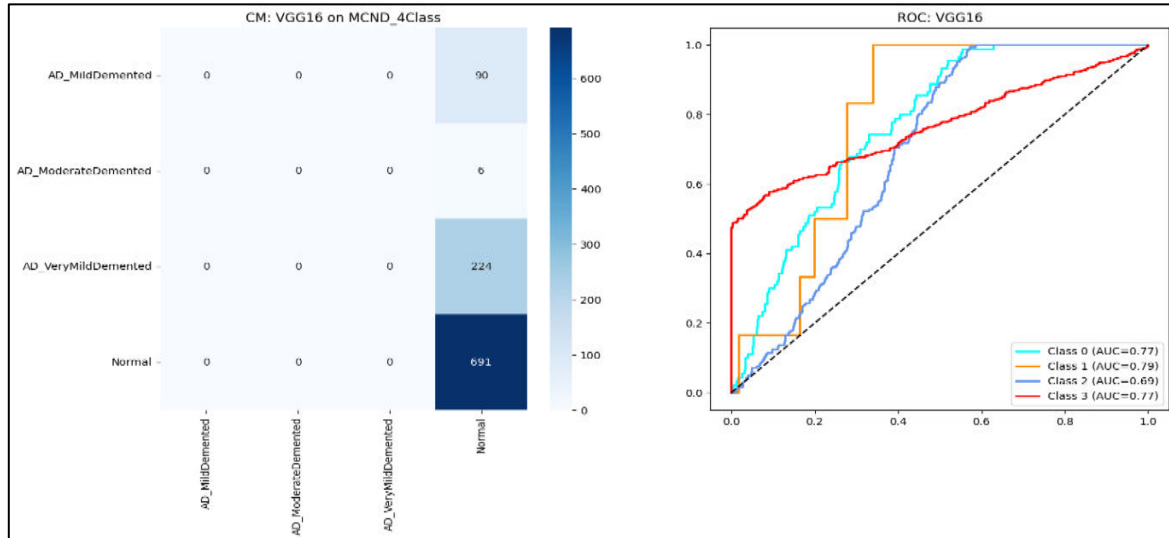


(d)

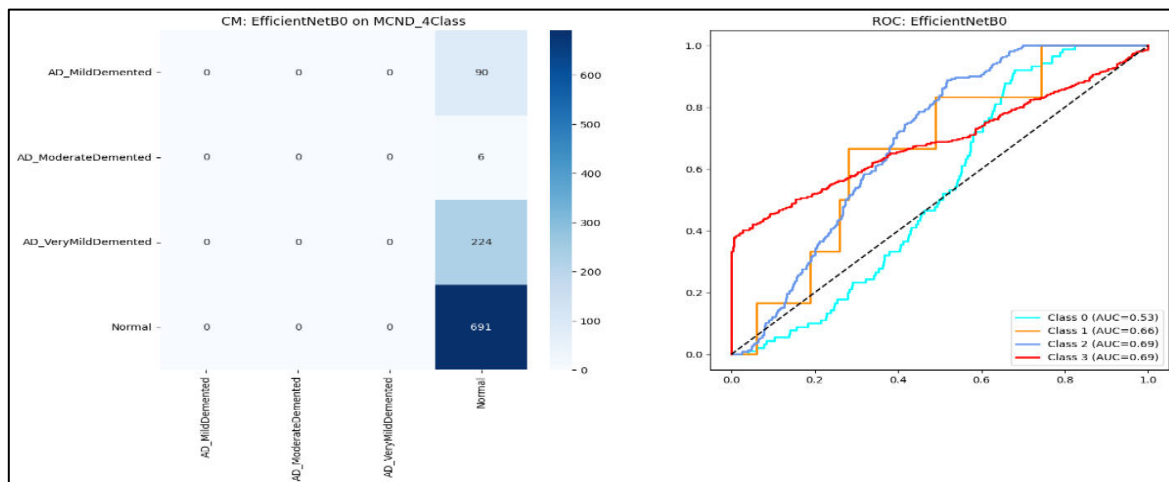
Figure 6: Classification fidelity across the OASIS dataset. Architectures (a) ResNet50V2, (b) DenseNet121, and (c) VGG16 demonstrated near-perfect separability, while (d) EfficientNetB0 struggled to capture distinct features.

Finally, the most critical visual evidence of architectural robustness is presented in **Figure 7**, representing the complex MCND dataset. Here, the contrast between residual and sequential architectures is undeniable. While ResNet50V2 maintains a strong diagonal prediction pathway with minimal spread into distant severity stages, VGG16 and EfficientNetB0 demonstrate a complete structural breakdown, paving the way for a deeper analysis of mode collapse.





(c)



(d)

Figure 7: Evaluation on the highly complex MCND dataset. (a) ResNet50V2 and (b) DenseNet121 maintained robust severity grading, whereas (c) VGG16 and (d) EfficientNetB0 exhibited severe structural failure.

5.2 Model-Specific Behaviors and Mode Collapse.

A critical architectural limitation was observed in the VGG16 and EfficientNetB0 models when evaluated on the MCND dataset. Across all data splits, both models plateaued at approximately 68.3% accuracy with a macro-averaged Recall of exactly 0.25 ($\frac{1}{4}$).

In a four-class classification setting, a recall of exactly 0.25 strongly suggests mode collapse, particularly when accompanied by predictions concentrated in a single class. Lacking the residual skip connections of ResNet50V2 and the dense feature reuse mechanism of DenseNet121, VGG16 and EfficientNetB0 were unable to effectively learn the subtle multi-class decision boundaries of the MCND severity stages. Consequently, both networks predominantly predicted a single majority class for most validation samples.

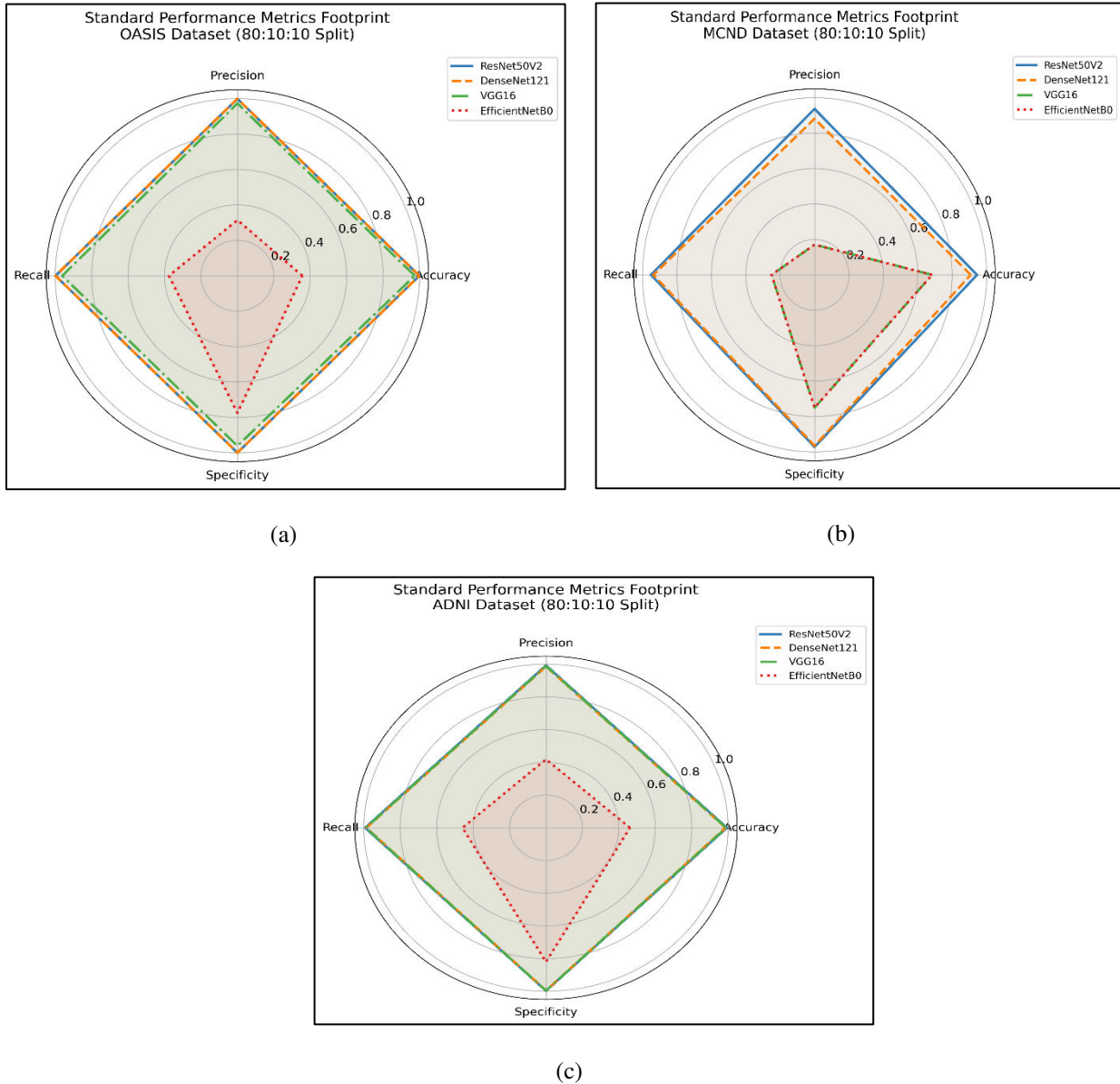
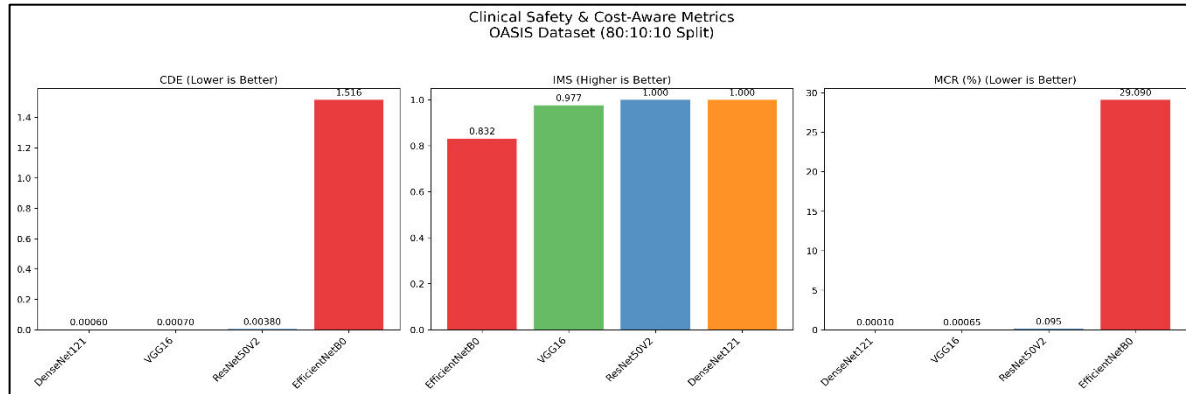


Figure 8: Radar charts detailing the standard performance footprints (Accuracy, Precision, Recall, and Specificity) of four evaluated deep learning architectures across the (a) OASIS, (b) MCND, and (c) ADNI datasets using the optimal 80:10:10 data split.

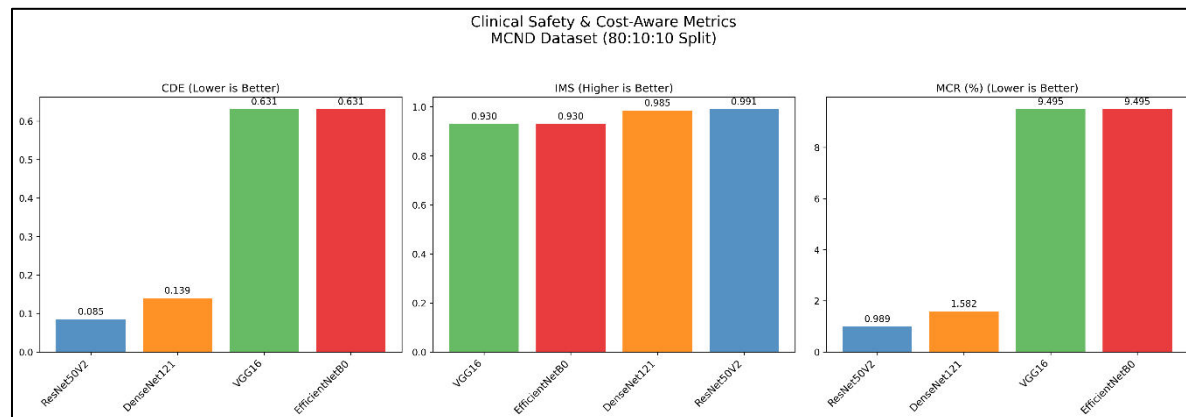
EfficientNetB0, in particular, demonstrated comparatively lower performance across all datasets (accuracies ranging from 28% to 47%), suggesting that compound-scaled, parameter-efficient architectures may require more specialized fine-tuning strategies for structural MRI classification. Conversely, ResNet50V2 effectively mitigated this degradation. Its residual connections facilitated stable gradient propagation throughout the network, enabling the extraction of discriminative feature representations required to differentiate adjacent Alzheimer's disease severity stages while maintaining robust classification performance.

5.3 Clinical Safety Analysis.

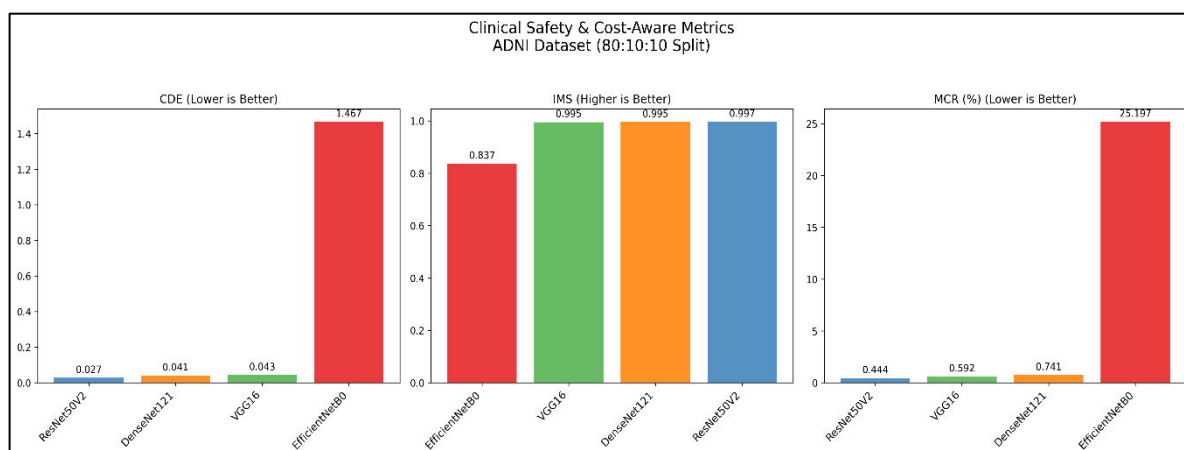
The primary objective of the NeuroCARE framework is to evaluate models based on clinical safety rather than raw statistical accuracy.



(a)



(b)



(c)

Figure 9: Clinical safety and cost-aware metric comparison across the ADNI, OASIS, and MCND datasets (80:10:10 split).

The Cost-Sensitive Distance Error (CDE) and Major Misclassification Rate (MCR) metrics provided profound insights into the true reliability of these systems. On the complex MCND dataset (80:10:10 split), **ResNet50V2** recorded an MCR of just 0.989% and a highly optimized CDE of 0.085. This translates to an Index of Model Safety (IMS) of 0.990. Clinically, this means that even when ResNet50V2 made an incorrect prediction, the error was almost exclusively an adjacent-stage misclassification (e.g., confusing Mild AD with Very Mild AD). It successfully avoided dangerous, stage-skipping errors.

In stark contrast, the collapsed **VGG16** model on the same MCND dataset yielded a catastrophic MCR of 9.49% and a CDE of 0.631. Nearly 10% of its predictions were major clinical failures (e.g., classifying a normal patient as having advanced AD, or vice versa). Standard evaluation metrics obscure this danger, presenting VGG16's 68% accuracy as merely "poor," while the NeuroCARE metrics expose it as clinically hazardous. By actively penalizing these severity-graded distances, the NeuroCARE framework definitively proves that residual-based architectures like ResNet50V2 are not just statistically superior, but vastly safer for deployment in automated neurological diagnostics.

VI. CONCLUSION AND FUTURE WORK

6.1 Summary of findings.

The transition from binary classification to multi-class severity grading represents a critical frontier in the automated diagnosis of Alzheimer's Disease. However, this study demonstrated that utilizing standard machine learning evaluation metrics for graded medical conditions is fundamentally flawed, as these metrics obscure severe, stage-skipping diagnostic errors. To rectify this, we introduced the NeuroCARE framework a cost-aware deep learning pipeline designed to prioritize clinical safety alongside predictive accuracy.

By benchmarking four distinct convolutional architectures across three heterogeneous, multi-institutional cohorts (ADNI, OASIS, and MCND), this research yielded several vital insights. First, spatial preprocessing via CLAHE effectively normalized varied MRI protocols, providing a robust feature baseline. Second, under the optimal 80:10:10 data split, residual architectures, specifically ResNet50V2, demonstrated unmatched stability and generalizability, peaking at 99.35% accuracy on ADNI and 94.46% on the highly complex MCND dataset.

Crucially, the implementation of our novel clinical safety metrics Cost-Sensitive Distance Error (CDE), Major Misclassification Rate (MCR), and the Index of Model Safety (IMS) proved that ResNet50V2 not only classified accurately but did so safely. It recorded an exceptionally low MCR of 0.989% on the MCND cohort. Conversely, the NeuroCARE metrics successfully exposed the latent clinical hazards of sequential networks like VGG16, which suffered from severe mode collapse (25% recall) and committed catastrophic, stage-skipping misclassifications in nearly 10% of its predictions on the same dataset. Ultimately, this study establishes that deep learning frameworks intended for neurological diagnostics must incorporate severity-aware penalties to be deemed clinically viable.

6.2 Limitations of the current study.

While the NeuroCARE framework establishes a robust standard for clinical reliability, several limitations must be acknowledged. First, the extraordinary 100% accuracy achieved by multiple models on the OASIS dataset highlights a systemic challenge in public medical datasets: the potential for slice-level data leakage. When 2D slices from the same 3D volumetric brain scan are distributed across both training and testing sets, models can achieve perfect memorization rather than true anatomical generalization [20].

Second, the current framework operates exclusively on 2D cross-sectional structural MRI data. While transfer learning on 2D networks is computationally efficient, it inherently discards the continuous 3D spatial context of the brain's internal architecture, which could hold vital volumetric biomarkers for the earliest stages of impairment. Finally, the fixed severity mapping (assigning integer values 0 through 3) assumes a linear progression of cognitive decline, which may oversimplify the highly individualized and fluctuating trajectory of neurodegenerative pathology.

6.3 Future directions.

Future iterations of the NeuroCARE framework will focus on expanding its multi-modal and temporal capabilities. Integrating structural MRI with functional imaging (fMRI) or metabolic scans (FDG-PET) via cross-modal attention mechanisms could drastically improve the detection of pre-symptomatic Alzheimer's. Furthermore, rather than relying solely on image data, fusing the convolutional pipeline with Recurrent Neural Networks (RNNs) or Transformers to

process longitudinal patient histories, clinical texts, and neuropsychological biomarker sub-scores will create a more holistic diagnostic engine.

To address the limitations of 2D processing, future research will explore the deployment of cost-aware 3D-CNNs, evaluating whether the computational overhead of volumetric convolutions translates to a statistically significant reduction in the Cost-Sensitive Distance Error (CDE). Finally, to circumvent the data leakage and privacy constraints of centralized healthcare datasets, implementing the NeuroCARE metrics within a Federated Learning architecture will be a crucial next step, allowing multi-institutional model training without compromising patient confidentiality.

REFERENCES

- 1) F. U. R. Faisal and G.-R. Kwon, "Automated detection of Alzheimer's disease and mild cognitive impairment using whole brain MRI," *IEEE Access*, vol. 10, pp. 65055–65069, Jun. 2022.
- 2) C. M. Chabib, L. J. Hadjileontiadis, and A. Al Shehhi, "DeepCurvMRI: Deep convolutional curvelet transform-based MRI approach for early detection of Alzheimer's disease," *IEEE Access*, vol. 11, pp. 44650–44661, May 2023.
- 3) Almohimeed, R. M. A. Saad, S. Mostafa, N. M. El-Rashidy, et al., "Explainable artificial intelligence of multi-level stacking ensemble for detection of Alzheimer's disease based on particle swarm optimization and the sub-scores of cognitive biomarkers," *IEEE Access*, vol. 11, pp. 123173–123194, Oct. 2023.
- 4) D. V. Puri, P. H. Kachare, S. B. Sangle, R. Kirner, et al., "LEADNet: Detection of Alzheimer's disease using spatiotemporal EEG analysis and low-complexity CNN," *IEEE Access*, vol. 12, pp. 113888–113899, Jul. 2024.
- 5) E. Kina, "TLEABLCNN: Brain and Alzheimer's disease detection using attention-based explainable Deep Learning (DL) and SMOTE using imbalanced brain MRI," *IEEE Access*, vol. 13, pp. 27670–27685, Feb. 2025.
- 6) D.-G. Lee and S.-B. Lee, "Diagnosis of Alzheimer's disease and frontotemporal AD from electroencephalography signals," *IEEE Trans. Neural Syst. Rehabil. Eng.*, Early Access, Jan. 2025. S. Gauthier, P. Rosa-Neto, J. A. Morais, and C. Webster, "World Alzheimer Report 2021: Journey through the diagnosis of AD," *Alzheimer's Disease Int.*, London, U.K., 2021.
- 7) J. Zhang, B. Zheng, A. Gao, X. Feng, et al., "A 3D densely connected convolution neural network with connection-wise attention mechanism for Alzheimer's disease classification," *Magn. Reson. Imag.*, vol. 78, pp. 119–126, May 2021.
- 8) S. Shameer Basha and B. Sathyanarayana, "Design of Neuropsychological Feature-Based Machine Learning Models for the Multi-Class Detection of Alzheimer's Disease Progression," *International Journal of Applied Mathematics*, vol. 38, no. 10s, pp. 1006–1025, 2025.
- 9) M. A. DeTure and D. W. Dickson, "The neuropathological diagnosis of Alzheimer's disease," *Mol. Neurodegeneration*, vol. 14, no. 1, pp. 1–18, Aug. 2019.
- 10) D. P. Veitch, M. W. Weiner, P. S. Aisen, et al., "Understanding disease progression and improving Alzheimer's disease clinical trials: Recent highlights from the Alzheimer's disease neuroimaging initiative," *Alzheimer's AD*, vol. 15, no. 1, pp. 106–152, 2019.
- 11) S. Lahmiri and A. Shmuel, "Performance of machine learning (ML) methods applied to structural MRI and ADAS cognitive scores in diagnosing Alzheimer's disease," *Biomed. Signal Process. Control*, vol. 52, pp. 414–419, Jul. 2019.
- 12) E. Hosseini-Asl, R. Keynton, and A. El-Baz, "Alzheimer's disease diagnostics by adaptation of 3D convolutional network," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 126–130.
- 13) J. Islam and Y. Zhang, "Brain MRI analysis for Alzheimer's disease diagnosis using an ensemble system of deep convolutional neural networks," *Brain Informat.*, vol. 5, no. 2, pp. 1–14, 2018.
- 14) M. Liu, D. Cheng, K. Wang, and Y. Wang, "Multi-modality cascaded convolutional neural networks for Alzheimer's disease diagnosis," *Neuroinformatics*, vol. 16, nos. 3–4, pp. 295–308, Oct. 2018.
- 15) D. Lu, K. Popuri, G. W. Ding, R. Balachandar, and M. F. Beg, "Multimodal and multiscale deep neural networks for the early diagnosis of Alzheimer's disease using structural MR and FDG-PET images," *Sci. Rep.*, vol. 8, no. 1, pp. 1–13, Dec. 2018.
- 16) T. Zhou, K.-H. Thung, X. Zhu, and D. Shen, "Effective feature learning and fusion of multimodality data using stage-wise deep neural network for AD diagnosis," *Hum. Brain Mapping*, vol. 40, no. 3, pp. 1001–1016, 2019.
- 17) S. Sarraf, D. D. DeSouza, J. Anderson, and G. Tofghi, "DeepAD: Alzheimer's disease classification via deep convolutional neural networks using MRI and fMRI," *BioRxiv*, 2017.

- 18) E. Maxwell, T. A. Warner, and F. Fang, "Implementation of machine learning (ML) classification in remote sensing: An applied review," *Int. J. Remote Sens.*, vol. 39, no. 9, pp. 2784–2817, 2018.
- 19) J. Samper-González, N. Burgos, S. Bottani, et al., "Reproducible evaluation of classification methods in Alzheimer's disease: Framework and application to MRI and PET data," *NeuroImage*, vol. 183, pp. 504–521, Dec. 2018.
- 20) W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11–26, Apr. 2017.
- 21) G. Litjens, T. Kooi, B. E. Bejnordi, et al., "A survey on Deep Learning (DL) in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, 2017.
- 22) J. Gu, Z. Wang, J. Kuen, et al., "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, 2018.
- 23) Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning (DL)," *Nature*, vol. 521, pp. 436–444, Nov. 2015.
- 24) Z.-Z. Wu, T. Weise, Y. Wang, and Y. Wang, "Convolutional neural network based weakly supervised learning for aircraft detection from remote sensing image," *IEEE Access*, vol. 8, pp. 158097–158106, 2020.
- 25) R. K. Mishra, S. Urolagin, J. A. A. Jothi, et al., "Deep Learning (DL)-based sentiment analysis and topic modeling on tourism during covid-19 pandemic," *Frontiers Comput. Sci.*, vol. 3, Nov. 2021.
- 26) J. Islam and Y. Zhang, "A novel Deep Learning (DL) based multi-class classification method for Alzheimer's disease detection using brain MRI data," in *Proc. Int. Conf. Brain Inform.*, Springer, 2017, pp. 213–222.
- 27) R. Jain, N. Jain, A. Aggarwal, and D. J. Hemanth, "Convolutional neural network based Alzheimer's disease classification from magnetic resonance brain images," *Cogn. Syst. Res.*, vol. 57, pp. 147–159, Oct. 2019.
- 28) H. T. Gorji and N. Kaabouch, "A Deep Learning (DL) approach for diagnosis of mild cognitive impairment based on MRI images," *Brain Sci.*, vol. 9, no. 9, p. 217, Aug. 2019.
- 29) J. P. Kim, J. Kim, Y. H. Park, et al., "Machine learning (ML) based hierarchical classification of frontotemporal AD and Alzheimer's disease," *NeuroImage: Clin.*, vol. 23, 2019.
- 30) X. Long, L. Chen, C. Jiang, L. Zhang, et al., "Prediction and classification of Alzheimer disease based on quantification of MRI deformation," *PLoS ONE*, vol. 12, no. 3, Mar. 2017.
- 31) H. Guo, F. Zhang, J. Chen, Y. Xu, and J. Xiang, "Machine learning (ML) classification combining multiple features of a hyper-network of fMRI data in Alzheimer's disease," *Frontiers Neurosci.*, vol. 11, p. 615, Nov. 2017.
- 32) L. Khedher, J. Ramírez, J. M. Górriz, et al., "Independent component analysis-based classification of Alzheimer's disease from segmented MRI data," in *Proc. IWANN*, Springer, 2015, pp. 78–87.
- 33) T. Tong, R. Wolz, Q. Gao, et al., "Multiple instance learning for classification of AD in brain MRI," *Med. Image Anal.*, vol. 18, no. 5, pp. 808–818, 2014.
- 34) Y. Gupta, K. H. Lee, K. Y. Choi, et al., "Early diagnosis of Alzheimer's disease using combined features from voxel-based morphometry and cortical regions of MRI T1 brain images," *PLoS ONE*, vol. 14, no. 10, Oct. 2019.
- 35) Esteva, B. Kuprel, R. A. Novoa, et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, pp. 115–118, Jan. 2017.
- 36) S. Vieira, W. H. L. Pinaya, and A. Mechelli, "Using Deep Learning (DL) to investigate the neuroimaging correlates of psychiatric and neurological disorders," *Neurosci. Biobehavioral Rev.*, vol. 74, pp. 58–75, Mar. 2017.
- 37) Esteva, A. Robicquet, B. Ramsundar, et al., "A guide to Deep Learning (DL) in healthcare," *Nature Med.*, vol. 25, no. 1, pp. 24–29, Jan. 2019.
- 38) R. Silva, G. S. Silva, R. G. de Souza, et al., "Model based on deep feature extraction for diagnosis of Alzheimer's disease," in *Proc. IJCNN*, Jul. 2019, pp. 1–7.
- 39) S. Liu, S. Liu, W. Cai, et al., "Early diagnosis of Alzheimer's disease with Deep Learning (DL)," in *Proc. IEEE ISBI*, May 2014, pp. 1015–1018.
- 40) S. Alinsaif and J. Lang, "3D shearlet-based descriptors combined with deep features for the classification of Alzheimer's disease," *Comput. Biol. Med.*, vol. 138, Nov. 2021.
- 41) S.-H. Wang, P. Phillips, Y. Sui, et al., "Classification of Alzheimer's disease based on eight-layer convolutional neural network with leaky ReLU and max pooling," *J. Med. Syst.*, vol. 42, no. 5, pp. 1–11, May 2018.
- 42) S. Ahmed, K. Y. Choi, J. J. Lee, et al., "Ensembles of patch-based classifiers for diagnosis of Alzheimer diseases," *IEEE Access*, vol. 7, pp. 73373–73383, 2019.
- 43) C. Feng, A. Elazab, P. Yang, et al., "Deep Learning (DL) framework for Alzheimer's disease diagnosis via 3D-CNN and FSBi-LSTM," *IEEE Access*, vol. 7, pp. 63605–63618, 2019.
- 44) S. Basaia, F. Agosta, L. Wagner, et al., "Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks," *NeuroImage: Clin.*, vol. 21, 2019.

- 45) S. Korolev, A. Safiullin, M. Belyaev, and Y. Dodonova, "Residual and plain convolutional neural networks for 3D brain MRI classification," in Proc. IEEE ISBI, Apr. 2017, pp. 835–838.
- 46) G. Folego, M. Weiler, R. F. Casseb, et al., "Alzheimer's disease detection through whole-brain 3D-CNN MRI," Frontiers Bioeng. Biotechnol., vol. 8, p. 1193, Oct. 2020.
- 47) Abrol, M. Bhattarai, A. Fedorov, et al., "Deep residual learning for neuroimaging: An application to predict progression to Alzheimer's disease," J. Neurosci. Methods, vol. 339, 2020.
- 48) Basher, B. C. Kim, K. H. Lee, and H. Y. Jung, "Volumetric feature-based Alzheimer's disease diagnosis from sMRI data using a convolutional neural network and a deep neural network," IEEE Access, vol. 9, pp. 29870–29882, 2021.
- 49) J. Liu, M. Li, Y. Luo, et al., "Alzheimer's disease detection using depthwise separable convolutional neural networks," Comput. Methods Programs Biomed., vol. 203, May 2021.
- 50) H. Wang, Y. Shen, S. Wang, et al., "Ensemble of 3D densely connected convolutional network for diagnosis of mild cognitive impairment and Alzheimer's disease," Neurocomputing, vol. 333, pp. 145–156, Mar. 2019.
- 51) X. Liu, Q. Xu, and N. Wang, "A survey on deep neural network-based image captioning," Vis. Comput., vol. 35, no. 3, pp. 445–470, Mar. 2019.
- 52) K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.
- 53) S. Shameer Basha and B. Sathyanarayana, "AD-GCRS: A Generalized Clinical Reliability System for Multistage Alzheimer's Classification Leveraging Transfer Learning across Heterogeneous MRI Datasets," International Journal of Science, Engineering and Technology, 2025.
- 54) S. Shameer Basha and B. Sathyanarayana, "A Current Perspective on Machine Learning Approaches for Early Alzheimer's Disease Detection," International Journal of Scientific Research and Engineering Development, vol. 8, no. 2, pp. 450, 2025.



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 9940 572 462  6381 907 438  ijirset@gmail.com



www.ijirset.com

Scan to save the contact details