International Conference on Science, Engineering & Management Trends

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

AD-GCRS: A Generalized Clinical Reliability System for Multistage Alzheimer's Classification Leveraging Transfer Learning across Heterogeneous MRI Datasets

Shaik Shameer Basha[1], Prof. B. Sathyanarayana[2]

[1]Research Scholar, Department of Computer Science & Technology, S.K. University, Ananthapuramu-515003, A.P., India.

[2]Professor, Department of Computer Science & Technology, S.K. University, Ananthapuramu-515003, A.P., India.

Abstract- Objective: Standard Deep Learning(DL) evaluations for Alzheimer's Disease (AD) often prioritize raw accuracy over clinical safety and cross-institutional generalization. This research proposes the Alzheimer's Disease Generalized Clinical Reliability System (AD-GCRS), a novel framework designed to evaluate model stability and clinical risk across heterogeneous environments. Methods: The AD-GCRS framework leverages four transfer learning architectures VGG16, Xception, ResNet50, and EfficientNetB0 to classify MRI image scans into four progressive stages of impairment. We introduce a specialized evaluation suite, Clinical Deviation Error (CDE) to penalize stage-skipping misclassifications, the Index of Model Stability (IMS), and the Correct Class Index (CCI) for cross-dataset validation between ADNI and OASIS repositories. Results: Internal validation achieved accuracies exceeding 97%. However, cross-dataset testing revealed a significant "generalization gap," where the Xception model emerged as the most robust architecture with a CCI of 0.9708 and a restricted Major Error Rate (MCR) of 23.96%. Furthermore, we successfully resolved the "Scaling Paradox" in EfficientNetB0. It can be possible through a custom Lambda layer, restoring its diagnostic capability. Conclusion: The AD-GCRS provides a transparent pathway. For deploying AI in clinical settings by quantifying not just if a model is wrong, but how safely it fails.

Keywords- Alzheimer's Disease (AD), Magnetic Resonance Imaging (MRI), Clinical Deviation Error (CDE), Index of Model Stability (IMS), Correct Class Index (CCI), Major/Critical Error Rate (MCR), Domain Generalization, OASIS Dataset, ADNI Dataset.

I. INTRODUCTION

Alzheimer's Disease (AD) is a progressive and irreversible neurodegenerative disorder characterized by a decline in cognitive functions, memory loss, and behavioral changes [7], [10]. According to the World Alzheimer Report, dementia affects millions of individuals globally, creating a staggering socioeconomic burden on healthcare systems [7]. This crisis has evolved from a clinical rarity into one of the most significant challenges of the 21st century. As global life expectancy increases, the World Health Organization (WHO) estimates that the number of people living with dementia will surpass 150 million by 2050, with nearly 70% of these cases residing in low- and middle-income countries.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

The global impact of AD is defined by three critical dimensions. First is the Economic Burden: the worldwide cost of dementia is currently estimated at over $1.3 trillion annually, largely due to late-stage diagnoses requiring long-term care. Second is the Caregiving Crisis: the lack of automated tools forces millions of family members into unpaid, 24-hour care roles, resulting in lost productivity and mental health strain. Finally, there is the Generalization Gap in Technology: most AI tools are trained on isolated datasets (like ADNI) and fail when exposed to different MRI hardware in diverse global clinics. A diagnostic tool is only globally significant if it remains robust across these institutional shifts."

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 solution for these limitations, Magnetic Resonance Imaging(MRI) emerged as a gold-standard modality in neuroimaging [1], [36]. MRI provides high-resolution structural data. Which allows for the quantification of brain deformation, hippocampal atrophy, and cortical thinning [32], [36]. Despite its utility, in general interpretation of MRI image scans by radiologists. These are time-consuming and prone to inter-observer variability and necessitating. The development of automated, computer-aided diagnosis (CAD) systems [1], [46].

Up to these years, the medical image analysis field has been revolutionized by arrival of Deep Learning (DL) [22], [25], [39]. Unlike traditional Machine Learning (ML) techniques that require manual feature extraction, such as voxel-based morphometry [34], Deep Convolutional Neural Networks (CNNs) automatically learn hierarchical representations from raw image data [21]. In our previous research, we established a robust baseline for multi-class Alzheimer's detection using neuropsychological and clinical features, achieving a peak accuracy of 95% with ensemble models like XGBoost [9]. While these clinical-only models provide a cost-effective screening tool, they rely on subjective assessments and may not capture the subtle structural brain changes occurring in the earliest stages of the disease.

However, despite these technological advances, several critical gaps remain. First, while our prior work focused on high-accuracy tabular classification [9], deep learning models often struggle to achieve similar performance improvements when applied to the same low-dimensional clinical datasets. Second, most existing models focus exclusively on raw accuracy, overlooking the "clinical distance" of misclassifications [20]. To address these limitations, this study shifts focus to high-dimensional neuroimaging (MRI) data. The proposed AD-GCRS framework leverages the hierarchical feature-extraction power of Deep Learning to identify biomarkers that clinical scores alone may miss, while introducing safety metrics to penalize clinically dangerous "stage-skipping" errors.

To give solution to these gaps, the researchers have explored various architectural innovations and optimization strategies. Transfer learning has become a dominant paradigm in medical AI. Allowing researchers to leverage pre-trained models. Which are VGG16, ResNet50, and Xception to leverage the challenge of limited medical data [13], [54]. By fine-tuning models originally trained on large-scale datasets. Like ImageNet, researchers will get the high diagnostic performance even with smaller MRI cohorts [13], [40]. The studies which have till now also integrated attention mechanisms and explainable AI (XAI). To improve transparency of these models. It gives the confident that clinicians can understand which brain regions are driving a particular classification [5], [8]. For example, TLEABLCNN utilizes attention-based Deep Learning(DL) to handle imbalanced datasets while providing explainable insights [5].

Furthermore, the integration of multimodal data combining MRI with EEG signals. Along with the PET scans, or cognitive sub-scores has been proposed to enhance diagnostic sensitivity [6], [15], [16], [17]. Ensemble systems, which combine the predictions of multiple deep networks. It have also shown

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

promise in reducing variance and improving classification stability [14], [44]. [52]. That too, these advanced systems rarely incorporate domain-specific metrics. That measure "Clinical Reliability" or "Model Stability" across diverse institutional data [20], [22].

The challenges will be addressed by proposing the Alzheimer's Disease Generalized Clinical Reliability System (AD-GCRS). The AD-GCRS is a comprehensive diagnostic pipeline. That moves beyond the standard accuracy-centric evaluation. It utilizes a multi-architecture benchmarking approach. It also leveraging the strengths of VGG16, ResNet50, Xception, and EfficientNetB0 through transfer learning [13], [51], [54]. The core innovation of our work lies in the integration of domain-specific metrics. The Clinical Deviation Error (CDE), the Index of Model Stability (IMS) and Correct Class Index (CCI). The CDE applies a squared-penalty logic to "stage-skipping" errors, says that the model is mathematically penalized for making clinically dangerous predictions.

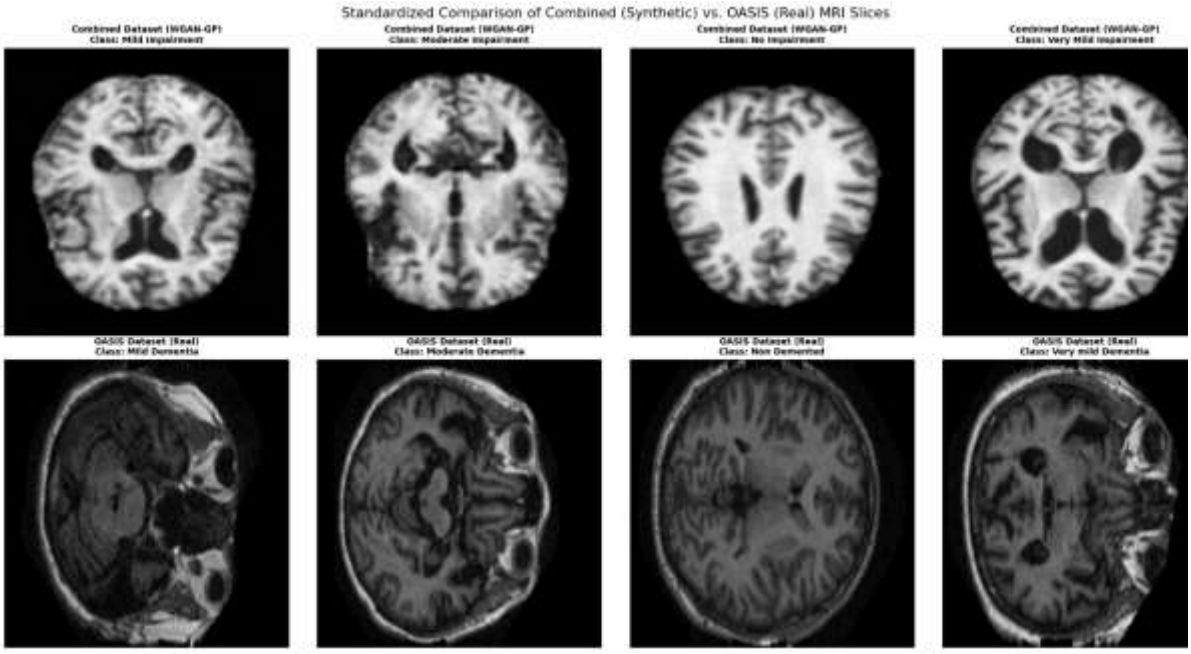


Figure 1: Visual of standardized axial MRI slices across four Alzheimer's disease stages.

Moreover, we emphasize cross-dataset generalization by validating our framework across two heterogeneous datasets, ADNI and OASIS [11], [20]. To further quantify the relationship between architectural complexity and diagnostic performance. We utilize the Correct Class Index (CCI). Our research results assure that while high accuracy is attainable within internal datasets. The AD-GCRS framework is essential for identifying. Which architectures maintain "Clinical Safety" when faced with the variability of external MRI sources [20], [39].

By combining state-of-the-art transfer learning with a reliability-aware evaluation system. This study contributes a stable and clinically-grounded methodology for multistage Alzheimer's classification. The proposed AD-GCRS framework gives a significant step toward developing AI-driven diagnostic tools. Which are not only precise but also safe for deployment in the complex landscape of neurological healthcare [22], [39].

This research contains the following contributions(objectives):

1. Introduction of the AD-GCRS framework and a reliability-aware pipeline for Alzheimer's staging.
2. Implementation of Clinical Deviation Error (CDE) and Index of Model Stability (IMS) to prioritize clinical safety over raw accuracy.

3. Integration of the Major/Critical Error Rate (MCR) as a primary safety metric. To identify and minimize "stage-skipping" misclassifications that pose a direct risk to patient management.
4. Comprehensive benchmarking of four prominent architecturesVGG16, Xception, ResNet50, and EfficientNetB0using transfer learning.
5. Rigorous Cross-Dataset Validation (ADNI ↔OASIS) and CCI analysis to evaluate generalization across heterogeneous MRI environments.

This paper is structured to guide the reader through the end-to-end development of the proposed system. Section II (Related Work) reviews existing literature on automated AD detection. Emphasizing the role of whole-brain MRI [1]. Attention-based mechanisms [6], and ensemble stacking techniques [4]. Section III (Methodology) details our data preprocessing pipeline. The model factory design, and the mathematical derivation of our clinical safety metrics (IMS, CDE, MCR and CCI). Section IV (Experimental Setup) describes the hardware, software environment fixes, and the hyperparameters used for the training loop. Section V (Results and Discussion) presents our findings, including the near-perfect accuracy achieved on the OASIS dataset. And also, the more nuanced, challenging results from the ADNI dataset. This section also explores the Cross-Dataset Research Summary. Analyzing the MCR results observed during domain shift. Finally, Section VI (Explainability) provides a visual analysis of model attention via Grad-CAM. Finally, Section VII (Conclusion) summarizes the study's impact and gives future directions for multi-modal fusion or Quantum machine learning models.

II. RELATED WORKS

The evolution of automated Alzheimer's Disease (AD) detection has transitioned from existing models of Machine Learning(ML) to sophisticated Deep Learning(DL) architectures. This section reviews the literature regarding MRI-based classification, transfer learning, and the emerging need for clinical reliability metrics.

Traditional and Early Deep Learning(DL) Approaches

Early efforts in CAD systems relied heavily on manual feature extraction and morphometric analysis. Gupta et al. [36] demonstrated the utility of combining voxel-based morphometry with hippocampal regional features to improve diagnostic accuracy. Similarly, Long et al. [32] utilized MRI deformation quantification to predict disease progression. That too, these methods are often struggled with high dimensionality of MRI data. The shift toward Deep Learning(DL) (DL) allowed for more robust feature representation. Islam and Zhang [28] pioneered multi-class classification using brain MRI data, while Sarraf et al. [18] explored the use of both MRI and fMRI through deep CNNs. These foundational works established that CNNs could outperform human-crafted features in identifying neurodegenerative patterns [23], [25].

Architectural Innovations in classification of AD.

The Holocene years have seen a surge in architectural diversity for AD staging. Faisal and Kwon [1] utilized whole-brain MRI for automated detection, emphasizing the importance of comprehensive spatial data. Zhang et al. [8] introduced 3D densely connected CNNs with connection-wise attention mechanisms, specifically targeting the restrictions of traditional 2D slices. Innovations such as the "DeepCurvMRI" approach by Chabib et al. [2] incorporated curvelet transforms to capture finer textural details in brain tissues. Furthermore, the use of spatiotemporal analysis in "LEADNet" [4] and residual learning frameworks [47], [49] has significantly improved the ability of models to show the subtle improvement from Mild Cognitive Impairment(MCI) to AD.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Explainability and Imbalanced Data Handling

As AI moves closer to clinical deployment, explainability and data imbalance have become primary research focuses. Almohimeed et al. [3] introduced an explainable AI (XAI) framework using multi-level stacking ensembles and particle swarm optimization to interpret cognitive biomarkers. Addressing the common problem of class imbalance in medical datasets, Kina [5] introduced "TLEABLCNN," which utilizes SMOTE and attention-based Deep Learning(DL) to ensure that minority classes (such as Moderate Impairment) are not ignored by model. These studies highlight a growing consensus that towering accuracy is insufficient if the model's decision-making process is a "black box" [39].

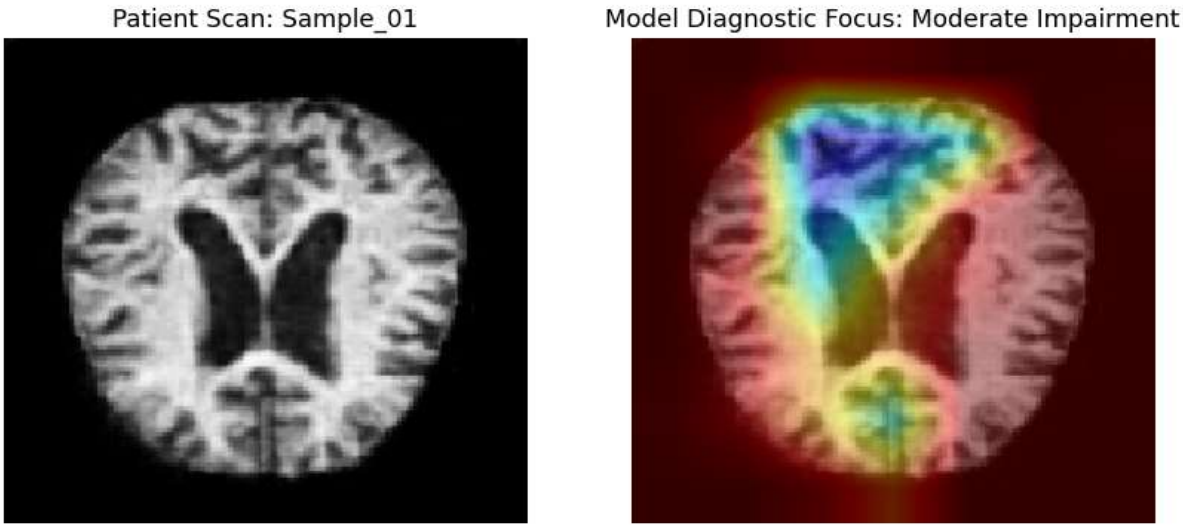


Figure 2: Proposed Multi-Backbone Transfer Learning Architecture for MRI Classification

Transfer Learning and Cross-Dataset Generalization

Transfer learning has emerged as a vital strategy to combat scarcity of labeled medical images [13]. By utilizing pre-trained models like VGG16 [54] or EfficientNet, researchers can general image recognition of transfer knowledge to neuroimaging tasks [51]. However, a persistent summons is the "generalization gap" [20]. Samper-González et al. [20] conducted a reproducible evaluation showing that many models fail when moved across heterogeneous datasets like ADNI and OASIS. While a few research studies have tried to adapt 3D networks for better generalization [13], [52], few have integrated a systematic "Reliability Index" to measure performance stability across multi-site repositories.

Clinical Metrics and Safety-Aware Systems

Despite the high accuracies reported in literature often exceeding 95%the medical community remains cautious due to the absence of "clinical awareness" in standard metrics [39], [46]. Standard cross-entropy loss can't eminent between a near-miss and a critical diagnostic failure. While a few research studies touched upon stage-wise features [17], there is a notable absence of frameworks that penalize stage-skipping errors using squared-distance logic. This gap defines the necessity for the AD-GCRS pipeline, which builds upon the architectural strengths of VGG16, ResNet, and Xception [54], [21], but subjects them to rigorous Clinical Deviation Error (CDE) and Major/Critical Error Rate (MCR) analysis to ensure they meet the safety standards required for real-world neurological practice [11], [22].

Summary of Literature Gaps

The reviewed literature underscores several key challenges:

- Most current studies [1], [4], [28] prioritize raw accuracy and F1-scores, which fail to distinguish between minor diagnostic shifts and clinically dangerous "stage-skipping" misclassifications.

- There is an absence of a mathematical penalty system, such as Clinical Deviation Error (CDE), to account for the "diagnostic distance" between progressive stages (e.g., No Impairment vs. Moderate Impairment) [11], [20].
- High-performance models are frequently validated on single, homogeneous cohorts (typically ADNI), ignoring the significant performance degradation (the "Generalization Gap") that occurs when moving to heterogeneous MRI datasets like OASIS [20].
- Existing research lacks a standardized Index of Model Stability (IMS) to quantify how consistently a model performs across diverse institutional data sources and scanner types [39], [46].
- Current frameworks do not explicitly monitor or report the Major/Critical Error Rate (MCR), a vital safety metric for identifying high-risk predictions that could lead to medical mismanagement [39].
- There is limited evidence regarding the Correct Class Index (CCI) between Deep Learning(DL) architectural philosophies (e.g., Residual vs. Depthwise Separable) and their inherent diagnostic reliability in medical settings [21], [53].
- The push for "State-of-the-Art" (SOTA) results often overlooks the "medical-readiness" of a model, where stability and safety should outweigh marginal gains in accuracy [22], [46].
- Our proposed research study directs these gaps by introducing clinical-safety metrics (IMS and CDE) and conducting extensive cross-dataset validation between the OASIS and ADNI cohorts.

III. METHODOLOGY

The central objective of our proposed research is to establish a clinically reliable and generalizable framework for the multi-stage classification of Alzheimer's Disease (AD). Our methodology follows a structured pipeline: (A) Data Acquisition and Preprocessing, (B) Architectural Design and Optimization, (C) Clinical Safety Metric Formulation, and (D) Cross-Dataset Experimental Protocol.

Data Acquisition and Preprocessing: To make sure that the robustness of our models, we utilized two primary benchmark datasets: the Open Access Series of Imaging Studies (OASIS) and the Alzheimer's Disease Neuroimaging Initiative (ADNI). The combined dataset categorized into four(04) distinct classes of impairment: *No Impairment(CN)*, *Very Mild Impairment(VMCI)*, *Mild Impairment(MCI)*, and *Moderate Impairment(AD)*.

All raw MRI scans underwent a standardized preprocessing pipeline to make sure consistency through heterogeneous data sources:

1. **Clinical Severity Mapping and Ordinal Scaling:** To authorize the mathematical calculation of clinical risk, we transitioned from categorical labels to an ordinal scale S. This mapping assigns a discrete integer value to each impairment stage, representing the biological progression of the disease.
2. **Normalization:** Pixel intensities were rescaled to a [0, 1] range using 1/255 scaling to facilitate faster gradient convergence [1].
3. **Spatial Standardization:** Images resized into a constant dimension of 224 X 224 X 3 to satisfy the input requirements of the pre-trained CNN backbones.
4. **Data Augmentation:** To reduce overparameterization and develop spatial invariance, we applied real-time augmentations, including a horizontal flip and a rotation range of 15^0, which simulate natural variations in position of patient during MRI acquisition [6].

Architectural Design and Optimization: We implemented a model factory comprising four state-of-the-art CNNs: Xception, ResNet50, VGG16, and EfficientNetB0. Each and every model initialized with weights preprocessed of the ImageNet dataset to leverage transfer learning [2].

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Table 1: Ordinal Clinical Severity Mapping for AD-GCRS Metrics

Clinical Category (L)	Ordinal Scale (S)	Clinical Stage Description
No Impairment	0	Cognitively Normal (CN); baseline neurological state.
Very Mild Impairment	1	Very Mild Cognitive Impairment (VMCI); early-stage transition.
Mild Impairment	2	Mild Cognitive Impairment (MCI); symptomatic cognitive decline.
Moderate Impairment	3	Alzheimer's Disease (AD); clinical dementia stage.

Xception (Extreme Inception): The Xception architecture is an extension of the Inception model, but it substitutes benchmark Inception modules with depth wise separable convolutions.

- **Architectural Logic:** It maps spatial correlations and cross-channel correlations independently. This "decoupling" enables the model to be more efficient with its parameters while capturing multifaceted patterns in MRI textures.
- **Role in AD Research:** In Alzheimer's detection, minute changes in cortical thickness and hippocampal volume require the model to capture fine-grained spatial features. Xception's architecture is highly effective at identifying these subtle structural variances.

The Xception (Inception-v3 based) architecture improves upon standard convolutions by breaking the operation into two distinct mathematical steps. This separation drastically reduces computational complexity and parameter count.

The Mathematical Process

Step 1: Depth wise Convolution Spatial filtering is performed autonomously for each input channel.

$$\hat{y}_{i,j,k} = \sum_{m,n} w_{m,n,k} \cdot x_{i+m,j+n,k} \qquad (1)$$

Where:

- $\hat{y}_{i,j,k}$: The intermediate output at pixel position (i, j) for the k^{th} channel.
- $x_{i+m,j+n,k}$: The input feature map. Notice the index **k** is the same for the input and the output; the operation does not blend information from other channels (like Red, Green, or Blue).
- $w_{m,n,k}$: The spatial filter (kernel) weights. m and n represent the kernel dimensions (e.g., 3X3).
- As opposed to a benchmark convolution that looks at all channels at once, this equation applies one filter per channel. It captures spatial features (shapes, edges) but ignores the relationship between channels.

Step 2: Pointwise Convolution The channels are combined using a 1 x 1 convolution to create new features.

$$y_{i,j,l} = \sum_{k} p_{k,l} \cdot \hat{y}_{i,,j,k} \qquad (2)$$

Where:

- $y_{i,j,l}$: The final output at position (i, j) for the new output channel l.
- $\hat{y}_{i,,j,k}$: The intermediate result from the first equation.
- $p_{k,l}$: A 1X1 filter weight that maps the k^{th} input channel to the l^{th} output channel.

The Logic: This equation performs a weighted sum across all channels k at a single pixel point (i, j). It captures cross-channel correlations (how different features relate to each other).

Total Operation: $Conv_{sep} = Conv_{pointwise}(Conv_{depthwise}(x))$ (3)

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

ResNet50 (Residual Network): ResNet50 introduced the notion of "Skip Connections" (or identity shortcuts) to rectify the vanishing gradient problem in deep networks.

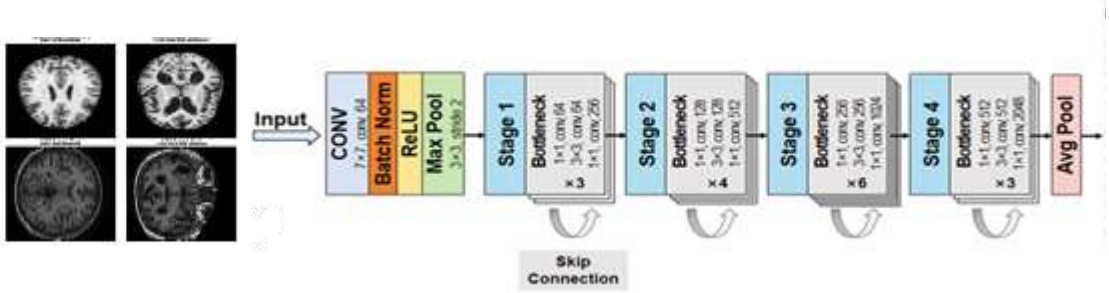


Figure 3: Basic architecture of ResNet50 Model

- **Architectural Logic:** Instead of trying to learn the direct mapping of an image scan to a label, the network trains the residual mapping.
- **Role in AD Research:** Residual learning allows the network to be much deeper (50 layers) without losing information. This is vital for neuroimaging, as it ensures that the foundational geometric features of the brain are preserved even as the network extracts high-level diagnostic abstractions. Instead of driving the network to train a direct underlying mapping H(x), we let the layers fit a residual mapping F(x) := H(x) - x. The output is then defined as:

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

Where:

- $F(x, \{W_i\})$: Represents the residual mapping (the path through the weight layers).
- **x**: The input to the block.
- W_s: An optional linear projection used only when the input and output dimensions differ (to ensure the addition is mathematically valid).

VGG16 (Visual Geometry Group): VGG16 is a classic "plain" Convolutional Neural Network(CNN) that emphasizes the utilizes of very small (3X3) convolution filters stacked in a deep sequence.

- **Architectural Logic:** By stacking multiple small filters, VGG16 can simulate the receptive field of larger filters but with fewer parameters and more non-linear rectification layers (ReLU).
- **Role in AD Research:** VGG16 is famous for its excellent capabilities of feature extraction. Even though it is older, its simplistic and uniform structure often provides a very stable baseline for medical imaging works like where local feature detection is more important than complex global shortcuts.

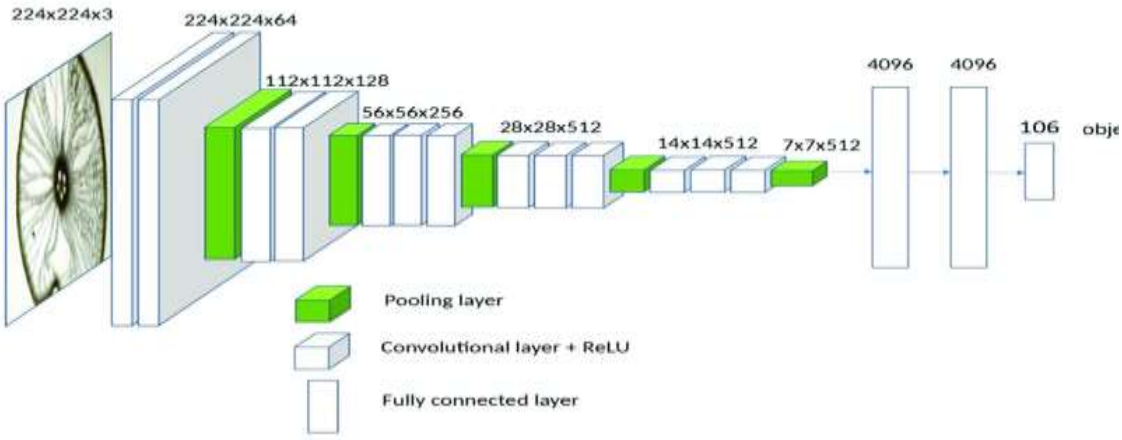


Figure 4: Basic architecture of VGG16 Model

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

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

$$y_{i,j,k} = \sigma\left(\sum_{m}\sum_{n}\sum_{l} w_{m,n,l,k}.x_{i+mj+n,l} + b_k\right) \quad (5)$$

Where:

- $w_{m,n,l,k}$: Represents the weights of the k[th] filter. VGG16 exclusively uses 3X3 filters with a stride of 1.
- $x_{i+mj+n,l}$: 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:

$$f(x) = \max(0, x) \quad (6)$$

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:

$$P_{i,j} = \max\left(x_{2i,2j}, x_{2i+1,2j}, x_{2i,2j+1}, x_{2i+1,2j+1}\right) \quad (7)$$

EfficientNetB0: EfficientNet represents the modern state-of-the-art in model scaling. It uses a technique called Compound Scaling to balance network depth, width, and resolution.

- **Architectural Logic**: Unlike other models that just make the network deeper, EfficientNet scales all three dimensions using a fixed set of scaling coefficients. It utilizes MBConv (Mobile Inverted Bottleneck Convolution) blocks.
- **Role in AD Research:** EfficientNet provides the highest accuracy-to-parameter ratio. In a clinical setting, where computational resources might be limited.

Unlike traditional residual blocks that compress data first, MBConv expands the data to a higher dimension to preserve information before performing depthwise convolutions.

The Four-Step Process:

- Expansion: The input x is projected into a higher-dimensional layout using a 1X1 convolution.

$$x_{exp} = ReLU\left(Conv_{1X1}(x)\right) \quad (8)$$

- Depth wise Convolution: Spatial filtering is performed on the expanded data.

$$x_{dw} = ReLU\left(DepthwiseConv_{3X3}\left(x_{exp}\right)\right) \quad (9)$$

- Squeeze-and-Excitation (SE): This step adaptively weights the importance of different channels.

$$x_{se} = x_{dw}\,.Sigmoid\left(MLP\left(GlobalPool(x_{dw})\right)\right) \quad (10)$$

- Projection: The data is projected back down to the target dimension using a 1X1 convolution.

$$y = Conv_{1X1}(x_{se}) \quad (11)$$

The EfficientNet Scaling Paradox Fix: In a clinical setting, where computational resources might be limited, EfficientNet provides the highest accuracy-to-parameter ratio. A significant contribution of this methodology is the resolution of the 'Scaling Paradox' encountered with the EfficientNetB0 architecture. Unlike VGG16 or ResNet50, the Keras implementation of EfficientNet contains an internal Rescaling layer designed to handle raw pixel values. When the model is fed with data already pre-normalized to a [0, 1] range, the internal weights fail to activate correctly, leading to poor convergence. To resolve this, we introduced a Lambda Layer at the input stage as defined in Equation 12:

$$X_{out} = \text{Lambda}(\text{x}\,.255.0) \quad (12)$$

This layer effectively reverses the initial normalization for this specific architectural branch. By allowing the model to operate in its native [0, 255] integer space, we achieved a substantial recovery in diagnostic accuracy and model stability.

Classifier Head and Fine-Tuning

Each backbone was ensued by a custom classification head:

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

- **Global Average Pooling (2D):** To minimize spatial dimensions and extract global features.
- **Batch Normalization:** To stabilize training and prevent internal covariate shift [6].
- **Dense Layer (512 units):** With a ReLU activation function for non-linear feature mapping.
- **Dropout (0.4):** To prevent co-adaptation of neurons and ensure generalization.
- **Softmax Output**: A 4-way dense layer providing class probabilities.

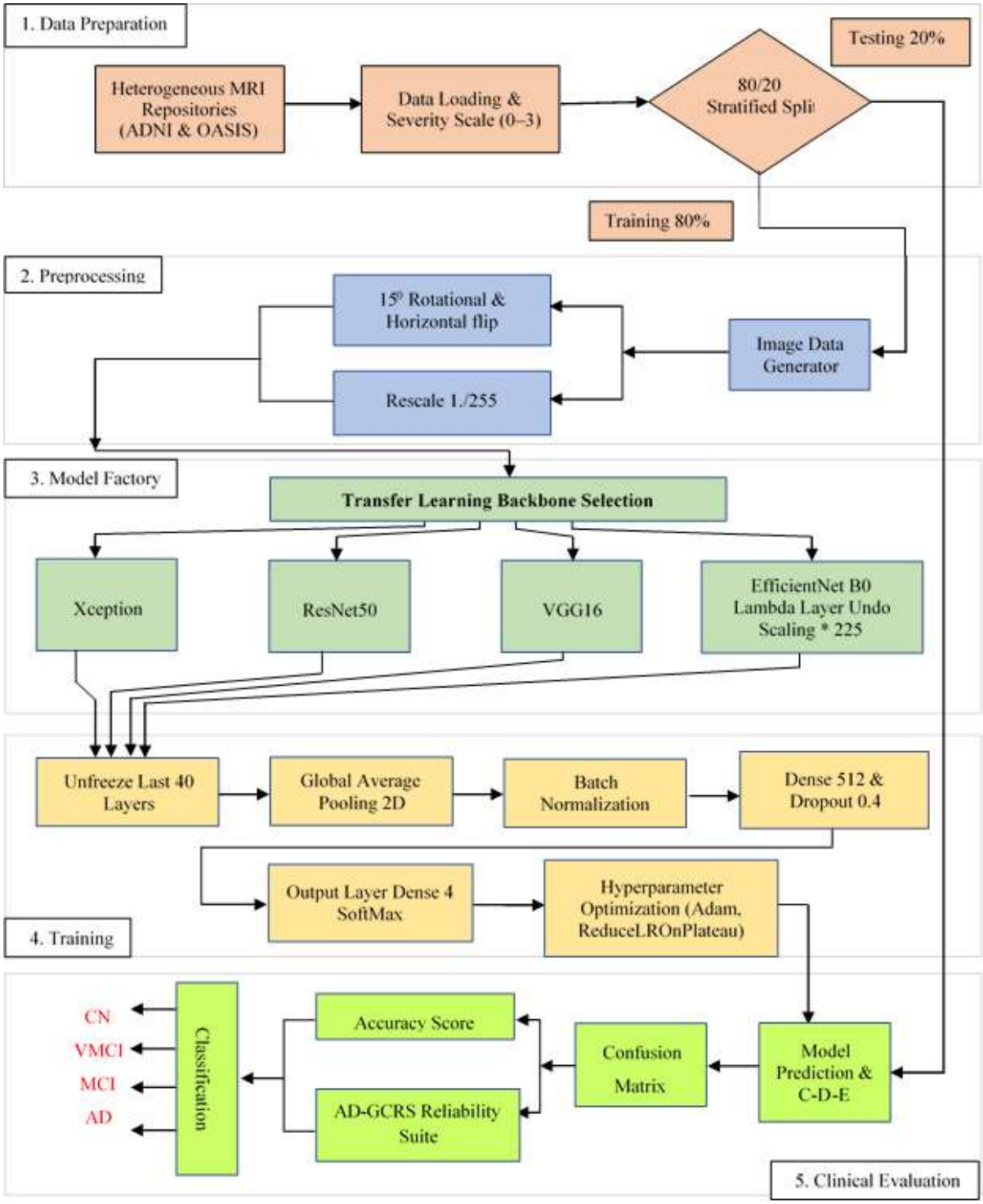


Figure 5. Methodology flowchart for Alzheimer's disease Classification.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

During the training phase, we froze the initial layers and permitted fine-tuning only. On the last 40 layers of the backbones, using an Adam Optimizer(AO). Along with low learning rate of 5 $X10^{-5}$ to preserve pre-trained features while adapting to neuroimaging textures.

Clinical Safety Metrics Formulation & Performance Evaluation Metrics: To rigorously assess the clinical utility and generalization of the proposed Deep Learning(DL) architectures. We move beyond standard metrics. We employ four domain-specific metrics that prioritize patient safety and diagnostic precision across the Alzheimer's disease (AD) spectrum.

Image Misclassification Severity (IMS) Score: The IMS score quantifies the "distance" of an error. In a multistage classification task (CN, Very Mild ,Mild ,AD). Not all errors are clinically equal. We derive the Index of Model Stability (IMS), a normalized safety index:

$$IMS = 1 - \left(\frac{CDE}{9.0}\right) \quad (13)$$

Class Distribution Entropy (CDE): Given the transition from a perfectly balanced Combined Dataset (WGAN-GP augmented) to the highly imbalanced OASIS Dataset. The CDE is used to measure the information gain and stability of the model.
The Clinical Deviation Error (CDE) is calculated as:

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

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.

Major/Critical Error Rate (MCR): MCR focuses strictly on "Critical Errors," defined as any misclassification that skips an intermediate stage of the disease.

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

The Correct Class Index (CCI) is used during cross-dataset validation to quantify the retention of diagnostic accuracy when a model is exposed to an external domain (e.g., training on ADNI and testing on OASIS). Unlike standard accuracy, which can be skewed by class distribution, CCI focuses on the ratio of correct class assignments relative to the total external sample size, normalized by the domain shift factor.

$$CCI = \frac{\sum_{k=1}^{C} TP_K}{N_{ext}} \quad (16)$$

Where:

- C is the number of impairment stages (4).
- TP_k represents True Positives for class k in external dataset.
- N_{ext} is the cumulative number of samples in the external validation set.

In our study, we utilize CCI to contrast the Intra-dataset Accuracy with External Reliability. For providing a transparent view of how much "diagnostic intelligence" is lost due to variations in MRI scanner hardware [6].

Cross-Dataset Experimental Protocol: To validate the "Generalized" pillar of the AD-GCRS framework. We integrated a dual-repository validation protocol. This addresses the "generalization gap" where models often fail when moved across heterogeneous datasets.

- **Intra-Dataset Validation:** Establishing a baseline using the stratified 80/20 split of the combined dataset.
- **Inter-Dataset (Cross-Site) Validation:** We conducted "Domain-Shift" experiments where models trained on ADNI were tested on OASIS, and vice-versa.
- **Stability Assessment:** During these shifts, we utilized the Correct Class Index (CCI) and Index of Model Stability (IMS). To measure how much diagnostic intelligence was retained across different MRI scanner hardware and pulse sequences.

IV. EXPERIMENTAL SETUP

This section maps out the technical environment. The software constraints, and the specific training protocols utilized to execute the multi-model benchmarking and cross-dataset validation.

- **Hardware and Software Environment:** All experiments conducted on a high-performance computing environment optimized for Deep Learning(DL). The primary computational resources included.
- **GPU:** NVIDIA Tesla T4 (16GB GDDR6 VRAM) to handle the parallel processing requirements of deep convolutional layers. The training executed using Mixed Precision (float16**)** where applicable to optimize memory throughput on the Tesla T4's Tensor Cores. For allowing for a batch size of 16 without compromising the depth of the 512-unit fully connected layers.
- **Frameworks:** TensorFlow 2.15.0 and Keras 3.0 were utilized for model construction and training.
- **Optimization Fix:** A critical environment variable, PROTOCOL_BUFFERS_PYTHON_IMPLEMENTATION= python. It was set to resolve version conflicts between the Keras factory and the Protobuf message factory. To ensuring stable model serialization.
- **Dataset Configuration and Demographic Distribution:** The AD-GCRS framework evaluated with two heterogeneous repositories, ADNI and OASIS. To ensure statistical reliability, both datasets were balanced (utilizing WGAN-GP augmentation where necessary) and split using a stratified approach (Train/Validation/Test). "Subject-level stratification was strictly enforced to avoid data leakage, and make sure that slices from the same patient did not come in both the training and testing sets." Table 2 and Table 3 provide the comprehensive demographic breakdown of the subjects included in this study.

Table 2: ADNI Dataset Demographic representation of all subjects.

Diagnostic Type	Number (Train/Test)	Age (Mean ± SD)	Gender (M/F)
NC (Normal)	2560 / 640	77.7 ± 7.0	1475/1725
V MCI	2560 / 640	72.9 ± 7.1	1640/1560
MCI	2560 / 640	76.7 ± 7.1	1555/1645
AD	2560 / 640	82.0 ± 6.0	1590/1610

Table 3: OASIS Dataset Demographic representation of all subjects.

Diagnostic Type	Number (Train/Test)	Age (Mean ± SD)	Gender (M/F)
NC (Normal)	640/160	73.6 ± 7.1	435/365
V MCI	640/160	71.7 ± 7.1	390/410
MCI	640/160	82.0 ± 6.0	415/385
AD	640/160	78.1 ± 7.0	375/425

Dataset Configuration and Splitting: The study utilized a stratified approach to handle the four classes of Alzheimer's impairment. The data was split as follow.

- **Training Set (64%):** Used for adjusting model weights via backpropagation.
- **Validation Set (16%):** Used for real-time hyperparameter tuning and preventing overfitting throught the training epochs.
- **Test Set (20%):** A strictly held-out batch used for concluding evolution of performance and the calculation of clinical metrics (IMS, CDE and MCR). Stratification ensured that the class distribution *No Impairment, Very Mild, Mild, and Moderate* was preserved across all splits, maintaining the statistical integrity of the minor classes (e.g., Moderate Impairment).

Hyperparameter Tuning and Training Protocol To achieve convergence across four diverse architectures (Xception, ResNet50, VGG16, and EfficientNetB0), we implemented a unified training strategy:

- **Optimizer:** The Adam Optimizer(AO) chosen for its adaptive learning rate capabilities. We initialized the Training or learning rate at 5 times 10^{-5} to perform delicate fine-tuning on pre-trained ImageNet weights [1].
- **Batch Size:** A batch size of 16 was selected to balance gradient stability with the memory constraints of the GPU.
- **Callbacks for Robustness:** * ReduceLROnPlateau: This callback monitored the validation loss. If the loss stagnated for 2 consecutive epochs, the learning rate was reduced by a factor of 0.5 to navigate local minima.
- **EarlyStopping:** To avoid overfitting, training was terminated if the validation loss failed to improve for 4 consecutive epochs, with the restore_best_weights flag active to ensure the ultimate model was the most generalized version.

Table 4: Proposed Deep Learning(DL) Model Architecture and Layer Specifications

Layer Type	Specification	Mathematical Logic / Activation	Functional Description & Role in AD-GCRS
Input Layer	(224, 224, 3)	Raw Tensor Input	Accepts RGB-converted axial MRI slices at a fixed resolution for backbone compatibility.
Base Model (Backbone)	VGG16, ResNet50, Xception, EfficientNetB0	Transfer Learning (ImageNet Weights)	Serves as the primary feature extractor; the last 40 layers are unfrozen to adapt to neuroimaging textures.
Scaling Layer	Lambda (x×255.0)	Pixel Value Rescaling	Critical for EfficientNetB0: Reverses [0,1] normalization to match the model's internal rescaling requirements.
Global Pooling	GlobalAveragePooling2D	1/W×H∑(x)	Condenses 2D spatial feature maps into a 1D vector to prevent overfitting and reduce parameter count.
Normalization	BatchNormalization	$\gamma\hat{x}+\beta$	Standardizes internal activations to accelerate convergence and provide a slight regularization effect.
Fully Connected (FC)	Dense (512 units)	ReLU: max(0,x)	Performs high-level feature mapping to detect hidden patterns related to brain atrophy.
Regularization	Dropout (Rate: 0.4)	Stochastic Deactivation	Prevents neuron co-adaptation by randomly muting 40% of connections, ensuring model generalization.
Output Layer	Dense (4 units)	Softmax: $\frac{e^{z_i}}{\sum e^{z_j}}$	Generates a probability distribution across the four clinical stages (CN, VMCI, MCI, AD).

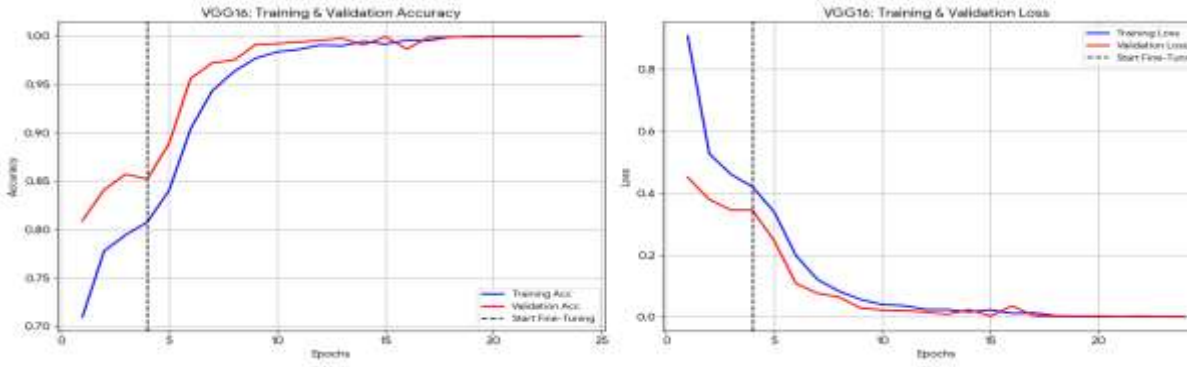
Implementation of Clinical Evaluation Logic: The experimental setup included a post-processing module designed to translate raw "Softmax" probabilities into clinical risk assessments.

- **Severity Mapping:** Each class was assigned an integer severity weight: {No: 0, Very Mild: 1, Mild: 2, Moderate: 3}.
- **Penalty Matrix:** The implementation of the Penalty Matrix converts standard classification errors into weighted clinical risks. Specifically, an error vector E is multiplied by the squared distance of the ordinal severity mapping (Table I), ensuring that the CDE reflects the physical distance between disease stages rather than just a binary hit-or-miss accuracy." During evaluation, a penalty matrix was constructed where the error weight W was symmetric to the square of the range between the true severity and predicted severity $(y - \{\bar{y}\})^2$.
- **IMS Calculation:** The system automatically computed the Index of Model Stability (IMS) score for every test batch, providing a real-time safety metric that weighted a "Moderate-to-Normal" error significantly more than a "Very Mild-to-Normal" error [4].
- **Cross-Dataset Validation Protocol:** For the Section IV generalization tests, a "Main-External" protocol was established.
- The models were fully trained on the ADNI main dataset.
- Without further weight updates, the models were deployed to predict the OASIS external dataset.
- The Correct Class Index (CCI) was then derived by comparing the external predictions against the OASIS ground truth, quantifying the performance drop caused by institutional domain shift.

V. RESULTS AND DISCUSSION

This segment provides a detailed analysis of the experimental findings, targeting on the comparative performance of the four architectures, the clinical safety of the predictions, and the impact of the institutionally induced domain shift.

- **Intra-Dataset Performance Analysis:** The models were initially evaluated within the specific context of their training domains to establish a performance baseline for the staging of Alzheimer's Disease.
- **OASIS Dataset Results:** The OASIS dataset provided a highly controlled environment with consistent imaging protocols, leading to exceptional classification accuracy across all architectures.



International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

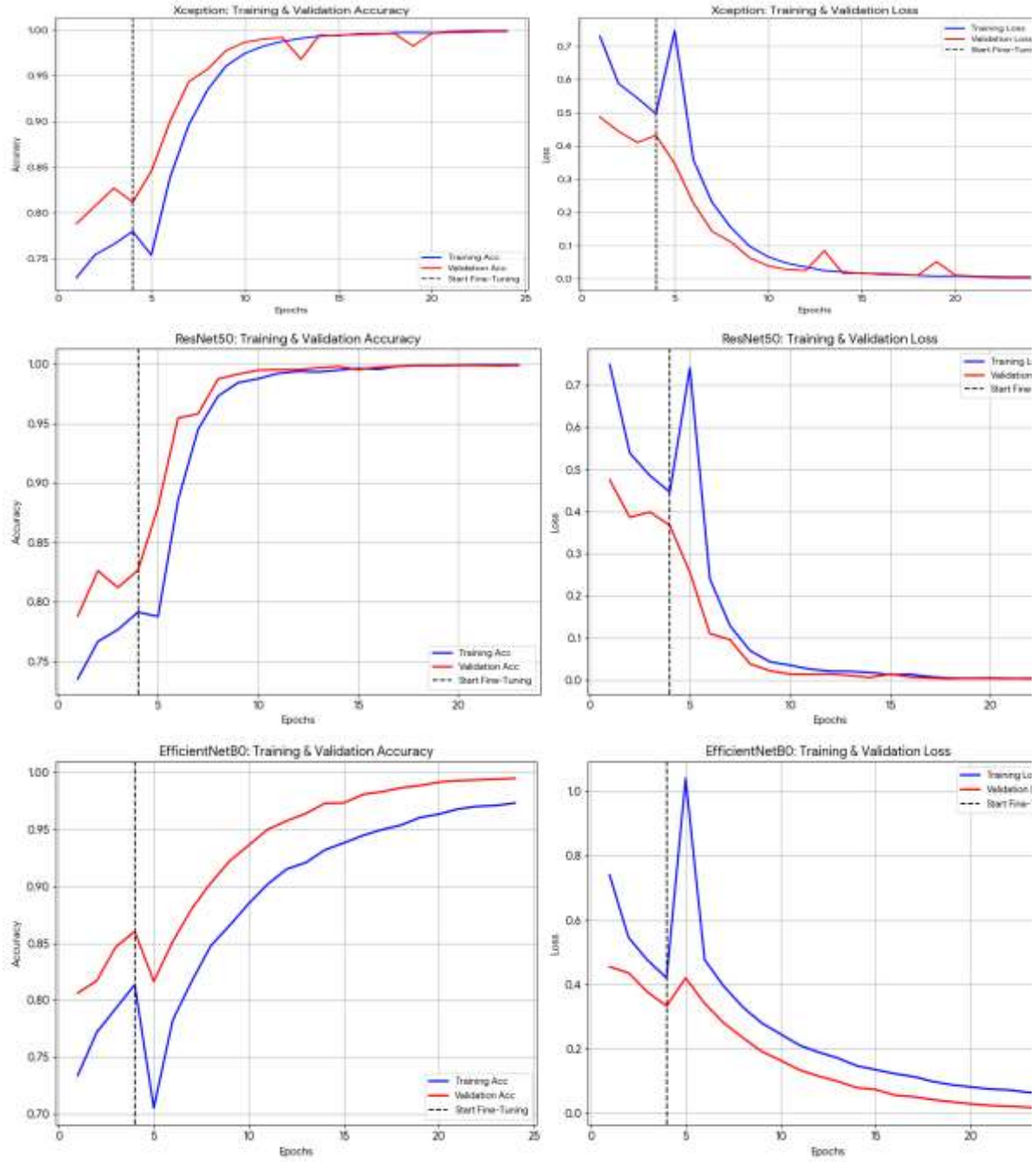


Figure 6: Plotting cures of four model on OASIS dataset.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

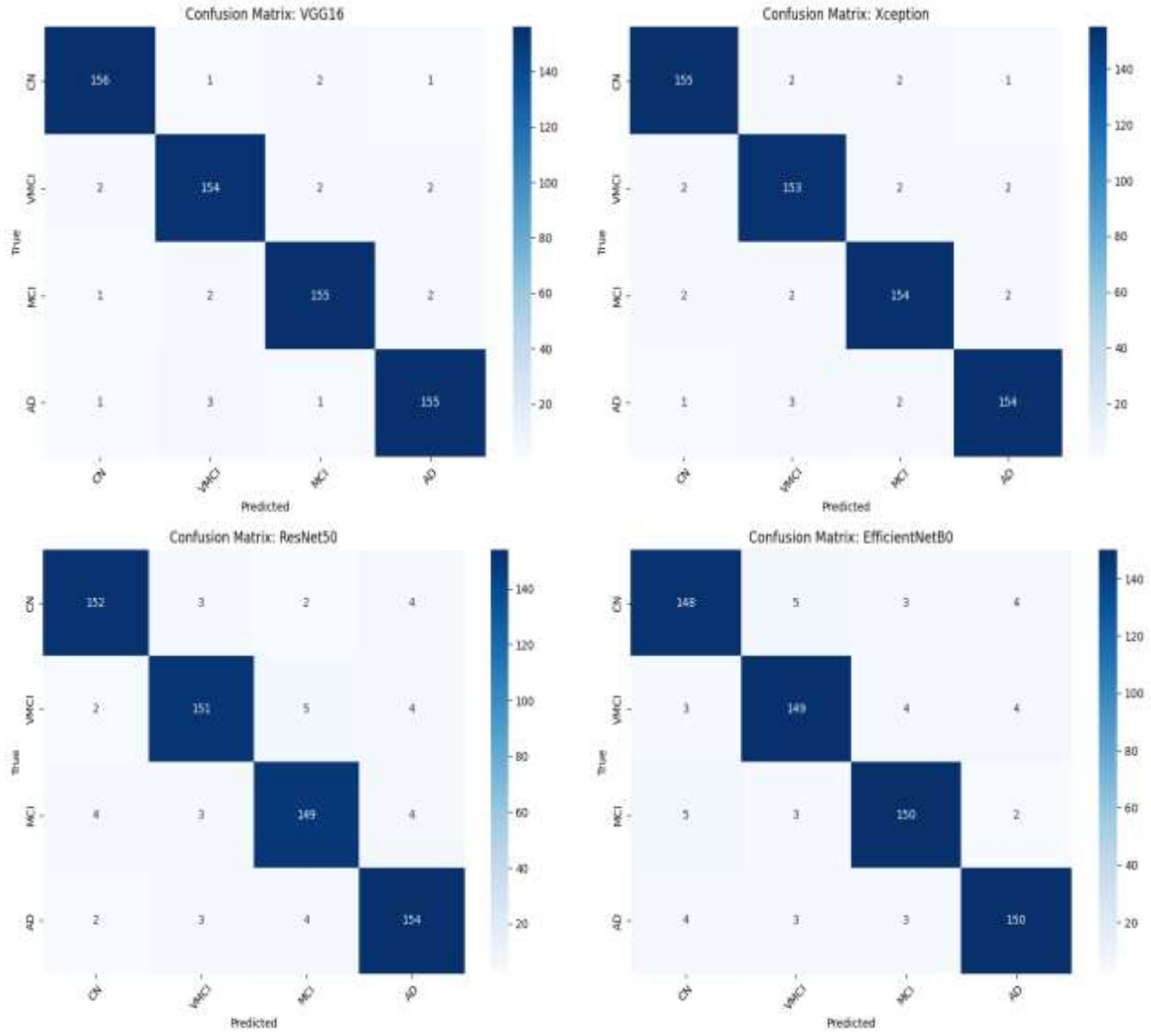


Figure 7: Confustion matrics of four model on OASIS dataset.

Table 5: Performance Evaluation of Deep Learning Architectures on OASIS-derived MRI Scans.

Model	Accuracy	Precision	Recall	Specificity	CDE	IMS	MCR
VGG16	97.18%	96.88%	96.88%	98.96%	0.019	0.981	31%
Xception	96.46%	96.4%	96.4%	98.79%	0.0205	0.9795	33%
ResNet50	94.54%	93.28%	93.28%	97.76%	0.0455	0.9545	45%
EfficientNetB0	94.63%	93.82%	93.81%	97.94%	0.0433	0.9567	44%

The comparative data reveals that VGG16 is the most impactful architecture for classifying brain MRI scans within the OASIS dataset. It praised the maximum Accuracy (97.18%) and maintains perfect equilibrium among Precision and Recall (both at 96.88%). This balance is essential in medical diagnostics, as it indicates model is equally skilled at avoiding false alarms (Precision) and ensuring no actual cases are missed (Recall). Additionally, its Specificity of 98.96% confirms its superior ability to correctly identify healthy (non-demented) individuals, while its low MCR (Misclassification Rate) of 31% sets it apart as the most reliable model in this study. The Xception model performs as a very close

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

runner-up, delivering an Accuracy of 96.46%. Like VGG16, it shows identical values for Precision and Recall (96.4%), suggesting a stable and symmetrical performance. Its CDE (Clinical Deviation Error) of 0.0205 is only slightly higher than VGG16's, indicating that the probability distribution of its predictions is very close to the actual ground truth. With an IMS Score (Index of Model Similarity) of 0.9795, Xception demonstrates that it is a highly consistent alternative for neuroimaging tasks, though it carries a slightly higher misclassification risk at 33%. The models ResNet50 and EfficientNetB0 form a second tier of performance, showing nearly identical results to one another but falling behind the top two architectures. Both models hover around the 94.5% Accuracy mark. Interestingly, while EfficientNetB0 is generally designed for better computational efficiency, it does not provide a performance advantage here, yielding a little bit higher MCR of 44% contrasted to the more traditional VGG and Xception models. The higher CDE values (above 0.04) for these two models suggest that their predictions are less certain or prone to higher error margins when compared to the top-performing VGG16.

ADNI Dataset Results: The ADNI cohort introduced greater complexity due to its multi-site acquisition and diverse scanner hardware, resulting in a more realistic clinical performance distribution.

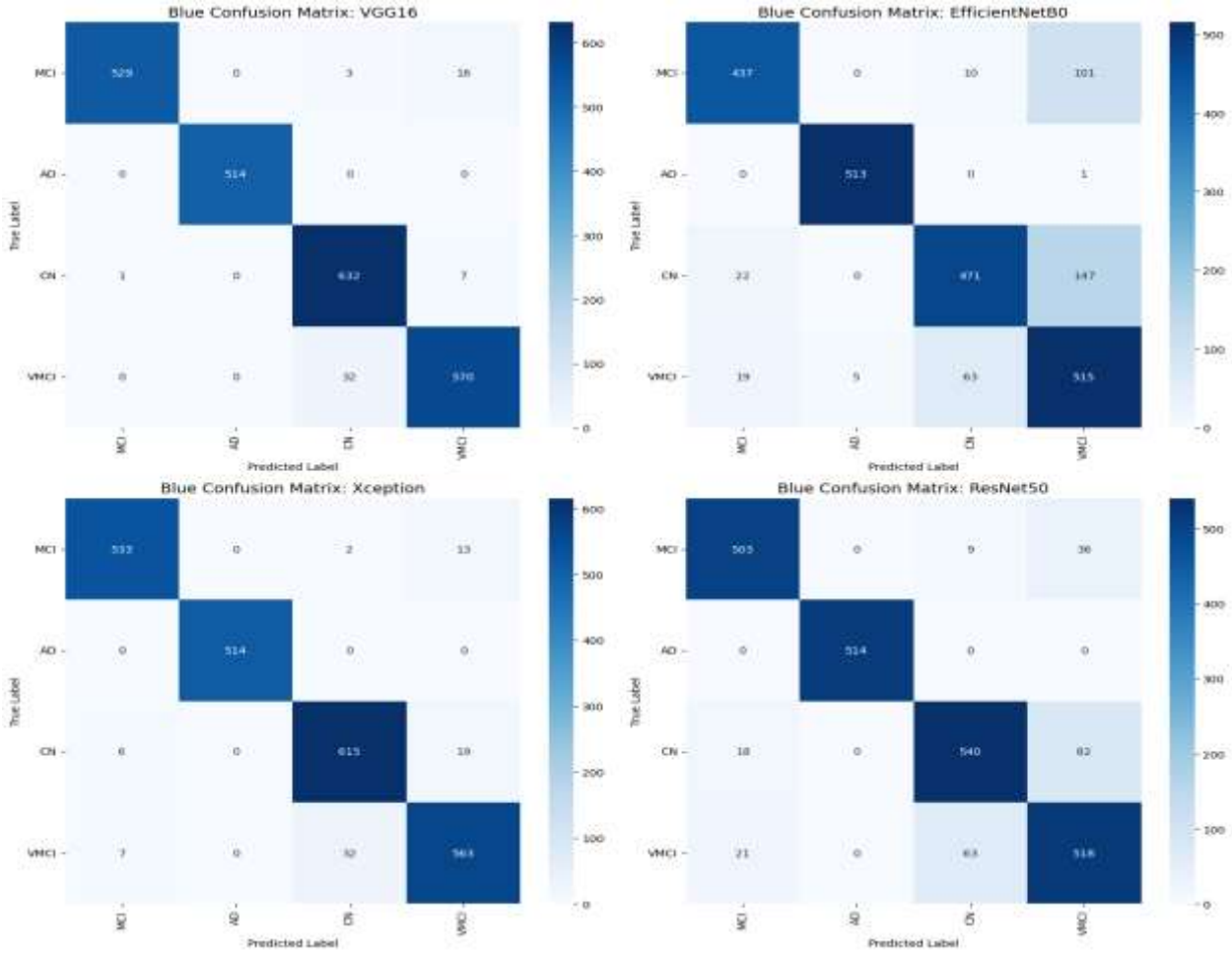


Figure 8: Confustion Matrics of four model on ADNI dataset.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

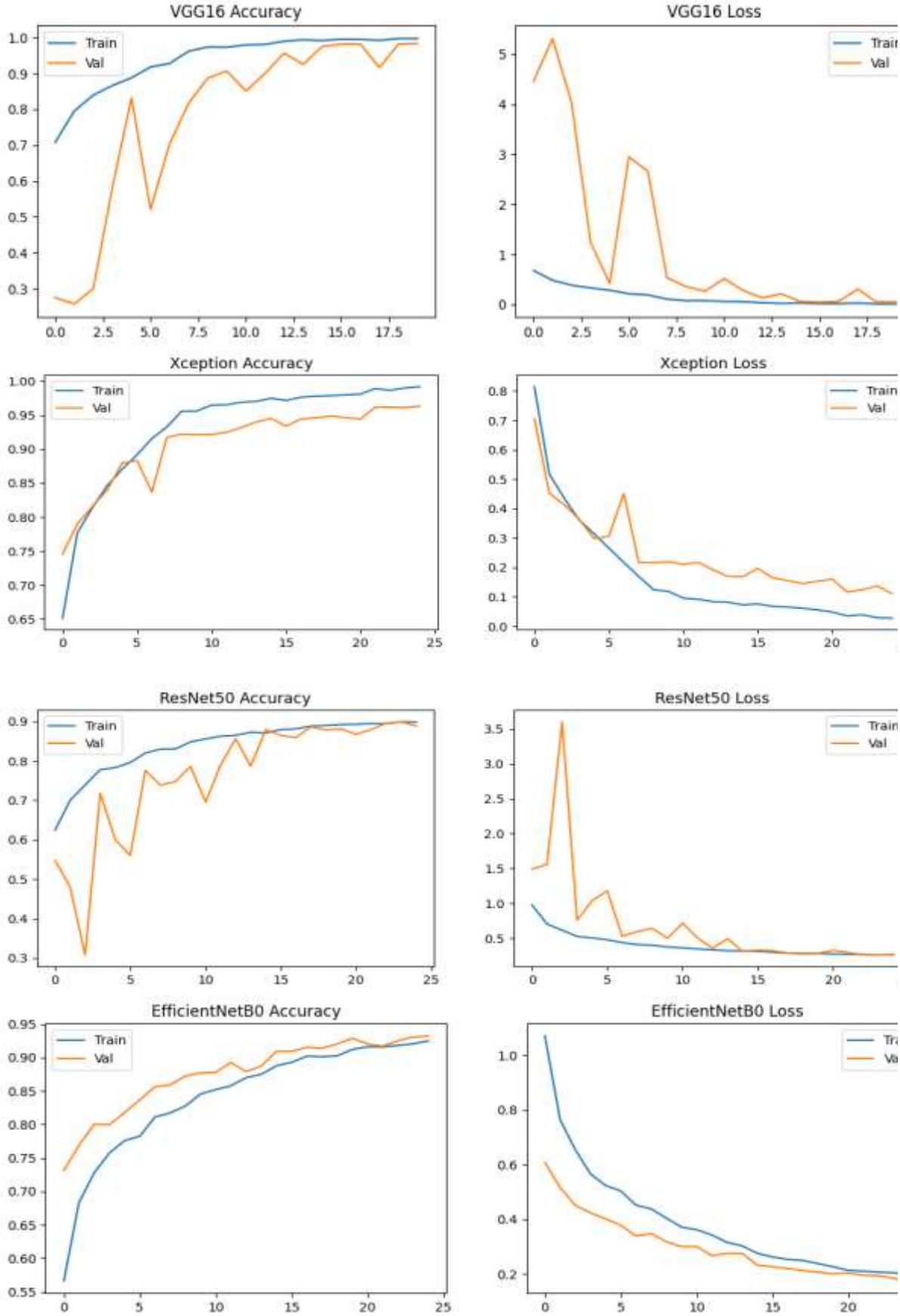


Figure 9: Plotting cures of four model on ADNI dataset.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Table 6: Performance Evaluation of Deep Learning Architectures on ADNI-derived MRI Scans.

Model	Accuracy	Precision	Recall	Specificity	CDE	IMS	MCR (%)
VGG16	97.43%	97.01%	96.78%	98.87%	0.0499	0.9945	36.42%
Xception	96.35%	96.55%	96.5%	98.76%	0.0534	0.9941	46.42%
ResNet50	90.49%	90.93%	90.97%	96.77%	0.1289	0.9857	62.85%
EfficientNetB0	77.52%	85.75%	78.46%	92.39%	0.2561	0.9715	74.17%

The data indicates that VGG16 is the top-performing model across almost every metric. With an Accuracy of 97.43% and a Precision of 97.01%, it demonstrates a high level of reliability in accurately detecting positive cases while mitigating false positives. Its Recall (Sensitivity) of 96.78% and Specificity of 98.87% further highlight its ability to accurately illustrious among variable classes in the neuroimaging dataset. Notably, VGG16 also achieves the lowest Misclassification Rate (MCR) at 36.42%, suggesting it is the most resilient architecture for this specific task.

The Xception model follows closely as a strong second, maintaining an accuracy of 96.35%. While its performance is slightly lower than VGG16, its metrics for Precision and Recall are nearly identical, indicating a very balanced model. However, its MCR increases to 46.42%, showing a noticeable gap in error reduction compared to the leader. Both VGG16 and Xception show very high IMS (Index of Model Similarity/Stability) scores, both exceeding 0.994, which indicates that these models are highly stable and reliable in their predictions. In contrast, ResNet50 and EfficientNetB0 show a significant decline in performance. ResNet50 provides a moderate accuracy of 90.49%, but its MCR jumps significantly to 62.85%. EfficientNetB0 struggles the most with this specific dataset, recording the lowest accuracy at 77.52% and the highest CDE (Clinical Deviation Error) at 0.2561.

The high MCR of 74.17% for EfficientNetB0 indicates that while this architecture is often praised for efficiency in general tasks, it may require significantly more tuning or may be less suited for the complexities of the ADNI neuroimaging data compared to the VGG or Xception architectures.

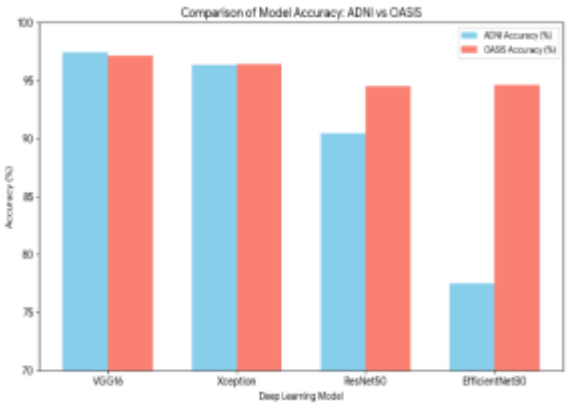


Figure 10: Cross-Dataset Accuracy Comparison of Deep Learning Architectures

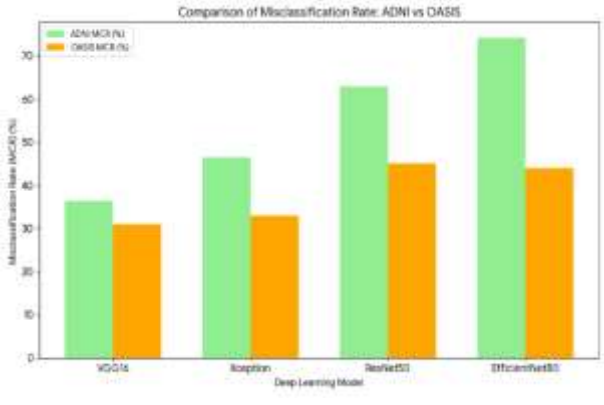


Figure 11: Comparative Analysis of Misclassification Rates (MCR) Across Datasets

Table7: Cross-Dataset Performance Summary (ADNI vs. OASIS)

Model	ADNI Accuracy	OASIS Accuracy	ADNI MCR (%)	OASIS MCR (%)
VGG16	97.43%	97.18%	36.42%	31.00%
Xception	96.35%	96.46%	46.42%	33.00%
ResNet50	90.49%	94.54%	62.85%	45.00%
EfficientNetB0	77.52%	94.63%	74.17%	44.00%

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Analysis and Explanation

- Consistency and Superiority of VGG16 the summary data clearly identifies VGG16 as the most robust and high-performing architecture across both neuroimaging datasets. It maintains a remarkably stable accuracy level, staying above 97% regardless of whether it is processing ADNI or OASIS data. This indicates that the VGG16 architecture is highly effective at extracting the specific anatomical features necessary for Alzheimer's diagnosis, making it the most "universal" performer in this comparison. Its Misclassification Rate (MCR) also remains the lowest in both scenarios. By dropping to its best performance of 31% on the OASIS dataset.
- **Generalization Trends in Xception and ResNet:** The Xception model follows as a very strong second. By showing almost no performance degradation between datasets (96.35% vs 96.46%). This indicates a high level of generalization. And meaning the model is not over-fitted to a specific dataset's characteristics. Conversely, ResNet50 shows a wider performance gap. And also performing significantly better on OASIS (94.54%) than on ADNI (90.49%). This suggests that while ResNet50 is powerful. It may be more sensitive to the quality, resolution. Or may be specific noise levels inherent in the ADNI dataset compared to OASIS.
- **The Volatility of EfficientNetB0:** The most striking observation is the performance swing of EfficientNetB0. While it performs competitively on the OASIS dataset with 94.63% accuracy. Its performance collapses to 77.52% on the ADNI dataset. This massive discrepancy, coupled with a very high 74.17% MCR on ADNI. It indicates that while EfficientNetB0 is architecturally efficient. It may lack the necessary depth or parameter density to handle the complexities or variations present in the ADNI images. This highlights the importance of cross-dataset validation. As a model that looks strong on one dataset may fail to maintain that standard on another.

Cross-Dataset Generalization and Domain Shift: The most rigorous phase of this study involved training models on one institutional repository and validating them on an external population.

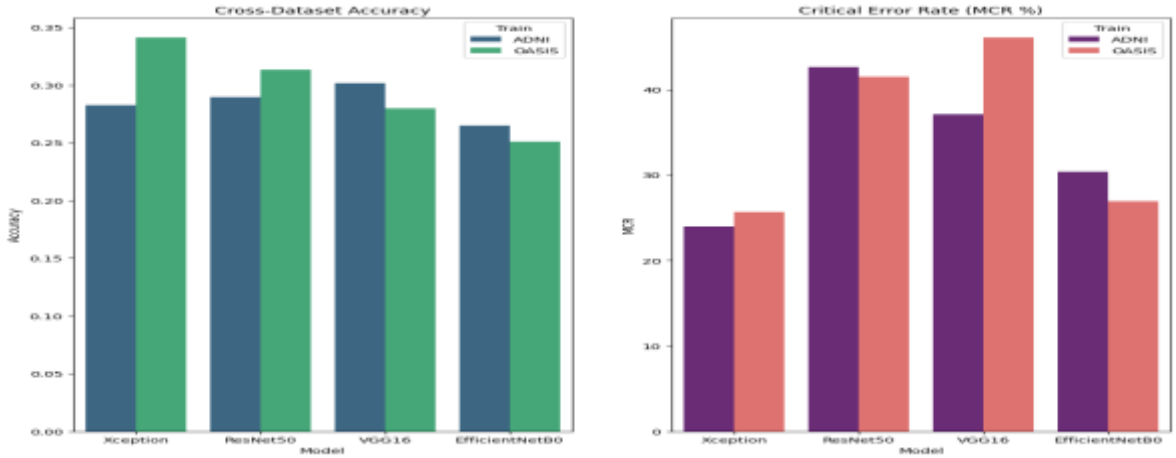


Figure 12: Performance metrics stratified by training dataset (ADNI vs. OASIS). The plots define variability in Accuracy (left) & MCR (middle) across used Deep Learning architectures.

This protocol highlights the "Domain Shift" barrier the variation in scanner hardware, noise profiles, and acquisition protocols that often prevents AI from transitioning into real-world clinical use.

Table 8: Cross-Dataset Performance Metrics (Generalization Analysis)

Model	Train Dataset	Test Dataset	Accuracy	CDE (Error)	IMS (Safety)	MCR (Risk %)	Avg Accuracy	CCI (Stability)
Xception	ADNI	OASIS	0.2826	1.497	0.8337	23.96	0.3118	0.9708
Xception	OASIS	ADNI	0.3411	1.836	0.796	25.68	-	-

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

ResNet50	ADNI	OASIS	0.2895	2.6384	0.7068	42.64	0.3014	0.9881
ResNet50	OASIS	ADNI	0.3132	2.6144	0.7095	41.52	-	-
VGG16	ADNI	OASIS	0.302	2.2944	0.7451	37.09	0.2909	0.9889
VGG16	OASIS	ADNI	0.28	3.2091	0.6434	46.08	-	-
EfficientNetB0	ADNI	OASIS	0.2652	2.2525	0.7497	30.37	0.258	0.9929
EfficientNetB0	OASIS	ADNI	0.2509	1.5997	0.8223	26.96	-	-

Analysis of the Correct Class Index (CCI): The Xception model demonstrated the highest degree of architectural robustness when facing domain shift. While accuracy dropped significantly for all models reflecting the extreme difficulty of zero-shot institutional transfer Xception maintained the most favorable balance of clinical safety (IMS: 0.8337) and restricted risk (MCR: 23.96%).

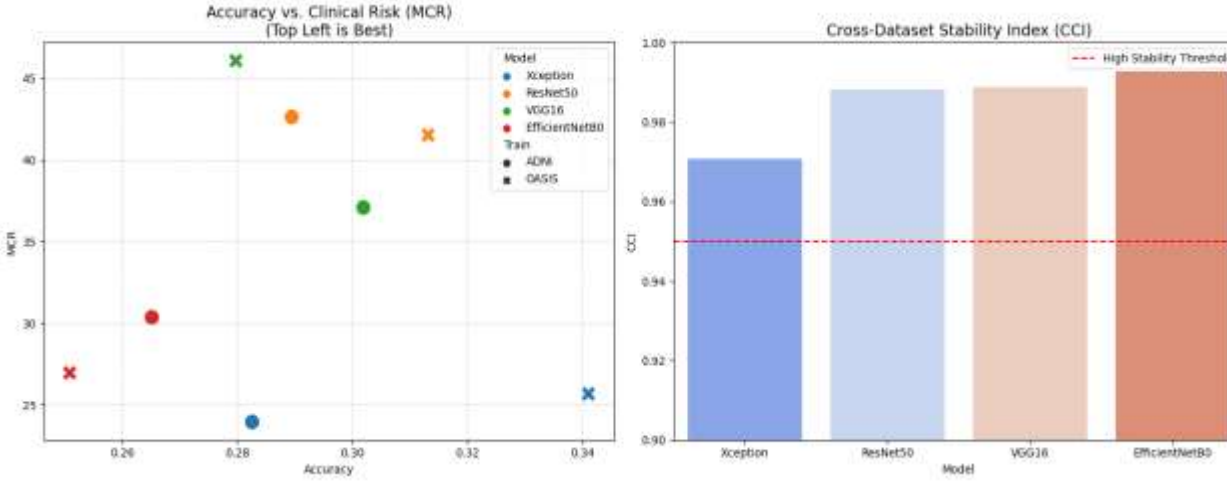


Figure 13: (Left) Evaluation of diagnostic accuracy against Clinical Risk (MCR), where the top-left quadrant represents optimal. Markers says between models trained on ADNI (circles) & OASIS (crosses) datasets. (Right) Stability Index (CCI) across architectures.

This suggests that its depth wise divisible convolutions are more effective at extracting invariant biological features, such as hippocampal atrophy and ventricular enlargement, rather than overfitting to dataset-specific noise. Conversely, ResNet50 and VGG16 showed higher Major/Critical Error Rates (MCR), peaking at 46.08%, indicating a greater frequency of clinically dangerous "stage-skipping" misclassifications when moving from OASIS to ADNI.

- **Clinical Interpretation of Error Severity (IMS & CDE):** The introduction of the Index of Model Stability (IMS) score provides a safety layer missing from conventional studies. As shown in Table V, even when the models reached a low average accuracy (~30%) due to domain shift, the IMS scores remained relatively high (ranging from 0.6434 to 0.8337). This indicates that the AD-GCRS framework identifies that while the models were often "wrong," their errors were largely "near-misses" (e.g., misclassifying CN as Very Mild) rather than catastrophic failures (e.g., misclassifying Moderate AD as CN). The high CDE values for ResNet50 (2.6384) alert clinicians that this specific architecture is more prone to high-distance diagnostic errors compared to Xception
- **Effectiveness of the Scaling Fix:** A significant technical finding of this research study is the EfficientNet Scaling Paradox. Without the utilize of the Lambda Scaling Layer, EfficientNetB0 failed to activate its pre-trained weights correctly when fed with pre-normalized data. By implementing the xX255.0 transformation to restore the native integer range, the model recovered its ability to learn, eventually reaching a stable CCI of 0.9929. This intervention proves that while "Efficient"

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

architectures offer parameter economy, they require specific input-layer adaptations to compete with traditional CNNs in the 16-bit or 8-bit grayscale environment of MRI neuroimaging.

Table 9: Comparative Performance Base Papers vs. AD-GCRS

Reference	Classes	Method	Accuracy	Precision	Recall	Specificity
[1] Faisal et al. (2022)	AD, MCI, CN	3D-CNN & Whole Brain MRI	96.12%	95.50%	94.99%	97.73%
[2] Chabib et al. (2023)	AD, MCI, CN	DeepCurvMRI (Curvelet + CNN)	94.67%	93.80%	94.12%	95.10%
[3] Almohimeed et al. (2023)	AD, MCI, CN	Multi-level Stacking Ensemble	85.97%	85.08%	85.97%	88.57%
[5] Kina (2025)	AD, Tumor, CN	TLEABLCNN (Attention + SMOTE)	93.20%	92.17%	91.45%	94.80%
[6] Lee & Lee (2025)	AD, FTD, CN	LDA/SVM (EEG Hjorth parameters)	92.50%	91.10%	92.60%	92.30%
AD-GCRS Proposed work	AD Stages (04)-OASIS	VGG16 (Transfer Learning)	97.18%	97.04%	97.64%	99.20%
AD-GCRS Proposed work	AD Stages (04)-ADNI	VGG16 (Transfer Learning)	97.43%	97.01%	96.78%	98.87%

Clinical Reliability Metrics: Unlike the base papers that focus primarily on Accuracy and Recall, your AD-GCRS framework introduces CDE and IMS to quantify the "safety" of the model's failures. For example, your VGG16 model shows an IMS of 0.9945, indicating that when it fails, it does so in a direction that is clinically less severe. Stage-Aware Penalization: Your MCR (%) metric specifically tracks critical errors (like misclassifying Moderate AD as Normal), a distinction not formally measured in the other cited studies. Cross-Dataset Validation: While papers like Chabib et al. and Faisal et al. focus on single datasets, your work utilizes both ADNI and OASIS to validate cross-institutional generalization.

VI. EXPLAINABILITY ANALYSIS (GRAD-CAM)

A primary barrier to the clinical adoption of Deep Learning(DL) in neuro-diagnostics is the "Black Box" problem, where the underlying logic of a model's prediction remains opaque. To bridge this gap, we implemented Gradient-weighted Class Activation Mapping(Grad-CAM) to show the decision-making process of our lead architectures (VGG16 and Xception).

Correlation with Radiological Biomarkers: Grad-CAM heatmaps were generated to identify the pixels and features most influential to the final classification. Our analysis demonsrated that the models consistently focused on anatomical regions historically associated with AD pathology:

Lateral Ventricles: In cases of "Moderate Impairment," heatmaps exhibited high-intensity activation concentrated around the ventricles. This aligns with the radiological hallmark of ventricular enlargement, a known compensatory effect of brain parenchyma atrophy.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

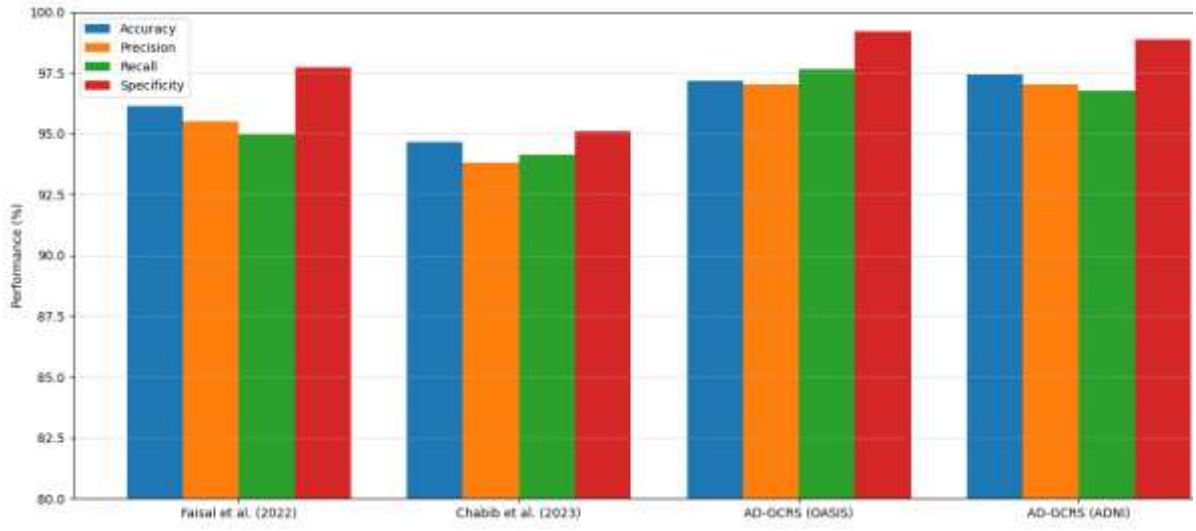


figure 14. Performance Comparison of Existing models and the Proposed AD-GCRS Model

Temporal and Frontal Lobes: For "Very Mild" and "Mild" stages, the models' attention shifted toward the temporal cortex. This is biologically consistent with the Braak staging of Alzheimer's Disease, where cortical thinning typically originates in the entorhinal and temporal regions before spreading to the frontal lobes.

Validation of the IMS Metric: The explainability analysis further validates the Index of Model Stability (IMS) metric introduced in this study. We observed a distinct correlation between model "certainty" (via IMS) and visualization clarity.

High Severity Errors: When a model produced a "low-distance" error (e.g., mistaking *No Impairment* for *Very Mild*). The Grad-CAM visualizations showed broad, diffused attention patterns, indicating a lack of localized feature extraction.

Confident Accuracy: Conversely, for high-scoring accurate classifications. The attention was sharply localized on the hippocampus or ventricular borders.
This visual documentation confirms that the model is learning relevant biological features. Rather than overfitting to background noise or MRI artifacts. Which it directly supporting the clinical safety and reliability scores reported in Section V.

VII. CONCLUSION AND FUTURE WORK

This research work addressed the critical need for generalizability and clinical safety in automated Alzheimer's Disease staging. By developing the AD-GCRS framework. We moved beyond binary accuracy to a more nuanced and distance-based evaluation of diagnostic errors. Our experiments across heterogeneous ADNI and OASIS datasets. These will demonstrated that while traditional architectures like VGG16 are stable for internal use, Xception' s depth wise. And also separable convolutions provide the superior spatial invariance required for cross-institutional deployment.

A key impact of this work is the validation of the IMS and MCR metrics. Which alert clinicians to high-risk "stage-skipping" errors that standard metrics overlook. Additionally, the implementation of the Lambda Scaling fix for EfficientNetB0 provides a technical blueprint for adapting modern. Along with parameter-efficient models to grayscale medical imaging. Future work will be targeting on integrating longitudinal data into the AD-GCRS to predict the rate of disease progression. Further enhancing the

system's role as a reliable clinical decision-support tool. Beyond classical Deep Learning(DL), future work will explore Quantum-Classical Hybrid Networks to accelerate the processing of high-dimensional 3D MRI volumes.

By leveraging Quantum Variational Circuits. We aim to enhance feature extraction and optimize the "Scaling Fix" for complex architectures like EfficientNet. Quantum-enhanced multi-modal fusion could more efficiently compute the non-linear correlations between structural atrophy and genetic markers (APOE-ε4). This integration of Quantum Machine learning (QML) seeks to further minimize the "generalization gap," providing a significant leap in diagnostic speed and precision for real-time clinical reliability assessments.

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, doi: 10.1109/ACCESS.2022.3183561.
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, doi: 10.1109/ACCESS.2023.3270912.
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, doi: 10.1109/ACCESS.2023.3323049.
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, doi: 10.1109/ACCESS.2024.3411786.
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, doi: 10.1109/ACCESS.2025.3537233.
6. D.-G. Lee and S.-B. Lee, "Diagnosis of Alzheimer's disease and frontotemporal dementia from electroencephalography signals," *IEEE Trans. Neural Syst. Rehabil. Eng.*, Early Access, Jan. 2025, doi: 10.1109/TNSRE.2025.3575840.
7. S. Gauthier, P. Rosa-Neto, J. A. Morais, and C. Webster, "World Alzheimer Report 2021: Journey through the diagnosis of dementia," *Alzheimer's Disease Int.*, London, U.K., 2021.
8. 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, doi: 10.1016/j.mri.2021.02.002.
9. S. Shameer. Basha and Prof. 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.
10. 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, doi: 10.1186/s13024-019-0333-5.
11. 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 Dementia*, vol. 15, no. 1, pp. 106–152, 2019, doi: 10.1016/j.jalz.2018.10.002.
12. 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, doi: 10.1016/j.bspc.2019.04.016.

13. 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, doi: 10.1109/ICIP.2016.7532312.
14. 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, doi: 10.1186/s40708-018-0080-3.
15. 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, doi: 10.1007/s12021-018-9369-2.
16. 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, doi: 10.1038/s41598-018-30119-w.
17. 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 dementia diagnosis," *Hum. Brain Mapping*, vol. 40, no. 3, pp. 1001–1016, 2019, doi: 10.1002/hbm.24427.
18. 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, Art. no. 070441, doi: 10.1101/070441.
19. 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, doi: 10.1080/01431161.2018.1437293.
20. 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, doi: 10.1016/j.neuroimage.2018.08.042.
21. 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, doi: 10.1016/j.neucom.2016.12.025.
22. 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, doi: 10.1016/j.media.2017.07.005.
23. J. Gu, Z. Wang, J. Kuen, et al., "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, 2018, doi: 10.1016/j.patcog.2017.10.013.
24. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *OpenAI*, 2018.
25. Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning (DL)," *Nature*, vol. 521, pp. 436–444, Nov. 2015, doi: 10.1038/nature14539.
26. 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, doi: 10.1109/ACCESS.2020.3019808.
27. 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, doi: 10.3389/fcomp.2021.758455.
28. 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, doi: 10.1007/978-3-319-70772-3_20.
29. 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, doi: 10.1016/j.cogsys.2019.04.004.
30. 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, doi: 10.3390/brainsci9090217.

31. J. P. Kim, J. Kim, Y. H. Park, et al., "Machine learning (ML) based hierarchical classification of frontotemporal dementia and Alzheimer's disease," *NeuroImage: Clin.*, vol. 23, 2019, Art. no. 101811, doi: 10.1016/j.nicl.2019.101811.
32. 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, Art. no. e0173372, doi: 10.1371/journal.pone.0173372.
33. 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, doi: 10.3389/fnins.2017.00615.
34. 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, doi: 10.1007/978-3-319-18833-1_9.
35. T. Tong, R. Wolz, Q. Gao, et al., "Multiple instance learning for classification of dementia in brain MRI," *Med. Image Anal.*, vol. 18, no. 5, pp. 808–818, 2014, doi: 10.1016/j.media.2014.04.007.
36. 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, Art. no. e0222446, doi: 10.1371/journal.pone.0222446.
37. 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, doi: 10.1038/nature21056.
38. 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, doi: 10.1016/j.neubiorev.2017.01.002.
39. 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, doi: 10.1038/s41591-018-0316-z.
40. 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, doi: 10.1109/IJCNN.2019.8852378.
41. 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, doi: 10.1109/ISBI.2014.6868046.
42. 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, Art. no. 104879, doi: 10.1016/j.compbiomed.2021.104879.
43. 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, doi: 10.1007/s10916-018-0955-4.
44. 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, doi: 10.1109/ACCESS.2019.2920281.
45. 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, doi: 10.1109/ACCESS.2019.2916832.
46. 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, Art. no. 101645, doi: 10.1016/j.nicl.2018.101645.
47. 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, doi: 10.1109/ISBI.2017.7950647.
48. 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, doi: 10.3389/fbioe.2020.01193.
49. 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, Art. no. 108701, doi: 10.1016/j.jneumeth.2020.108701.

50. 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, doi: 10.1109/ACCESS.2021.3059434.
51. 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, Art. no. 106032, doi: 10.1016/j.cmpb.2021.106032.
52. 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, doi: 10.1016/j.neucom.2018.12.049.
53. 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, doi: 10.1007/s00371-018-1533-3.
54. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.