Deep Transfer Learning Framework for Alzheimer's Disease Classification with Explainable AI Insights

Supervised By

Ms. Raiyan Janik Monir
Lecturer

Department of Software Engineering

Daffodil International University

Submitted By

Sadia Tasnim Prema
ID:221-35-1073

Department of Software Engineering

Daffodil International University

This thesis report has been submitted in fulfilment of the requirements for the Degree of Bachelor of Science in Software Engineering.

APPROVAL

APPROVAL

This thesis titled on "Deep Transfer Learning Framework for Alzheimer's Disease Classification with Explainable AI Insights", submitted by Sadia Tasmin Prema (ID: 221-35-1073) to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

Dr. Fazla Elahe
Assistant Professor & Associate Head
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Chairman

Dr. Marzia Ahmed
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 1

Dr. Shabnom Mustary
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 2

Md. Rajib Mia
Lecturer (Senior Scale)
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 3

Mohammod Abul Kashem, PhD
Professor
Department of Computer Science and Engineering
DUET, Bangladesh

External Examiner

Deep Transfer Learning Framework for Alzheimer's Disease Classification with Explainable AI Insights

Sadia Tasnim Prema
ID:221-35-1073

Bachelor of Science

DAFFODIL INTERNATIONAL UNIVERSITY

DAFFODIL INTERNATIONAL UNIVERSITY

DECLARATION OF THESIS AND COPYRIGHT

Author's Full Name	:	Sadia Tasnim Prema
Date of Birth	:	08/10/2002
Title	:	Deep Transfer Learning Framework for Alzheimer's Disease Classification With Explainable AI Insights
Academic Session	:	2022-2025

I declare that this thesis is classified as:

☒	CONFIDENTIAL	(Contains confidential information under the Official Secret Act 1997)*
☐	RESTRICTED	(Contains restricted information as specified by the organization where research was done)*
☐	OPEN ACCESS	I agree that my thesis to be published as online open access (Full Text)

I acknowledge that Daffodil International University reserves the following rights:

1. The Thesis is the Property of Daffodil International University.
2. The Library of Daffodil International University has the right to make copies of the thesis for the purpose of research only.
3. The Library of Daffodil International University has the right to make copies of the thesis for academic exchange.

Certified by:

(Student's Signature)

221-35-1073

Student ID
Date:27/11/2025

(Supervisor's Signature)

Ms.Raiyan Janik Monir

Name of Supervisor
Date: 27/11/2025

SUPERVISOR'S DECLARATION

I hereby declare that I have reviewed this thesis entitled "**Deep Transfer Learning Framework for Alzheimer's Disease Classification with Explainable AI Insights"**, and in my opinion, it is adequate in terms of scope and quality for the award of the degree of Bachelor of Science in Software Engineering.

(Supervisor's Signature)

Full Name : Ms. Raiyan Janik Monir

Position : Lecturer, Department of SWE, DIU

Date : 26 November 2025

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Daffodil International University or any other institution.

(Student's Signature)

Full Name : Sadia Tasnim Prema

ID Number : 221-35-1073

Date : 26 November 2025

Deep Transfer Learning Framework for Alzheimer's Disease Classification with Explainable AI Insights

Sadia Tasnim Prema

ID: 221-35-1073

Thesis submitted in fulfillment of the requirements
for the award of the degree of
Bachelor of Science

Department of Software Engineering (Major in Data Science)

DAFFODIL INTERNATIONAL UNIVERSITY

DECEMBER 2025

ACKNOWLEDGEMENTS

I would like to take this opportunity to list my heartfelt thanks for the help and encouragement from all the people who helped me finish this thesis titled “Deep Transfer Learning Framework for AD Classification with Explainable AI Insights”. I am indebted primarily to Ms. Raiyan Janik Monir, Lecturer, Department of Software Engineering, Daffodil International University (DIU) for her supervision, guidance and encouragement during the work on this study. Her valuable suggestions, guidance and expertise have been absolutely vital to the success of this work. I am also thankful to the Department of Software Engineering, Daffodil International University for providing a conducive academic atmosphere, facilities and resources which enabled me to carry out this research. I owe a heartfelt debt to all my friends and colleagues for their valuable comments, cooperation and encouragement throughout the research work. Last but not least, my sincere thanks go to my dear family members for their whole-hearted love, moral support and encouragement. It’s that unwavering faith in me that has given me the strength to carry on and to push myself forward when the going gets tough.

DEDICATION

This work is dedicated to my beloved parents, whose boundless love, support and sacrifices have formed the bedrock of my accomplishment. Their undying support, prayers and believing in my capabilities pushed me to be at my best despite the obstacles. Dedication I further dedicate this thesis to my dear supervisor, Ms. Raiyan Janik Monir, for her sincere guidance, support and priceless supervision during the period of this research. Without their love, support and inspiration this would not have been possible.

ABSTRACT

Alzheimer's disease (AD) is a progressive neurodegenerative disorder that requires early and reliable diagnosis, especially in low-resource settings. This study proposes an explainable deep learning framework for multi-class classification of AD stages from MRI brain images. A publicly available four-class Mendeley MRI dataset (Non-Demented, Very Mild, Mild, and Moderate-Demented) was preprocessed through resizing, normalization, augmentation, and an 80/10/10 train–validation–test split. Four transfer learning models GoogLeNet, DenseNet121, ResNet101, and VGG16 were fine-tuned using the Adam optimizer and evaluated with accuracy, precision, recall, F1-score, AUC, and mean categorical hinge loss. Among all architectures, GoogLeNet achieved the best performance with 0.98 accuracy, macro-F1 of 0.98, AUC of 1.00, and the lowest hinge loss (0.048), clearly outperforming the other models. Explainable AI techniques, Grad-CAM and LIME, were applied to highlight discriminative brain regions, consistently focusing on clinically relevant structures such as the hippocampus, temporal lobe, and ventricles. These pictorial elucidations uphold the clinical suitability of the model's decisions. All in all, the suggested GoogLeNet based architecture is a mixture of high strong interpretable diagnostic accuracy, proving that it can be a assistance screening device in MRI-based AD detection under data constrained clinical.
such environments as Bangladesh.

Keywords: ResNet, GoogleNet, DenseNet, Alzheimer's disease, explainable AI.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

$H_l(.)$	the transition function
W_f	Weight factor
b_f	Bias factor
$L^c{}_{Grad-CAM}$	Final Grad-CAM heatmap
$ReLU(x)$	Rectified Linear Unit activation function
$\triangle$	margin parameter
w_{ij}	Weight connecting neuron i to neuron j
L	Loss function
$\pi_x(z)$	Proximity measure between instance x and perturbed sample z
$\Omega(g)$	Complexity penalty in LIME
α_k^c	weight for the *k-th* feature map for class *c*
$A_{ij}{}^k$	Activation at location (i, j) in feature map k

LIST OF ABBREVIATIONS

AD	Alzheimer's Disease
AI	Artificial Intelligence
ANN	Artificial Neural Network
AUC	Area Under the Curve
BD	Brain Disorder
CAD	Computer-Aided Diagnosis
CNN	Convolutional Neural Network
CSF	Cerebrospinal Fluid
CT	Computed Tomography
DL	Deep Learning
FN	False Negative
FP	False Positive
FPR	False Positive Rate
F1-score	F1 Measure (Harmonic Mean of Precision and Recall)
GAN	Generative Adversarial Network
GPU	Graphics Processing Unit
Grad-CAM	Gradient-weighted Class Activation Mapping
HPC	High-Performance Computing
KNN	K-Nearest Neighbour
LIME	Local Interpretable Model-Agnostic Explanations
MAE	Mean Absolute Error
ML	Machine Learning
MRI	Magnetic Resonance Imaging
MSE	Mean Squared Error
NN	Neural Network
PET	Positron Emission Tomography
PPV	Positive Predictive Value
ReLU	Rectified Linear Unit
ResNet	Residual Network
ROC	Receiver Operating Characteristic
SD	Standard Deviation

SGD	Stochastic Gradient Descent
SNR	Signal-to-Noise Ratio
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
TNR	True Negative Rate
TPR	True Positive Rate
VGG	Visual Geometry Group Network
XAI	Explainable Artificial Intelligence

LIST OF APPENDICES

INTRODUCTION

1.1 Background of the Study

Alzheimer's disease (AD) is a progressive neurodegenerative disorder that gradually destroys memory, cognitive function, and independence, and is the most common cause of dementia worldwide[1], [2]. Globally, more than tens of millions of people live with dementia, and this number is projected to rise sharply as populations age, creating a major health, social, and economic burden. In countries with low and middle income levels, health systems often lack specialized services, neuroimaging facilities, and national dementia policies, which widens the gap in early detection and care[3]. Bangladesh is undergoing rapid demographic aging, with an increasing number of older adults and an estimated dementia burden already in the millions, yet public awareness remains low, specialist care is concentrated in urban centers, and formal Alzheimer's services are scarce[4],[5]. Many patients are diagnosed late or remain undiagnosed, and their symptoms are often attributed to "normal aging," which increases caregiver stress and delays appropriate management[6]. Financial constraints, stigma, and limited trained manpower further restrict access to comprehensive dementia services in the country[7],[8]. At the same time, advances in AI, especially DL and TL applied to MRI brain images, offer new opportunities for earlier, more objective, and more scalable diagnosis. Explainable AI techniques such as Grad-CAM and LIME can further help clinicians trust and understand model decisions, making AI-based tools particularly attractive for resource-constrained settings like Bangladesh[9],[10].

1.2 Problem Statement

Although the problem of AD is increasing both globally and in Bangladesh, the disease is challenging to diagnose at an early and reliable stage. Traditional diagnosis is based on the expert interpretation of neuroimaging, neuropsychological tests, and clinical judgments which may be time-consuming, subjective and very reliant on the presence of the specialists. These problems are compounded by the lack of access to neurologists, neuroimaging centers, and formal memory clinics in rural areas, particularly not in big cities. In recent research, it was demonstrated that deep learning models (with transfer learning, especially CNN are highly accurate in classifying AD based on MRI data. Nevertheless, most current models are trained using Western-centric data, and are not strictly tested in multi-class scenarios, and are typically black boxes with no clear description of their predictions at all. Such inability to be interpreted and contextually validated restricts clinical trust and adoption, particularly within low-resource settings. It is thus urgently needed that an explainable, high-performing, and

computationally efficient deep learning architecture to multi-class AD staging with the ability to run on publicly available MRI data and be scaled to the Bangladeshi healthcare setting.

1.3 Aim and Objectives

Aim

This study will primarily aim to design and test an explainable deep learning system, using transfer learning structures, to classify the stages of AD accurately, interpretably and computationally efficiently using MRI images of the brain, with a view to potentially use them in resource limited environments, like Bangladesh.

Specific Objectives

To achieve this aim, the study pursues the following objectives:

1. To collect, pre-process, and utilize a large and balanced four-class MRI dataset from the Mendeley Data repository for AD classification.

2. To implement and compare four transfer learning architectures: GoogLeNet, DenseNet121, ResNet101, and VGG16 under consistent pre-processing, training, and evaluation settings.
3. To evaluate each model using key performance metrics, including accuracy, precision, recall, F1-score, AUC, and mean categorical hinge loss.
4. To apply explainable AI techniques, specifically Grad-CAM and LIME, to visualize model attention regions and enhance the interpretability of MRI-based predictions.
5. To identify the model that offers the best trade-off between diagnostic accuracy, computational efficiency, and clinical transparency for potential use in real-world healthcare environments, particularly in Bangladesh.

1.4 Research Scope

The study targets structural MRI brain images in the multi-classification of the Alzheimer disease into four levels, Non-Demented, Very Mild Demented, Mild Demented, and Moderate Demented. The experiment applies an openly available, labelled MRI set of Mendeley Data, which has original and augmented images available sorted by disease stage. The dataset is partitioned into training, validation, and test subsets following an 80:10:10 split, and image preprocessing includes resizing, normalization, and augmentation.

On the modelling side, the scope is limited to four widely used transfer learning architectures: GoogleNet, DenseNet-121, ResNet-101, and VGG-16, all implemented in a 2D CNN framework. The models are fine-tuned using the Adam optimizer and evaluated via standard classification metrics and mean categorical hinge loss. Explainability is addressed through Grad-CAM and LIME visualizations of model decisions. The study does not include other imaging modalities (e.g., PET, fMRI), non-imaging clinical data, or longitudinal progression modelling, and it does not propose any treatment or intervention strategies.

1.5 Research Contribution

This study makes several present to the fields of medical imaging, artificial intelligence, and explainable DL for AD. First, it provides a Systematic and fair comparison of four popular transfer learning architectures—GoogLeNet, DenseNet121, ResNet101, and VGG16—using the same dataset, pre-processing pipeline, training protocol, and evaluation metrics, thereby establishing reproducible benchmarks for four-class AD classification from MRI.

Second, the study proposes a GoogLeNet-based framework that achieves high diagnostic performance, including very high accuracy, macro-F1, and AUC, while maintaining relatively low mean categorical hinge loss, demonstrating strong margin separation between AD stages. Third, it integrates explainable AI techniques, Grad-CAM and LIME, to generate complementary visual explanations that highlight clinically relevant brain regions such as the hippocampus, temporal lobe, and ventricles. This strengthens the clinical plausibility of model decisions and supports trust in AI-assisted diagnosis.

Finally, the work situates these technical findings within the context of Bangladesh and other low-resource settings, emphasizing the potential of transfer learning and explainable AI to support early AD screening where specialist availability and infrastructure are limited. By focusing on publicly accessible data and computationally feasible architectures, the thesis contributes a practical and adaptable framework that can inform future implementations and policy discussions on AI-driven dementia care in developing countri

1.6 Structure of the Thesis

This study is structured into five principal chapters. The first chapter introduces the background of AD, outlines the research problem, defines the study's aims and objectives, describes the scope of the work, highlights the research contributions, and provides an overview of the overall organization of the thesis.It introduces the motivation for using deep learning and explainable AI, and explains why this research is important, especially for a resource-limited context like Bangladesh.

Chapter 2 reviews previous work related to AD diagnosis using MRI, DL, TL, and XAI. It summarizes key papers, highlights their main findings and reported performance values, and identifies existing gaps and limitations in the literature. This chapter shows how the present study is related to earlier research and where it offers something new.

Chapter 3 shows the methodology used in the study. It explains the dataset and pre-processing steps, the train–validation–test splitting strategy, and the design of the transfer learning models (GoogLeNet, DenseNet121, ResNet101, VGG16). It also presents the training setup, the evaluation metrics (Accuracy, F1-Score, AUC, hinge loss), and the use of Grad-CAM and LIME for model explanation.

Chapter 4 reports and discusses the experimental results, including quantitative performance of each model, confusion matrices, ROC curves, hinge loss behavior, and visual explanations. It compares the models, identifies the best-performing architecture, and discusses the implications in the Bangladesh context. Chapter 5 then provides an overview of the study, summarizes the main conclusions, states the limitations, and suggests directions for future work and possible real-world application of the proposed framework.

LITERATURE REVIEW

1.7 Overview

AD is a gradually neurodegenerative disorder where early and accurate diagnosis is essential for proper management and planning. Neuroimaging, especially MRI and PET, is widely used to detect structural and functional changes in the brain related to different stages of the disease. Earlier computer-aided systems mainly depended on handcrafted features and traditional classifiers, which required expert knowledge and complex pre-processing. With the development of DL, researchers have shifted toward CNN, residual architectures, transformers, and graph-based models, often supported by transfer learning from large-scale image datasets. Recently, there has also been growing interest in explainable AI, where visual interpretation methods are used so that model decisions become more transparent and clinically acceptable.

1.8 Previous Work

One study used an ensemble of DCNN on the OASIS structural MRI dataset to classify four dementia stages (non-demented, very mild, mild, moderate). The model combined several CNN branches with dense connectivity and cost-sensitive training to handle class imbalance. The final ensemble achieved about 93% accuracy, with average precision and recall around 94% and 93%, respectively, showing that deep ensembles can work well even with limited data[11] .

Another work designed a 3D CNN classifier on the ADNI MRI dataset to separate AD, mild cognitive impairment, and healthy controls. After careful pre-processing and slice selection, the network reached about 94.97% Accuracy for AD vs. normal and 91.98% accuracy for AD vs. MCI, outperforming traditional machine learning baselines. These results highlight the benefit of 3D feature learning for subtle brain changes[12].

A lightweight densely connected CNN focusing on hippocampal patches was proposed to reduce model size while keeping good performance. By using dense connections, filter

sharing and patch-based input, the improved DenseCNN2 model achieved an accuracy of about 92.5%, sensitivity around 88.2%, specificity near 95%, and AUC close to 97.9% on structural MRI. This indicates that concentrating on key regions like the hippocampus can still give strong diagnostic power with fewer parameters[13].

Multimodal fusion has also been explored using both MRI and FDG-PET. One method created a new GM-PET modality by fusing gray-matter information from MRI and PET, then trained 3D CNNs on this composite image. For AD vs. normal classification, the proposed image fusion framework reached about 94.1% accuracy, with sensitivity around 92–93% and specificity around 94–95%, consistently outperforming single-modality MRI or PET and simple feature-level fusion[14], [15].

A recent multi-modal approach combined 3D MRI and amyloid PET in a dual-branch CNN where features are fused before classification. On the OASIS-3 dataset, a 3D MRI-only model achieved roughly 91.7% Accuracy, while the MRI+PET fusion model improved performance to about 95% Accuracy, 93.3% Sensitivity, 96.7% Specificity, and 93% AUC. The authors also used Grad-CAM to show that the model focuses on temporal and parietal regions linked to Alzheimer's pathology[16].

Another structural MRI study developed a multistage, 26-layer CNN to both detect dementia and identify its severity. The network combined residual and dense connections and was trained on OASIS and ADNI data. It reported around 98.2% accuracy for binary dementia detection and about 99.7% Accuracy for subclassifying dementia stages, and even achieved 100% accuracy on a smaller independent dementia cohort, indicating a very strong fit to the available datasets[17].

There are also works that focus on multi-class CNNs using 3D MRI but implemented as 2D networks for efficiency. One such model used a 2D deep CNN fed with slices from 3D volumes and reported about 99.9% accuracy for three-class Alzheimer's stage classification, showing how careful pre-processing and architecture design can push performance close to perfect on benchmark data. However, such very high numbers may not directly transfer to more diverse, real-world clinical populations[18].

Another paper used a DenseNet-201 transfer learning model on a five-class Alzheimer's MRI dataset (non-demented, very mild, mild, moderate, severe). After pre-processing

and heavy data augmentation, the model reached an average Accuracy of about 98.24%, with per-class accuracies often above 95% and an AUC around 97.36%. These results show that deep transfer learning on DenseNet-201 can give very strong multi-class performance for dementia staging[19], [20].

A recent hybrid DL study combined a ResNet50 CNN branch with a Vision Transformer branch using an adaptive feature fusion layer for multi-stage Alzheimer's classification from T1-weighted MRI. On the AD5C dataset (five classes), the hybrid model achieved around 99.42% Accuracy, with Precision about 99.55%, Recall 99.46%, and F1-Score 99.50%, outperforming earlier single-model baselines. The work showed that dynamically fusing local CNN features and global transformer features can significantly boost stage-wise classification accuracy[21].

METHODOLOGY

1.9 Introduction

In this study, we aim to detect AD using DL models. The workflow begins with the collection of the Alzheimers disease dataset from Mendeley, which contains images of infected and non-infected individuals from different villages in Bangladesh. The images are preprocessed using normalization, resizing, batching, caching, and prefetching to ensure uniformity and efficient training. We then apply both standard pre-trained convolutional neural network architectures, including VGG16, DenseNet121, and ResNet101, as well as a Customized GoogleNet model designed specifically for this classification task. The models are trained and evaluated using performance metrics such as accuracy, precision, recall, F1-score, and ROC–AUC. Furthermore, explainable AI techniques, including Grad-CAM and LIME, are employed to visualize and interpret the regions of the images that influence the model's predictions, ensuring transparency and reliability of the classification. As shown in Figure 1, the proposed architecture design consists of dataset collection, preprocessing, model training using pre-trained and customized CNN models, and evaluation using performance metrics and explainable AI techniques.

1.9.1 Dataset Description

The proposed research was used the AD MRI Dataset that is available in Mendeley Data Repository that is publicly accessible. Kaggle Data has been an effective open-source research platform, hosted by Kaggle, that is aimed at the sharing and conservation of scientific datasets across all fields. It guarantees the transparency and reproducibility of data and their accessibility in the long-term by assigning Digital Object Identifiers (DOIs) to every dataset to enable them to be duly referenced and validated. The data on the AD employed in the present study was chosen in Mendeley because it is authentic, well labelled, big in size and openly available, therefore, it is very practical in DL based medical image analysis.

The dataset contains a large set of magnetic resonance images (MRI) of human brains, which were divided into four different classes, corresponding to certain phases of the development of the AD:

- Non-Demented - This is a healthy subject group that does not show any apparent cognitive deficits or structural abnormalities in brain tissue.
- Very Mild Demented- Refers to the initial, subtle signs of cognitive impairment with slight shrinkage of the hippocampal and slight cortical thinning.
- Mild Demented - Is characterised by observable structural abnormalities, with increased atrophy of the hippocampal as well as enlarged ventricles.
- Moderate Demented - Appearing in the later stages with a severe cortical atrophy, high tissue loss and a severally enlarged ventricles.

The data is separated into two major folders:

- Original Images - these are the original MRI scans that are to be used primarily in validation and testing to see generalisation through testing on unseen data.
- Augmented Images - incorporating artificially enriched samples generated with the use of augmentation methods, i.e., rotation, flipping, scaling, and adjustment of brightness to enhance robustness and deal with the problem of class imbalance.

Overall, the dataset has about 40,000 MRI images in each of the four classes and, therefore, it has balanced representation, which is conducive to effective training and testing. It is also enabled by the fact that it incorporates both the training and testing subsets that allow the full assessment of model accuracy and generalisability.

Error! Reference source not found. shows examples of MRI of each of the four classes and visually depicts how the brain structure degenerates with the advancement of the disease. The images emphasize the change in distinct cortical structures of healthy individuals to massive cortical atrophy and ventricular fibrillation in moderate cases of dementia, which stresses the validity of the dataset in terms of the ability to determine disease progression patterns.

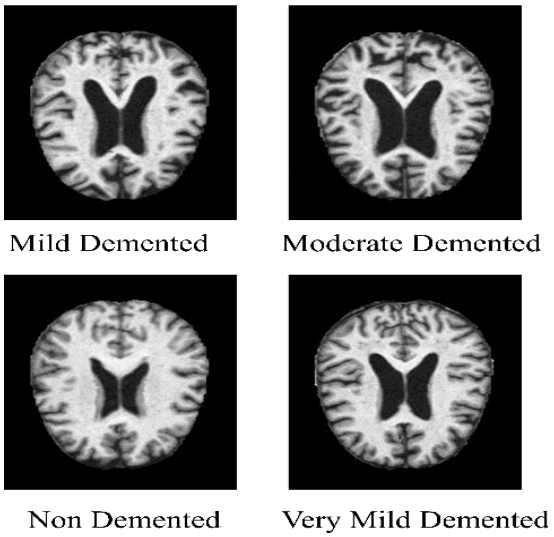


Figure 1 Sample MRI Images Representing Different Stages of Alzheimer's Disease

1.9.2 Proposed Framework

This flowchart, presented in Error! Reference source not found., illustrates the complete methodological pipeline used in this study, beginning with Alzheimer MRI data acquisition followed by preprocessing steps such as normalization, resizing, batching, caching, prefetching, and augmentation. The dataset is divided into three subsets like training, validation, and testing. Multiple transfer learning models (DenseNet121, VGG16, ResNet101, and a customized GoogLeNet) are evaluated using Accuracy, Precision, Recall, Test-Loss, and AUC. The best-performing model is then interpreted using explainable AI techniques, including LIME and Grad-CAM.

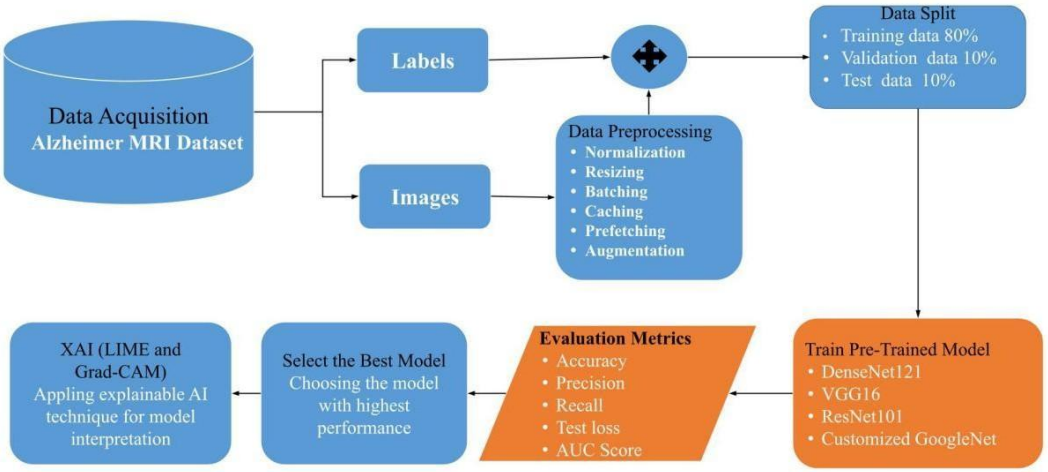


Figure 2 Workflow of the Proposed AD Classification Framework

1.9.3 Data Preprocessing

The dataset were initially of varying sizes, so all images were resized to a uniform dimension of 224×224 pixels. Several preprocessing techniques were applied to improve model training and convergence.

1.9.3.1 Normalization (Scaling Pixel Values)

The pixel values of all input images were normalized to the range [0,1] by dividing each pixel intensity by 255:

where x represents the original pixel intensity, and x′ is the normalized value. This ensures that the neural network receives inputs on a consistent scale, improving convergence speed, training stability, and reducing numerical instability during gradient updates.

1.9.3.2 Resizing

All images were resized to 224 × 224 pixels using bilinear interpolation to maintain image quality. Resizing ensures uniformity across the dataset, enables efficient batch processing, and maintains compatibility with pretrained network weights[22].

1.9.3.3 Batching

The dataset was divided into mini-batches of size 32. Batching allows the model to process multiple images in parallel, maximizing GPU utilization while keeping memory requirements manageable. Additionally, batching stabilizes gradient estimates during optimization, leading to smoother and faster convergence[23].

1.9.3.4 Caching

To reduce preprocessing overhead, the preprocessed dataset was cached in memory. This stores already-processed images, preventing repeated transformations in subsequent epochs and improving training efficiency[24].

1.9.3.5 Prefetching

Prefetching was employed to overlap data loading and model execution. While the GPU processes one batch, the next batch is prepared in the background, reducing idle GPU time. TensorFlow's AUTOTUNE feature was used to automatically determine the optimal prefetch buffer size[25].

Table 1 Model Training Parameters

Parameter /Technique	Value /Range
Image input shape	224×224×3
Number of classes	4
Batch size	32
Number of epochs	30
Train-Test: Validation split	80:10:10
Optimizer	Adam
Learning rate	0.001
Dropout rate	0.3
Early stopping criterion	Validation accuracy

1.9.4 Data Split

The dataset was partitioned into three subsets to ensure robust model evaluation and to mitigate overfitting. Specifically, 80% of the data was allocated for training, 10% for validation, and the remaining 10% for testing. Data augmentation techniques, including random rotation, horizontal and vertical flipping, and scaling, were applied exclusively to the training images to improve the model's ability to generalize and adapt to variations in input data. In contrast, the validation and test sets were left unaltered to ensure an unbiased assessment of the model's performance. This configuration allows the model to learn discriminative and generalizable feature representations while ensuring reliable evaluation on unseen data[26].

1.9.5 Transfer Learning Models

1.9.5.1 Customized GoogleNet

In this study, a customized version of the GoogLeNet (Inception v1) architecture was implemented using TensorFlow to classify MRI brain images into four categories of AD: Non-Demented, Very Mild-Demented, Mild-Demented, and Moderate-Demented. The network begins with a 7×7 convolutional layer (64 filters, stride 2) followed by 3×3 max pooling and batch normalization to normalize feature activations. Subsequent convolutional layers with 64 and 192 filters further refine low-level representations before entering the core Inception modules. The Inception structure consists of parallel 1×1, 3×3, and 5×5 convolutional paths alongside a max-pooling branch, all concatenated to capture both local and global spatial features efficiently. Dimensionality reduction is achieved using 1×1 convolution preceding higher-dimensional operations, minimizing computational complexity[27].

Multiple stacked Inception modules progressively increase feature depth while reducing spatial resolution through intermediate max-pooling layers. The final feature maps are processed by a Global Average Pooling layer to replace dense connections, followed by a Dropout layer (rate = 0.4) to prevent overfitting. A Dense output layer with softmax activation produces the final four-class prediction. The model was trained using the Adam

optimizer with a learning rate of 0.0001, employing sparse categorical cross-entropy loss over 30 epochs. This configuration ensures a balanced trade-off between convergence speed and generalization[28].

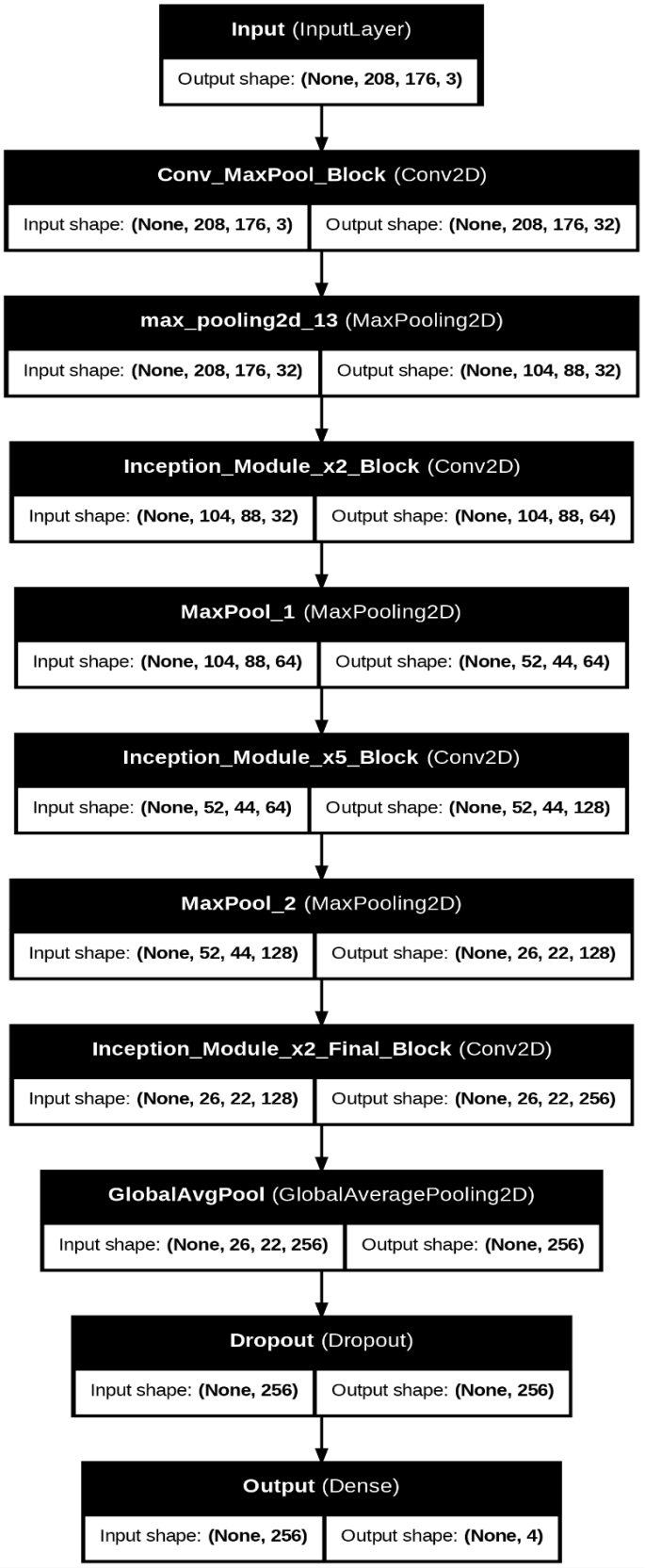


Figure 3 Architectural Overview of the Customized GoogLeNet Model

1.9.5.2 DenseNet121

A deep convolutional neural network architecture that was introduced by Hasan et al. (2021) and is known as DenseNet-121 (Dense Convolutional Network) is an effective way of solving the vanishing gradient issue as well as encouraging features reuse in all layers. In contrast to the traditional architectures, where every layer is fed by the layer before it, DenseNet provides dense connectivity between layers with each layer receiving the feature maps of all the previous layers and sending its own feature maps to all the following layers. Such a design promotes the flow of information as much as possible across the network and promotes feature propagation, reuse, and efficiency[29].

Formally, the output of the l^{th} layer,x_l, can be expressed as:

$$x_l = H_l([x_0, x_1, x_2, \ldots x_{l-1}]) \quad 3.1$$

Where $[x_0,\ x_1,\ x_2,\ \ldots\ x_{l-1}]$ denotes the concatenation of the feature maps produced by later 0 through $l - 1$ and $H_l(.\ )$ Represents a composite function of operations such as Batch normalization, Rectified linear unit (ReLU), and Convolution.

Each dense block is followed by a transition layer, which performs feature compression and spatial down-sampling to control the network's complexity. The transition layer output x_l'can be expressed as:

$$x_l' = H_l(Conv_{1\times1}(AvgPool_{1\times1}(x_l)) \quad 3.2$$

Where $H_l(.\ )$ denotes a transition operation that consists of a convolutional layer followed by an average pooling step.

The final feature representation undergoes global average pooling, followed by a fully connected layer with a softmax activation to produce class probabilities:

$$\hat{y} = softmax(W_f.\, x_{global} + b_f) \quad 3.1$$

The DenseNet-121 model is a model that has 121 layers with convolutional, pooling and fully connected layers arranged in four dense blocks with transition layers in between that perform down-sampling. The dense blocks consist of several layers of convolutional networks in which feature maps are not summed up but concatenated, which results in a better use of the parameters and gradient flow. This architecture starts by a convolution and pooling layer, which is then followed by dense and transition blocks and finally finishes with a global average pooling layer and fully connected output layer with the loss of a softmax function that classifies the output[30], [31].

1.9.5.3 ResNet101

ResNet50 is a DCNN developed by He et al. (2015) as part of the residual networks (ResNets), which won the ILSVRC 2015 competition. It has 101 layers comprising convolutional blocks with residual connections. The key innovation of ResNet101 is residual learning, in which shortcut (skip) connections bypass one or more layers. A residual block can be expressed as:

$$y = F(x, \{W_i\}) + x \tag{3.1}$$

where x and y are the input and output vectors of the block, and F represents the learned residual mapping. This architecture alleviates the vanishing gradient problem and enables deeper networks to be trained effectively. ResNet101 consists of 3×3 convolutional layers, batch normalization, ReLU activations, pooling layers, and bottleneck residual blocks to balance computational efficiency and representation power. It is pretrained on ImageNet and demonstrates state-of-the-art performance on various image classification tasks[32], [33].

1.9.5.4 VGG16

VGG16 is CNN builded by the VGG and was one of the top performers in the 2014 ILSVRC competition . The final convolutional feature maps $F \in R^{h \times w \times d}$ are flattened into a one-dimensional vector $x \in R^{h.w.d}$:

$$x_n = F_{i,j,k} \tag{3.1}$$

where, $n = k.h.w + i.w + j$

This vector is passed through a sequence of fully connected Dense layers each followed by Batch Normalization, LeakyReLU activation, and dropout regularization to prevent overfitting. The final output layer uses a softmax activation to predict class probabilities:

$$\hat{y}_i = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}} \quad 3.1$$

where C is the number of classes, z_i is the input to the softmax for class i, and $\hat{y}_i$ is the predicted probability. The model optimization employed the Adam optimizer with categorical cross-entropy loss:

$$\mathcal{L} = \sum_{i=1}^{C} y_i \log(\hat{y}_i), \quad 3.1$$

where y_i is the true label. Model Checkpoint and Early Stopping callbacks were applied to preserve the optimal model and reduce overfitting.[34], [35].

1.9.6 Performance Metrics

1.9.6.1 Accuracy

The accuracy of any predictive method is the foundation for assessing its performance in machine learning. It basically computes the proportion of correctly predicted overall data points. The accuracy score interpreted as the maximum possible accuracy is 1.0, while the minimum possible accuracy is 0.0. It is simple to calculate by dividing the number of correctly predicted by the total of projections[36]. Also, it can be expressed as,

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad 3.1$$

1.9.6.2 Precision

The Precision focuses on the quality of the model's positive predictions. It tells us how many of the "positive" predictions were actually correct. It is important in situations

where false positives need to be minimized, such as detecting spam emails or fraud[37]. The formula of precision,

$$Precision = \frac{TP}{TP + FP} \quad 3.2$$

1.9.6.3 Recall

Recall is a performance metric that measures the ability of a classifier to identify all relevant instances. It is defined as the ratio of true positives to the sum of true positives and false negatives, indicating the model's completeness.

$$Recall = \frac{TP}{TP + FN} \quad 3.3$$

1.9.6.4 F1-Score

The F-Score is the harmonic mean of precision and recall in binary classification, serving as a single summary measure of a classifier's performance, particularly useful when the prevalence of the positive class is low[38].

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad 3.4$$

1.9.6.5 AUC (Area Under the ROC Curve)

Area Under the Curve (AUC) evaluates a binary classifier's ability to distinguish between positive and negative classes. It is computed as the area under the ROC curve, which plots the TPR versus the FPR. A higher AUC indicates better discriminative capability[39].

$$AUC = \int_0^1 TPR(FRP^{-1}(x))dx \quad 3.5$$

Where

$TPR = \frac{TP}{TP+FN}, FPR = \frac{FP}{FP+TN}$

The above equations are constructed for all the classifiers that consist of true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

1.9.7 Hinge Loss

Hinge loss for multiclass classification is an extension of the hinge loss used in binary classification. It is commonly used in support vector machines (SVMs) for multiclass problems and works by ensuring a margin between the score of the true class and the scores of the other classes.

The multiclass hinge loss for an example with true class y_i and score function output $s(x_i, y_i)$ for each class y is defined as:

$$L(y, f(x)) = \max(0,1 - y * f(x)) \quad 3.6$$

where

- y the actual class (-1 or 1).
- $f(x)$ the output of the classifier for the datapoint.

The loss sums over each incorrect class, penalizing those that score within the margin of the correct class score.

1.9.8 Grad-Cam

Explainable Artificial Intelligence (XAI) aims to provide human-interpretable explanations of DL models, especially CNN . Gradient-weighted Class Activation

Mapping (Grad-CAM), proposed by Selvaraju et al. (2017), is one of the most widely used XAI techniques. Grad-CAM computes the importance weights for each feature map in the final convolutional layer by aggregating the gradients of the target class score with respect to the feature maps. Mathematically, for a class c and feature map Ak, the weight αc k is defined as:

$$\alpha_k{}^c = \frac{1}{z}\sum_i \sum_j \frac{\partial y^c}{\partial A^k{}_{ij}} \qquad 3.7$$

where y^c is the score for class c, $A_{ij}{}^k$ is the activation at location (i, j) in feature map k, and Z is the total number of spatial locations. The final Grad-CAM heatmap $L^c{}_{Grad-CAM}$ is obtained by a weighted combination of feature maps:

$$L^c{}_{Grad-CAM} = ReLU\left(\sum_k \alpha^c{}_k A^k\right) \qquad 3.8$$

The ReLU ensures that only positive contributions are considered, emphasizing regions that positively influence the prediction.

1.9.9 Lime

LIME enhances interpretability by approximating complex models with simpler, interpretable surrogate models locally around the prediction of interest. For images, LIME divides an image into interpretable components called super pixels and perturbs them randomly. The resulting changes in prediction probabilities are used to train a linear or simple model that approximates the original model locally.

Mathematically, let f be the original model, x the input image, and x' a perturbed version. LIME minimizes the following objective:

$$L(f, g, \pi_x) = \sum_{x'} \pi_x\,(x')(f(x') - g(x'))^2 + \Omega(g) \qquad 3.9$$

where g is the interpretable surrogate model, π_x is a proximity measure between x and x', and $\Omega(g)$ is a complexity penalty to keep g simple. The weights of g then indicate which superpixels contribute most strongly to the prediction, providing insight into the model's decision-making.

RESULTS AND DISCUSSION

1.10 Result

This chapter reports the experimental findings of the experimental procedures and testing of four transfer learning structures including GoogLeNet, DenseNet121, ResNet101 and VGG16, to classify the AD within the MRI brain scan. The key goal of the analysis is to evaluate the performance of each of the models in several categories of diagnosis: Non-Demented, Very Mild-Demented, Mild-Demented, and Moderate-Demented. The assessment is based on the well-known performance metrics as Accuracy, Precision, Recall, and F1-Score and also on the sophisticated ones, such as AUC and mean categorical hinge loss. All of these metrics will be used to form a complete picture of every model in terms of its classification ability, sensitivity and overall reliability. Besides the quantitative assessment, the visual performance analysis is presented in radar charts, confusion matrixes, and ROC curves to make the strong and weak sides of each architecture apparent in various dementia stages. In addition, training and validation curve is evaluated to determine the learning stability and convergence behaviour in optimizing the model. In order to improve interpretability, XAI algorithms, such as Grad-CAM and LIME are used to show the regions of attention in MRI scans that are the most relevant in model predictions. Such explainability studies make sure that the decision-making process corresponds to the clinically significant brain structures, which supports the belief in AI-based diagnostic systems.

1.10.1 Model Performance

Table 2 Model-wise and Class-wise Performance Comparison

Model Name	Class	Accuracy	Precision	Recall	F1 score	Support

DenseNet121	0(Non-Demented)	0.8355	0.9876	0.6374	0.7748	1743
	1(Very Mild Demented)		0.8989	0.9992	0.9464	1282
	2(Mild Demented)		0.9189	0.7714	0.8387	1938
	3(Moderate Demented)		0.6838	0.9771	0.8046	1833
GoogLeNet	0(Non-Demented)	0.9800	0.98	0.99	0.99	1743
	1(Very Mild Demented)		1.00	1.00	1.00	1282
	2(Mild Demented)		0.98	0.97	0.97	1938
	3(Moderate Demented)		0.97	0.97	0.97	1833
ResNet	0(Non-Demented)	0.7310	0.6906	0.7390	0.7140	1743
	1(Very Mild Demented)		0.9438	0.9821	0.9625	1282
	2(Mild Demented)		0.7506	0.6847	0.7161	1938
	3(Moderate Demented)		0.5861	0.5848	0.5855	1833
VGG16	0(Non-Demented)	0.6826	0.6821	0.6254	0.6525	1743
	1(Very Mild Demented)		0.8882	0.9789	0.9314	1282
	2(Mild Demented)		0.6153	0.7534	0.6773	1938
	3(Moderate Demented)		0.5907	0.4550	0.5140	1833

Table 2 presents the detailed performance evaluation of four transfer learning architectures, DenseNet121, GoogLeNet, ResNet, and VGG16, for AD classhification using MRIi images. Each model's effectiveness is assessed through accuracy, precision,

Rtecall, and F1-Score across four classes: Non-Demented, Very Mild Demented, Mild Demented, and Moderate Demented.

Among all models, GoogLeNet achieved the highest overall accuracy of 0.98, demonstrating exceptional and consistent results across every class. For the Non-Demented category, GoogLeNet gain a Precision of 0.98, a Recall of 0.99, and an F1-Score of 0.99, shows a nearly perfect balance among prediction accuracy and sensitivity. In the Very Mild-Demented class, the model gain perfect scores of 1.00 for precision, Recall, and F1-Score, showing its strong ability to detect early-stage dementia. Similarly, for the Mild-Demented class, GoogLeNet attained a precision of 0.98, a recall of 0.97, and an F1-score of 0.97. The Moderate Demented class also showed consistently high results, with a Precision of 0.97, a Recall of 0.97, and an F1-Score of 0.97. The DenseNet121 model ranked second, achieving an overall accuracy of 0.8355. It displayed robust performance in the Very Mild Demented and Mild Demented categories, recording precision, recall, and F1-score values of 0.8989, 0.9992, and 0.9464 for the former, and 0.9189, 0.7714, and 0.8387 for the latter. For the Non-Demented class, the precision was 0.9876, the recall was 0.6374, and the F1-score was 0.7748, indicating that while DenseNet121 was highly accurate in identifying Non-Demented samples, it occasionally misclassified a few instances. Meanwhile, the Moderate Demented category achieved a precision of 0.6838, a recall of 0.9771, and an F1-score of 0.8046, demonstrating good sensitivity but moderate precision. The ResNet model attained an overall accuracy of 0.731, reflecting moderate but consistent classification ability. For the Non-Demented class, ResNet achieved a Precision of 0.6906, a Recall of 0.7390, and an F1-score of 0.7140. In contrast, its performance improved significantly for the Very Mild Demented class, where it gain a Precision of 0.9438, a Recall of 0.9821, and an F1-Score of 0.9625. For the Mild Demented class, precision was 0.7506, recall was 0.6847, and F1-score was 0.7161, while for the Moderate Demented class, Precision, Recall, and F1-Score were 0.5861, 0.5848, and 0.5855, respectively. The VGG16 model achieved the lowest overall accuracy of 0.6826 among the four architectures, suggesting relatively weaker generalization performance. In the Very Mild Demented class, however, VGG16 performed relatively well, achieving a Precision of 0.8882, a Recall of 0.9789, and an F1-Score of 0.9314. For the Non-Demented category, the model recorded a Precision of

0.6821, a Recall of 0.6254, and an F1-Score of 0.6525. In the Mild Demented , precision was 0.6153, recall was 0.7534, and F1-score was 0.6773, while in the Moderate Demented class, the values dropped further to a precision of 0.5907, a recall of 0.4550, and an F1-score of 0.5140.

1.10.2 Confusion Matrics

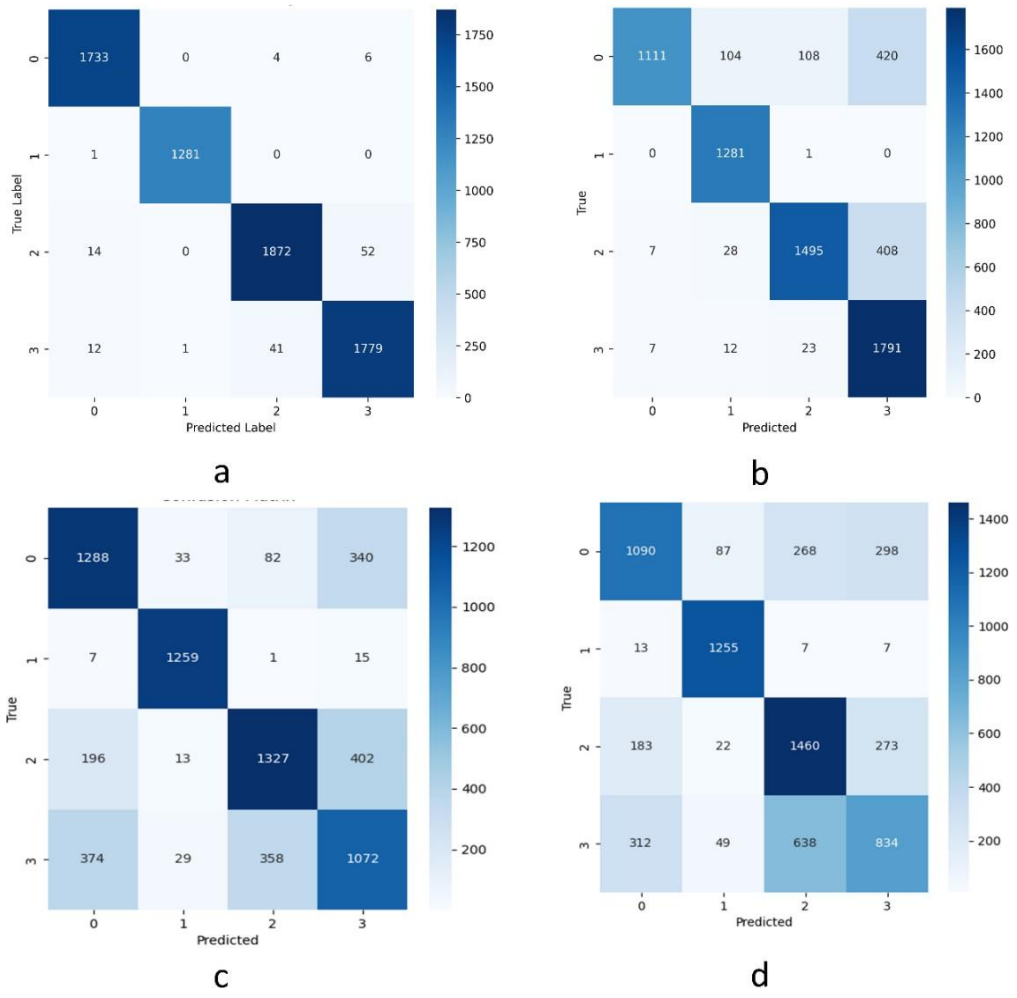


Figure 4 Confusion Matrix of Transfer Learning Models

The Figure 4 displays the confusion matrix of four TL architectures, GoogLeNet, DenseNet121, ResNet101, and VGG16, in recognizing MRI images as one of four categories of AD Non-Demented (Class 0), Very Mild-Demented (Class 1), Mild-Demented (Class 2), and Moderate-Demented (Class 3). Both matrices give a graphical display of the instances that were correctly classified as misclassified in the off-diagonal cells which allows a full comparison of both models classification behavior at all the levels of dementia.

Figure 4(a), where we have GoogLeNet model, we see that the model has very good classification accuracy with the lowest number of misclassifications. This model was able to predict 1,733 Non-Demented, 1,281 Very Mild Demented, 1,872 Mild Demented, and 1,779 Moderate Demented samples correctly. There were 4 Non-Demented samples that were wrongly classified as Mild Demented, 6 Non-Demented that were wrongly classified as Moderate Demented, 14 Moderate Demented that were misclassified as Non-Demented and 52 Non-Demented that were incorrectly classified as Moderate Demented. In the same way, 12 Moderate Demented were wrongly grouped as Non-Demented, 1 as Very Mild and 41 as Mild Demented. These low error rates show that there are very steady and consistent predictions. The presence of more of the dark blue color at the diagonal proves the strength of GoogLeNet in identifying faint differences in MRI patterns, which also matches its accuracy of 0.98 that was reported before. The structure of the model which was based on inception was successful in capturing the multi-scale features of the spatial environment hence almost perfect classification of all the four stages of AD.

Figure 4(b) depicts DenseNet121 model, which also had a high predictive performance but it had a few more off-diagonal confusions compared to GoogLeNet. The model was right in its classification of 1,111 Non-Demented, 1,281Very Mild Demented, 1,495Mild Demented and 1,791Moderate Demented. It however, poorly categorized 104 Non-Demented as Very Mild, 108 as Mild and 420 as Moderate Demented. Also, 7 Mild Demented were classified as Non-Demented, 28 as Very Mild and 408 as Moderate. As these misclassifications were rather narrow, they demonstrate that DenseNet121 was a little bit troubled with the distinction of similar features between the Mild and Moderate

categories. However, the combined accuracy of 0.8355 with a macro-average F1-score of 0.8411 indicate the high reliability of it and its propagation of features effectively through dense connectivity between layers.

The confusion matrix of ResNet101 model is presented in Figure 4(c), and the model obtained moderately better results than the two initial architectures. The correct sample classification of ResNet101 was as follows: Non-Demented (1,288), Very Mild Demented (1,259), Mild Demented (1,327), and Moderate Demented (1,072). But there were also more misclassifications especially with the Moderate Demented cases where 374 cases were misclassified as Non-Demented, 29 as Very Mild and 358 as Mild Demented cases. On the same note, 196 Mild Demented cases were misclassified as Non-Demented, 13 as Very Mild and 402 as Moderate Demented. These findings show that although ResNet101 was very effective in detecting early and mid-stage dementia, it was not effective in severe cases because of an overlap of features in the late phases of the disease. The general accuracy of 0.731 with the preciseness and the recall of 0.7427 and 0.7474 respectively indicate the average yet reliable performance of the tool at all levels.

Lastly, Figure 4(d) illustrates the confusion matrix of the VGG16 model which had the least performance of the four architectures. This model yielded 1,090 Non-Demented, 1,255 Very Mild Demented, 1,460 Mild Demented and 834 cases of Moderate Demented. There were however a lot of misclassifications across categories. One of the examples is the 87 Non-Demented samples that wrongly marked as the Very Mild, 268 marked as Mild, and 298 as Moderate Demented. Also, 183 Mild Demented samples were mistaken to be Non-Demented, 22 were Very Mild and 273 were Mild. Likewise, 312 Instances of Moderate Demented were also incorrectly classified as Non-Demented, 49 asVery Mild and 638 asMild Demented. The general low accuracy of 0.6826 and F1-score of 0.6748 can be attributed to the general low level of misclassification in VGG16, indicating that the relatively low level of misclassification might be due to the relatively shallow architecture of the model that failed to capture the intricate structural variations and textural features inherent in the Alzheimer MRI scans.

1.10.3 ROC Curves for Model Comparision

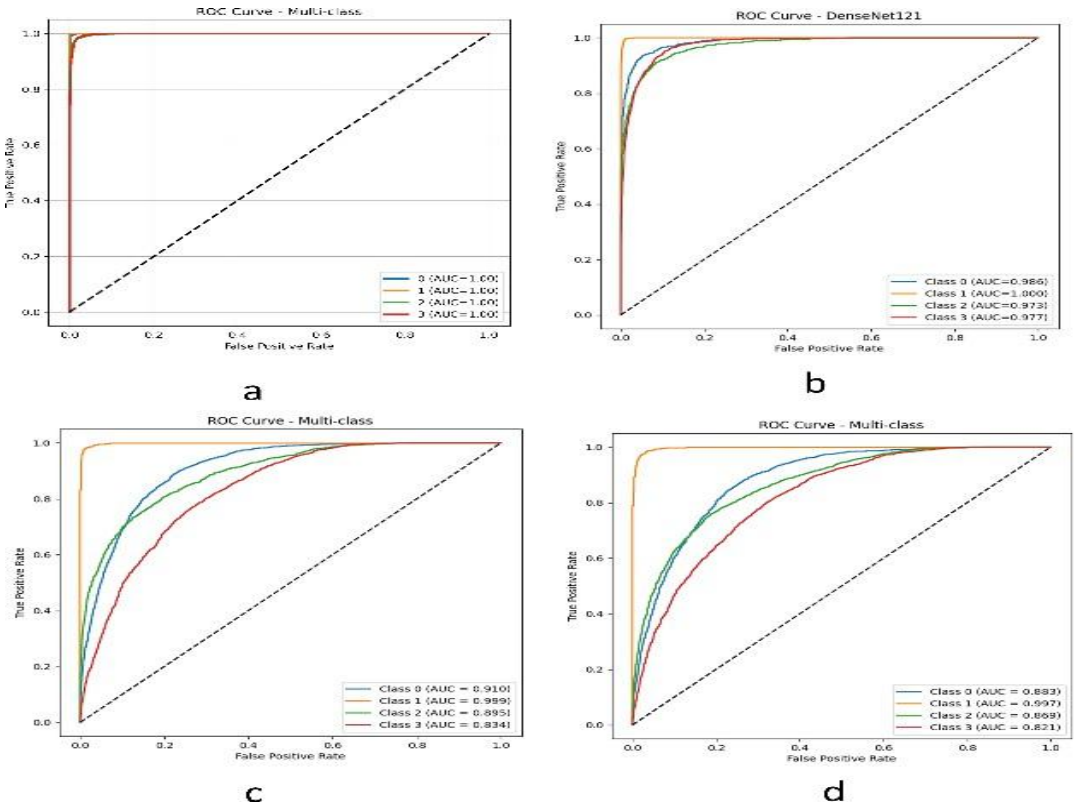


Figure 5 ROC-AUC Curves for Transfer learning Models

Figure 5 shows the Receiver Operating Characteristic (ROC) curves and the corresponding Area Under the Curve (AUC) values of the four transfer learning architectures GoogLeNet, DenseNet121, ResNet101, and VGG16, when applied to the MRI brain images to classify the images into four diagnostic categories that include Non-Demented (Class 0), Very Mild Demented (Class 1), Mild Demented (Class 2) and Moderate Demented (Class 3). The ROC curve is an important performance measure that illustrates the trade-off between the true positive rate (sensitivity) and the false positive

rate (1 - specificity) of each of the classes. The model whose AUC is more similar to the value of 1.0 has better discriminative power and accuracy of classification.

Figure 5(a) shows that the GoogLeNet model is almost flawless in its discrimination of all classes with AUC values of 1.00 of Class 0 (Non-Demented), Class 1 (Very Mild Demented), Class 2 (Mild Demented) and Class 3 (Moderate Demented). The four ROC curves are very close to one another in the upper-left part of the plot, which means that GoogLeNet can practically perfectly identify all the dementia type with minimum false identifications. Such an excellent result is in line with its earlier reported overall accuracy of 0.98 and F1-score of 0.98, which proves the high reliability and capacity of the model to generalize successfully on different MRI patterns and disease stages.

Figure 5(b) shows the ROC curves of DenseNet121 model that also exhibit outstanding classification capability. Class 0 has AUC values of 0.986, Class 1 has AUC values of 1.000, Class 2 has AUC values of 0.973 and Class 3 has AUC values of 0.977. These values reveal that DenseNet121 had very high sensitivity and specificity values across all four classes and its ROC curves are almost near the top-left corner of the graph. The slightly lower AUC of Class 2 (Mild Demented) indicates some slightness over the differenceiation of middle case cases of dementia, but it is rather impressive overall. This finding supports the general accuracy of DenseNet121 of 0.8355, which makes it the second-most efficient model in this comparative study.

The ROC curves of the ResNet101 model are shown in Figure 5(c) and this indicates that the model has a moderately strong classification ability. The model has obtained AUC of 0.910, 0.999, 0.834 respectively with Classs 0, 1,2, and 3. The ROC curve of Class 1 (Very Mild Demented) is almost ideal meaning that ResNet101 is very effective in detecting Alzheimer at an early stage. The decrease in AUC of Class 3 is however, indicative of lower sensitivity of the model at higher stages of dementia. The results are similar to those of its confusion matrix where there was more overlap between the Mild and Moderate categories. Having an overall accuracy of 0.731, ResNet101 offers solid performance but fails to perform as powerfully as GoogLeNet and DenseNet121.

Finally, in Figure 5(d), the models of VGG16 are used to illustrate the ROC curves and the resulting AUC values are lower, demonstrating the poorer predictive abilities of this model. The model scored AUC of 0.883 in Class 0, 0.967 in Class 1, 0.821 in Class 2 and 3 respectively. Although the Very Mild Demented group had relatively high AUC, the decrease in AUC of the other classes, in particular, the Moderate Demented one, suggests the lack of sensitivity and increased false positive rate. ROC curves of VGG16 are made even more distant to the upper-left corner, which indicates that it is unable to distinguish similar MRI patterns. These results correspond with its reduced overall accuracy of 0.6826 and show that the model has not enough depth and advanced feature extraction ability in comparison to more complex networks.

1.10.4 Learning Curve

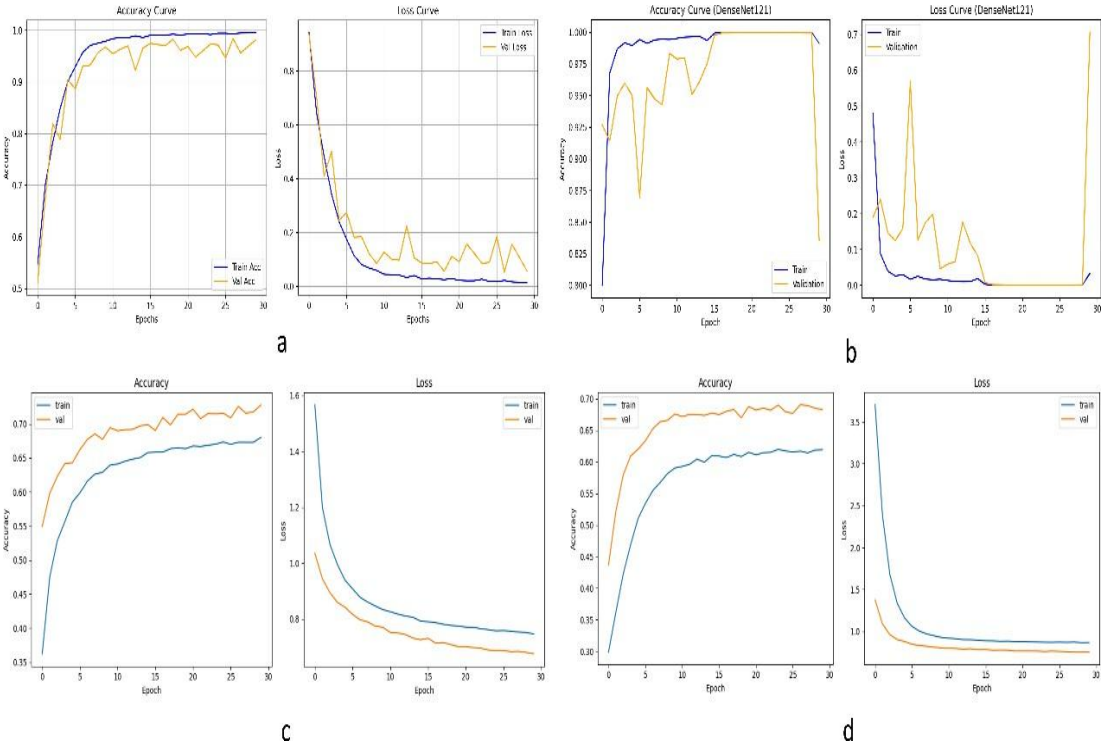


Figure 6 Training and Validation Accuracy–Loss Curves of Transfer Learning Models

Figure 6 is a plot that shows both training and validation accuracy and loss curves of the four transfer learning models, GoogLeNet, DenseNet121, ResNet101, and VGG16, when trained on the MRI brain image data on the classification of AD. All the models have been trained over 30 epochs and the plots illustrate the learning efficiency, convergence and generalization of the models. These graphs offer a profound understanding of the ability of each architecture to learn the discriminative patterns, curb overfitting, and stabilize throughout training by comparing the changes of the training accuracy and loss with the validation accuracy and loss.

In Figure 6(a), GoogLeNet model has the most desirable training dynamics with high level of stability. In the initial few epochs, training and validation accuracy increase rapidly at the range of 0.55 to 0.90 and ultimately levels to about 0.99 in epoch 10 and 25, respectively. At the same time, the training loss drops to approximately 0.0 instead of almost 1.0 and the validation loss drops to approximately 0.05 towards the end of the training. It is worth noting that the two accuracy curves, training and validation, nearly overlap, which means that there is little overfitting and it is well-generalized to unseen data. The fact that the two loss curves decrease in parallel is yet another confirmation that the model is stable and that GoogLeNet has been able to optimize its parameters to achieve the lowest possible misclassification rates, as well as ensure that the training and validation curves look very similar. The outstanding learning trend is in agreement with its close-to-perfect quantitative learning performance (accuracy = 0.98, AUC = 1.00 across all classes), indicating that GoogLeNet is capable of differentiating all stages of an AD with outstanding accuracy. This high learning efficiency and balanced convergence is due in part to its inception-based architecture, which is a learning architecture that extracts features over a range of receptive scales.

The accuracyloss curves of the DenseNet121 model are presented in Figure 6(b) that also indicate the strong and fast convergence with certain fluctuation in validation loss. The training accuracy starts with the value of about 0.80, then immediately rises to 0.95 at epoch 10 and then nearly 1.00 at epoch 25. The validation accuracy is also on the same trend and increases between 0.85 and almost 0.98 but with small variations in the same interval between epochs 10 and 25. The training loss, on the other hand, declines at the

first epoch with approximately 0.70, and the validation loss continuously drops to approximately 0.05 during epoch 10 but then oscillates between 0.05 and 0.7 towards the end of training. These small variations in the validation loss can be due to a small overfitting or instability caused by the high reuse of features and dense connectivity of DenseNet121, which can occasionally enhance noise in difficult MRI images. Nevertheless, the general convergence is high and the general accuracy of the model is 0.8355 and the AUC of the model ranges 0.973 to 1.000, which proves that the model can effectively learn the appropriate features of MRI to stage the disease. The high density of connection between layers provides better propagation of features resulting in very discriminative representation learning on all classes particularly in early stages of dementia.

ResNet101 model (Figure 6(c)) has slower but slow progression in learning in comparison with GoogLeNet and DenseNet121. The training accuracy begins at about 0.40 in the first epoch and slowly grows to 0.67 by epoch 30, the validation accuracy also grows with time beginning at about 0.55, and eventually reaching 0.70 in the 30 th epoch. The loss during the training is reduced to almost 1.6 initially to about 0.75 whereas the validation loss is minimized to about 1.2 to about 0.65 respectively. The narrow difference between the training and the validation accuracy curves indicates fairly good generalization but the slower convergence and premature plateau in the accuracy are indicative of the fact that more time has taken the model to achieve optimal performance. The skip connections and deep structure of ResNet101 enable powerful hierarchical features learning, and the complexity and deep level of its parameters render the model hyper sensitive to modifications in the learning rate and range of data. That is why it has an average accuracy of 0.731 and AUC class-wise of 0.834 to 0.999. Although ResNet101 was able to capture early features of dementia well, its learning rate and poor optimization performance limited its eventual performance as compared to GoogLeNet and DenseNet121.

Lastly, Figure 6(d) shows the accuracy-loss curve of the VGG16 model, which documents the slowest convergent performance and lowest performance of all the four architectures. The training accuracy commences at a rate of about 0.30 and it slowly

climbs to some 0.63 after 30 epochs whereas the validation accuracy starts at around 0.45 and levels to about 0.70 in the end of the training. The training loss reduces to almost 1.0 compared to 3.5 and the validation loss is reduced to 0.8 compared to 2.8 in the same time period. Although there is steady improvement, the values of losses remain relatively high and the difference between the accuracy curves is larger which means that there is less generalization and the chances of underfitting. The model learns gradually but cannot learn the higher-level discriminative features, in part because it has a shallow layer structure and is not very well able to derive the complex spatial hierarchies of MRI data. All these learning properties can be explained by the fact that VGG16 provides less quantitative performance (accuracy = 0.6826, AUC range = 0.821-0.967) compared to the other two, which proves the fact that this model is less flexible when it comes to medical image classification.

1.10.5 Hinge Loss

Table 3 Mean Categorical Hinge Loss of Transfer Learning Models

Model Name	Mean Categorical Hinge Loss
GoogLeNet	0.048368
DenseNet121	0.335367
ResNet101	0.713692
VGG16	0.769410

Table 3 contains the mean categorical values of the hinge loss of the four transfer learning models, including GoogLeNet, DenseNet121, ResNet101, and VGG16 in AD classification using MRI data. The hinge loss measure quantifies how well a model can distinguish the classes in the correct manner and is used to penalize those that are wrong as well as quantify the distance between the score of the predicted and the true class. Reduced values of hinge loss are associated with increased classification confidence, a superior margin separation, and enhanced generalization.

Of all the models, GoogLeNet had the lowest mean hinge loss of 0.048368, which indicated that it had a high classification accuracy and the best decision boundaries between the four dementia classes. This very low loss value indicates that GoogLeNet was effective in reducing errors in misclassification as well as attaining the almost perfect margin separation between actual and predicted labels. The outcome supports the excellent quantitative performance of the model in the past analysis- a total accuracy of 0.98, a macro-average F1-score of 0.98 and an AUC of 1.00 across all classes. GoogLeNet is able to effectively extract multi-scale spatial features using the inception modules which result in highly confident and well-calibrated predictions.

DenseNet121 with an average hinge loss of 0.335367, the separation between classes was quite good, but not ideal, in comparison with GoogLeNet. Although the value of the loss is low it points to the fact that still some overlap between feature representations of the neighbor classes can be observed, Mild and Moderate Demented categories in particular. Nevertheless, DenseNet121 highly interconnected its features and offered a means of reuse, which gave it a good representation learning capability, which matches its overall accuracy of 0.8355 and AUC of between 0.973 and 1.000. This loss of moderation in hinge helps to verify the moderate performance of the model-low accuracy in the initial stages of dementia and reduced inaccuracy in the advanced ones.

ResNet101 had a greater hinge loss of 0.713692, which implies that it is relatively challenging to produce accurate separation between classes. This value indicates that as ResNet101 could learn meaningful hierarchical information in its residual learning structure, it had increased prediction uncertainty and reduced certainty of differentiating between neighboring dementia stages. The outcome is in agreement with its intermediate level of classification performance in terms of accuracy (= 0.731, AUC range = 0.834-0.999). It suggests that more significant architectures such as ResNet101 can be improved with more fine-tuning or regularization when used on more sophisticated medical image datasets.

The VGG16 model had the highest value of hinge loss of 0.769410 which shows that the model has the weakest separation of classes as contrast to the other three models. This higher loss shows a higher rate of misclassification and reduced confidence, which is probably because it has a smaller network depth and lacks the ability to extract features as deeply. This is in line with the lower accuracy of VGG16 (0.6826) and lower AUC scores (0.821-0.967). Simple sequential design of the model is computationally efficient but limits its capability of modeling complex differences in MRI intensity and texture between stages of dementia with an increase in loss and a decrease in classification reliability.

1.10.6 Radar Chart Comparison

1.10.6.1 Radar Chart Comparison of Class-Wise Accuracy

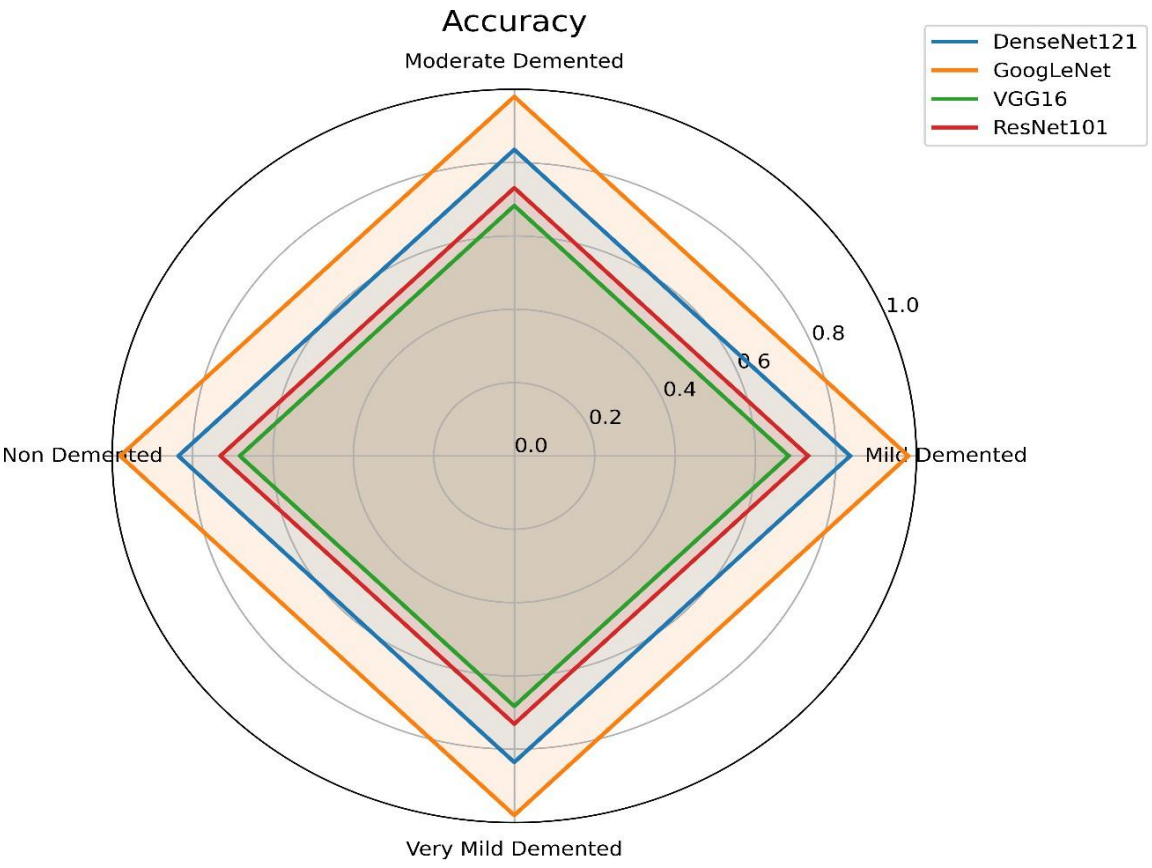


Figure 7 Comparative Radar Chart of Accuracy Across Transfer Learning Models

Figure 7 shows a radar chart to compare the accuracy of the four transfer learning architectures, namely, DenseNet121, GoogLeNet, VGG16, and ResNet101 in classifying the four categories of the AD: Non-Demented, Very Mild Demented, Mild Demented, and Moderate Demented. The different lines in the chart denote each dementia class and the colored lines denote the accuracy of each model in that particular category. The shape of the lines and their dispersion in general gives the consistency and capability to generalize of the model. GoogLeNet (orange line) is the only model that shows the largest and most equal accuracy of all classes and creates an almost perfect outer boundary at the 1.0 mark, which suggests the close to the perfection of classification performance. DenseNet121 (blue line) comes next with high accuracy with the slightest difference among classes. ResNet101 (red line) in comparison displays moderate performance, having a smaller coverage area, which indicates the inability to detect dementia at its advanced stages with the accuracy it needs. VGG16 (green line) has the lowest accuracy and creates the innermost shape that indicates its poorer nature of feature extraction and a weak extrapolation ability. In general, it is evident that the radar chart shows that GoogLeNet is the most effective compared to the other models, followed by DenseNet121, and that ResNet101 and VGG16 have much less reliable classification.

1.10.6.2 Radar Chart Comparison of Class-Wise Recall

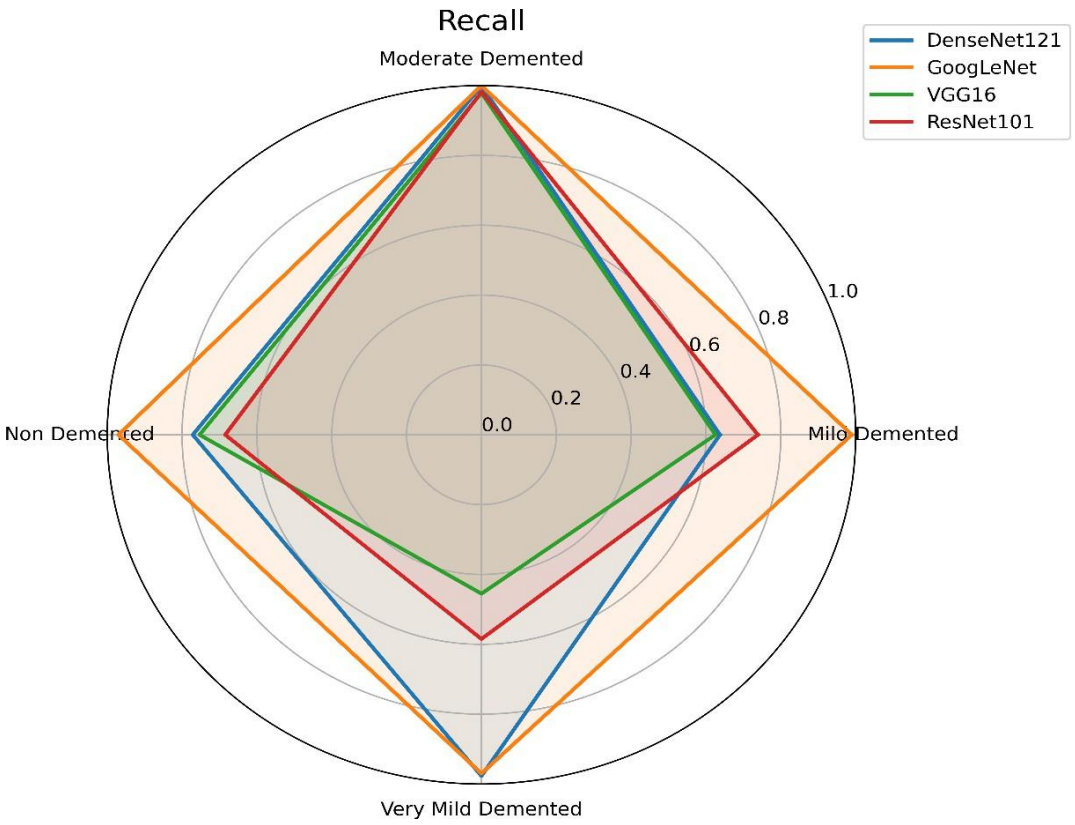


Figure 8 Comparative Radar Chart of Recall Across Transfer Learning Models

Figure 8 depicts a radar chart between the sensitivity (recall) of four transfer learning models, DenseNet121, GoogLeNet, VGG16, and ResNet101 in four diagnostic categories of AD: Non-Demented, Very Mild-Demented, Mild-Demented and Moderate-Demented. Recall measures the capacity of a model to detect all the positive cases, which is especially significant in medicine when false negativity may have severe clinical consequences. As we can see in the chart, GoogLeNet (orange) has the highest and the most consistent recall in each category with a relatively close value to the maximum value of 1.0. This means that GoogLeNet is the most sensitive and it is able to recognize the patients in all the stages of dementia with high false negative. DenseNet121 (blue line) is also a good model with a similar pattern of recall values as GoogLeNet but with a little more interclass variance. ResNet101 (red line) demonstrates moderate recall power, indicating a certain inability to identify true positives very consistently, particularly when

the case is in its early stages. VGG16 (green line) has the lowest recall especially on Very Mild and Non-Demented classes, thus it often fails to identify cases with positive results in the stages. In general, the radar chart shows that GoogLeNet has the strongest and most dependable recall, then DenseNet121, and ResNet101 and VGG16 portray a relatively lower sensitivity in the context of identifying the disease in the framework of Alzheimer.

1.10.6.3 Radar Chart Comparison of Class-Wise Precision

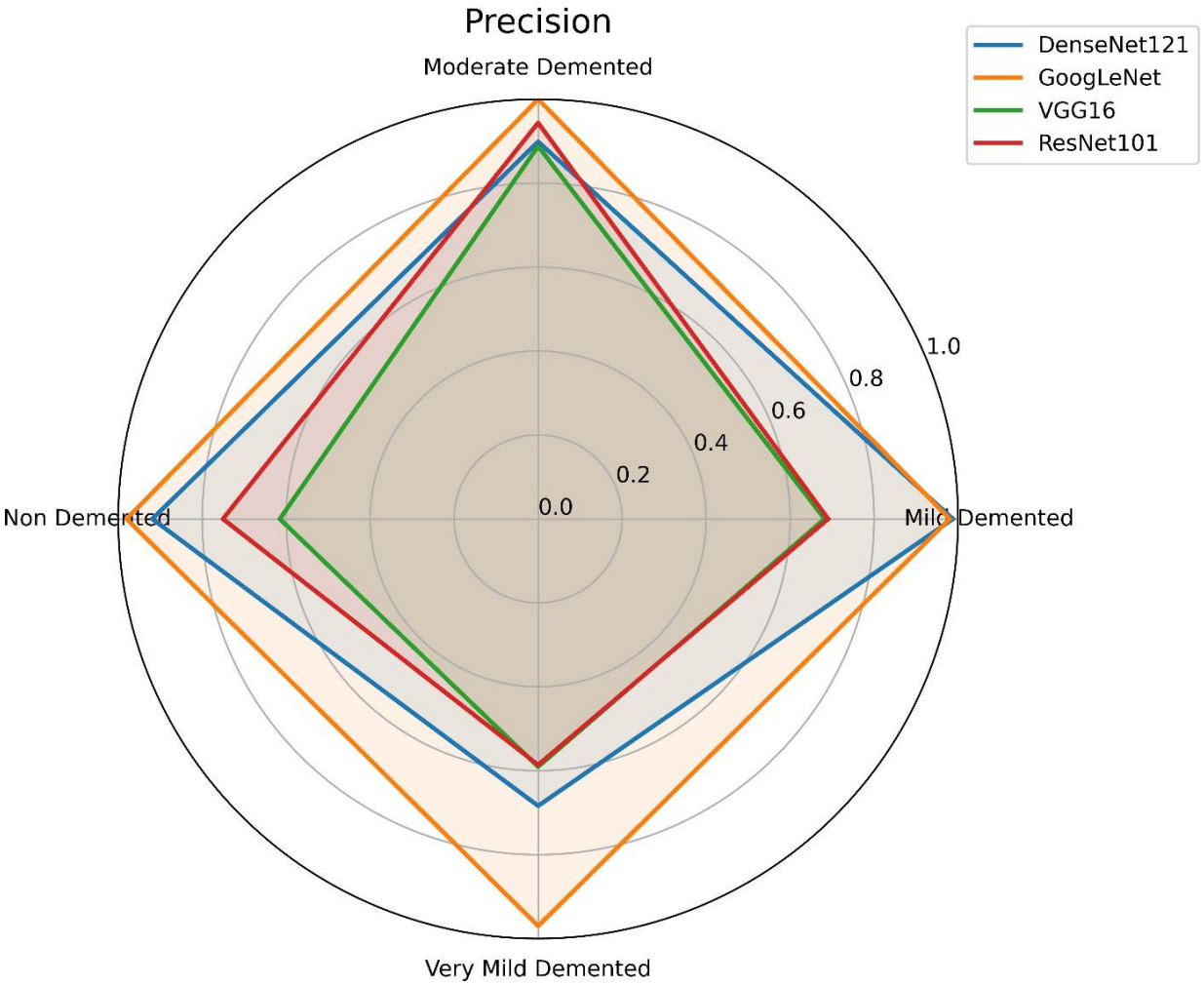


Figure 9 Comparative Radar Chart of Precision Across Transfer Learning Models

A radar chart of the performance of four transfer learning models, including DenseNet121, GoogLeNet, VGG16, and ResNet101, in relation to the four diagnostic groups of AD, Non-Demented, Very Mild-Demented, Mild-Demented, and Moderate-

Demented is shown in figure 4.2. Precision measures the level of how each model can detect true positive cases with minimal false positive cases and therefore it is a vital measure of reliability in medical diagnosis. It is evident in the chart that GoogLeNet (orange line) is the best among all categories of dementia and the outer polygon is almost a perfect polygon just short of the 1.0 mark. This shows that GoogLeNet will be able to differentiate diseased and non-diseased cases with high precision. DenseNet121 (blue line) is also a good performer as it retains a high level of precision particularly on the Mild and Moderate Demented classes. ResNet101 (red line) has moderate accuracy indicating that it has a larger false-positive rate than GoogLeNet and DenseNet121. VGG16 (green line) has the worst accuracy of all classes, especially in the Very Mild and Non-Demented groups, which implies that it has a false positive with non-demented samples. By and large, the radar chart supports the fact that GoogLeNet produces the most precise and reliable classifications, and with a slight difference, DenseNet121 has to be mentioned as the ones that have rather weak results in differentiating the stages of dementia.

1.10.6.4 Radar Chart Comparison of Class-Wise F1-Score

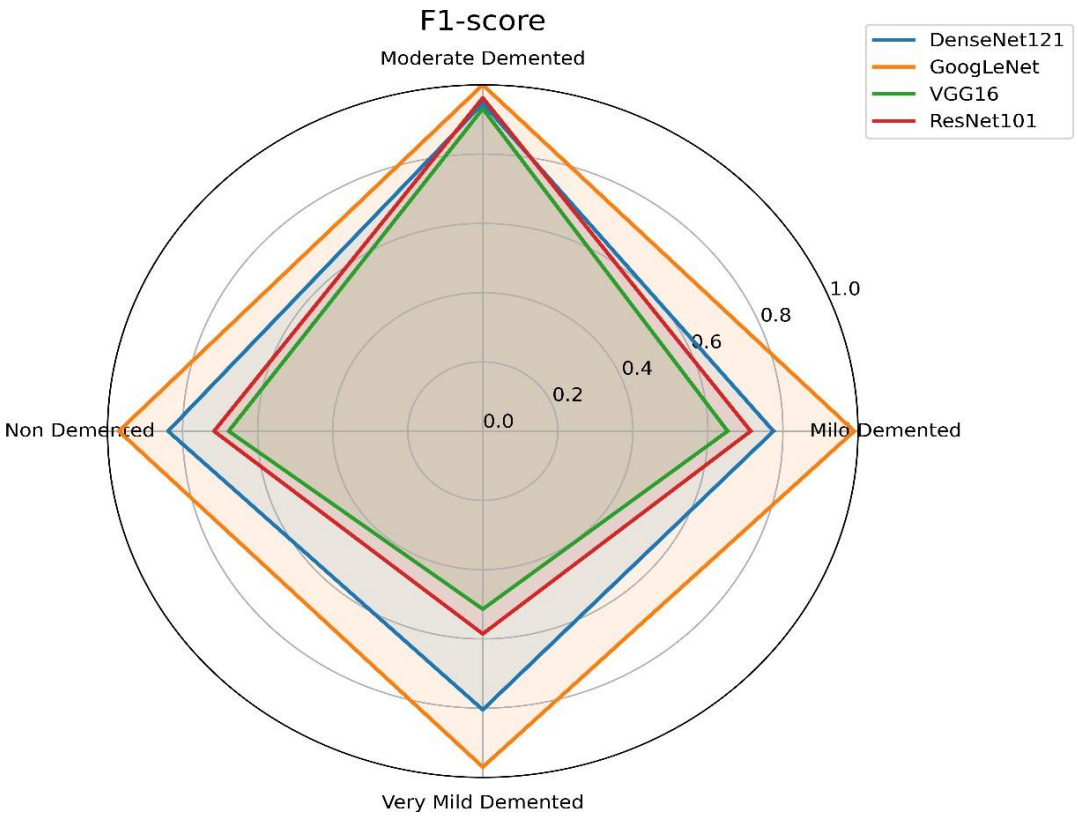


Figure 10 Comparative Radar Chart of F1-Score Across Transfer Learning Models

Figure 10 represents a radar chart of four transfer learning models, DenseNet121, GoogLeNet, VGG16 and ResNet101, when applied in the four diagnostic sub-categories of the AD of Non-Demented, Very Mild Demented, Mild Demented, and Moderate Demented. The F1-score is a harmonic mean of the precision and recall values, which provides a balanced score of the accuracy of a model in identifying the TP and reducing the FP and FN. In this figure, GoogLeNet (orange line) continuously attains the maximum F1-scores in all classes, which creates a wide and homogenous outer polygon that demonstrates high overall results and stability in multi-class classification. DenseNet121 (blue line) is also performing well, keeping the F1-scores relatively high in all of the categories with slight deviations, especially in Non-Demented and Very Mild Demented categories. ResNet101 (red line) shows a moderate performance with a report of

inconsistencies in the precision and the recall values. The lowest F1-scores are obtained with VGG16 (green line) especially in the Non-Demented and Very Mild Demented classes, which proves its inability to balance correct detections with misclassifications. Generally, the radar chart demonstrates that GoogLeNet is the strongest and the best-generalized model, then we have DenseNet121, with ResNet101 and VGG16 having relatively lower stability and efficiency with respect to classifying the various stages of AD.

1.10.7 Grad-CAM and LIME Visualization

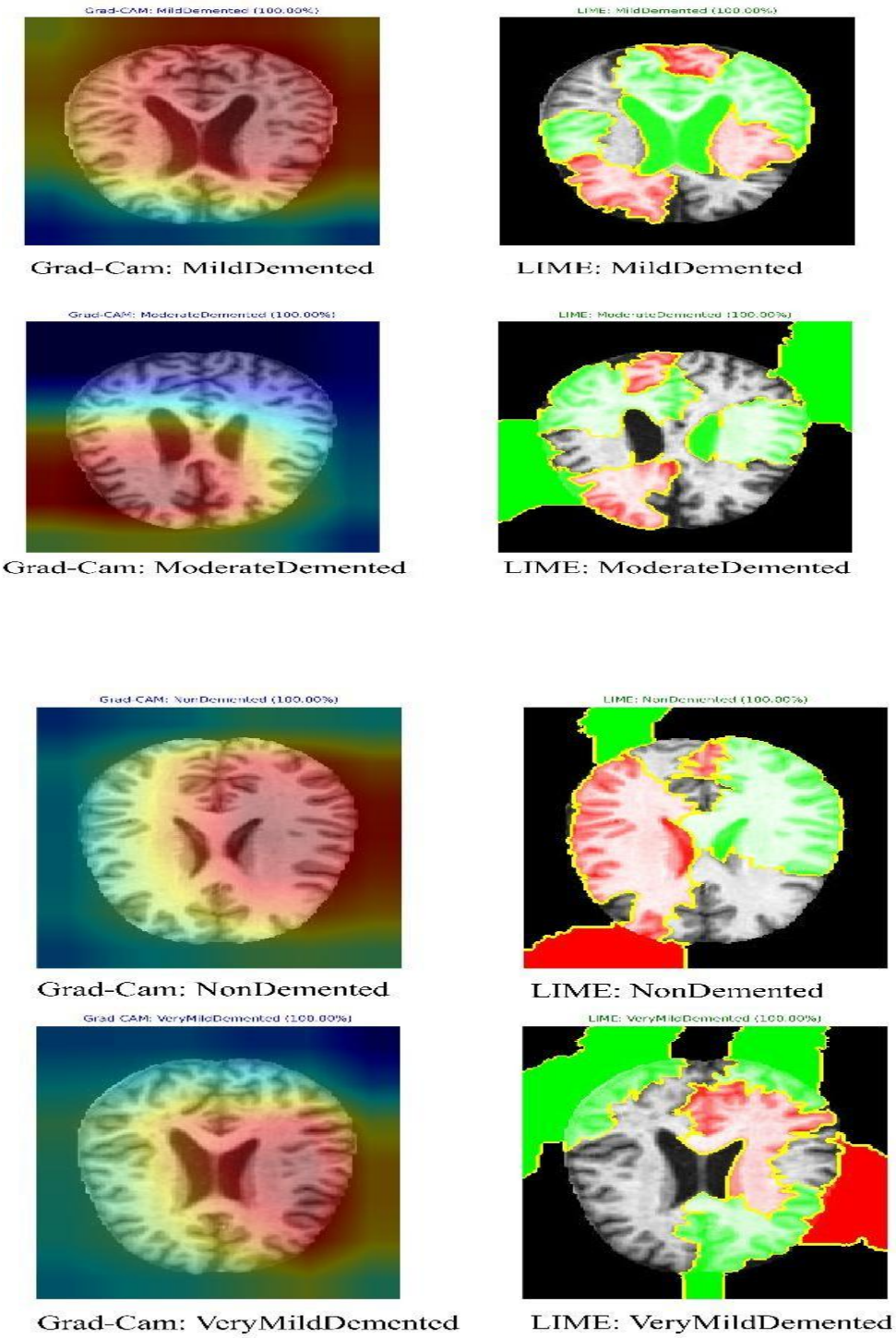


Figure 11 Grad-CAM and LIME Visualization for MRI-based Alzheimer's Disease Classification

The results of explainability visualizations of the four categories of AD, namely, Mild Demented, Moderate Demented, Non-Demented, and Very Mild Demented, are provided in figure 11 using Gradient-weighted Class Activation Mapping (Grad-CAM) and Local Interpretable Model-Agnostic Explanations (LIME). The most effective GoogLeNet model that was used to create these visual explanations delivered an overall accuracy of 0.98 and AUC of 1.00 in all classes. These visualizations are meant to understand how the model finds and concentrates on disease-relevant regions in MRI scans and thereby confirm that its predictions are guided by clinically meaningful brain structures and not irrelevant features.

In the Mild Demented case (top row), Grad-CAM heatmap shows important activations (red and yellow areas) around the medial temporal lobe and hippocampal areas which are established biomarkers of early-stage Alzheimer-related pathology. These areas carry out the memory processing and are one of the first to degenerate because of neurodegenerative processes. This finding is supported by the LIME visualization that indicates the same regions with green color, which depicts positive contributions towards the correct classification of the model. A high degree of consistency between Grad-CAM and LiME shows that the model is rightly concerned with early structural degradation patterns of mild dementia.

In case of the Moderate Demented (second row), both the visual explanation depicts larger activity, especially in the ventricular and parietal cortical areas. Grad-CAM map shows extensive red and orange areas and it demonstrates that the model pays attention to the areas that are characterized by ventricular enlargement and loss of gray matter which are characteristic symptoms of the disease at the advanced stages. In line with this, the LIME map indicates several green-filled spots in both the cortex and subcortical areas, meaning that the model identified many atrophic areas to classify it. The prediction confidence (100 percent) and space correlation of these highlighted regions attain the known medical imaging results of moderate progression of dementia.

The Non-Demented visualization (third row) has a different pattern since Grad-CAM displays predominantly cooler color areas (blue and light yellow) which means that the high-activation areas are not many. This implies that the model failed to identify pathological traits which are common to the AD. This is further supported in the LIME explanation, since in the image there are sparse green parts with fewer neutral parts which indicates that the model identified the brain structure as normal and healthy. The view of the magnetism of balanced gray and white matter boundaries suggests that the model accurately characterizes normal brain anatomy without being overfit to non-disease-specific characteristics.

In the case of Very Mild Demented (bottom row), the Grad-CAM map shows a moderate response at the location of the hippocampus and posterior cingulate cortex, which are typically associated with the initial functional dysfunction in memory and orientation. The red and yellow heat zones also show slight variations that are utilized by the model to identify early cognitive deterioration. The visualization of the corresponding LIME visualization determines that similar regions are green and the features that had a positive impact on the classification. The fact that both interpretability methods overlap indicates that the model reasonably represents the onset trends of atrophy that distinguishes between the cases of Very Mild Demented and Non-Demented subjects despite the fact that structural variations are marginal.

1.11 Discussion

This study shows that transfer learning with pre-trained CNNs can classify four stages of AD from MRI with very good performance. Among the four models, GoogLeNet performed best, with overall accuracy around 0.98 and F1-scores between 0.97 and 1.00 for all classes, as well as the lowest mean categorical hinge loss. DenseNet121 showed moderate performance, while ResNet101 and VGG16 were clearly weaker under the same training and data conditions. Explainability analysis using Grad-CAM and LIME indicated that the better models mainly focused on clinically relevant regions such as the hippocampus, temporal lobes, and ventricular areas. This overlap with known Alzheimer's pathology suggests that the networks are learning meaningful medical features rather than relying on spurious patterns. Overall, a GoogLeNet-based transfer learning framework appears to provide a strong balance of accuracy, stability, and interpretability for four-class MRI-based Alzheimer's staging.

In Bangladesh, dementia is rising with population aging, but early diagnosis is limited by low awareness, few specialists, and concentration of services in large cities. An automated, explainable MRI-based system could support radiologists and physicians by flagging scans suggestive of Very Mild or Mild Demented stages, especially in centers where specialist neurology support is not always available. Because the model is based on transfer learning, it can be deployed on relatively modest hardware, which suits many Bangladeshi hospitals. At the same time, the present model is trained on international data and must be validated and fine-tuned on Bangladeshi MRI scans before real clinical use. Local studies are needed to check performance across different scanners, protocols, and patient profiles. If this is done carefully, an AI-assisted tool of this kind could help improve early detection, guide referral decisions, and support more equitable dementia care in Bangladesh within the constraints of existing resources.

CONCLUSION

1.12 Overview

DL–based analysis of MRI images was applied in this thesis to classify four stages of AD: Non-Demented, Very Mild, Mild, and Moderate Demented. Four transfer learning models, GoogLeNet, DenseNet121, ResNet101, and VGG16, were trained and evaluated using the same dataset, preprocessing steps, and training strategy. Among these, GoogLeNet showed the strongest overall performance, with high accuracy and balanced F1-scores for all classes, and the lowest mean categorical hinge loss, indicating confident and well-separated predictions. In addition to numerical performance, the work focused on the interpretability of the models. Explainable AI techniques, Grad-CAM and LIME, were used to highlight important brain regions, and the most discriminative areas were mainly found around the hippocampus, temporal lobes, and ventricular spaces, which are consistent with known Alzheimer's pathology. These findings suggest that the proposed framework is not only accurate but also clinically meaningful, and therefore has the potential to act as a decision-support tool, especially in resource-limited environments such as Bangladesh.

1.13 Limitations

This work is based on a single publicly available MRI dataset and does not include images from hospitals in Bangladesh or other local sources. As a result, the model may not fully capture variations in scanners, imaging protocols, and patient demographics, so performance may drop on new data. Only four CNN-based transfer learning architectures were considered, and other promising models, such as transformer-based or more advanced hybrid approaches, were not explored. The study also focused only on structural MRI and did not include PET or clinical variables, and the system has not yet been tested in a real clinical environment, so its practical usability and acceptance remain unknown.

1.14 Future Work

External validation and local adaptation on MRI data from different hospitals, especially within Bangladesh, are needed to check whether the model generalizes well to new settings. Testing and fine-tuning on local datasets will help the system handle different scanners, protocols, and patient profiles and is essential before routine clinical use. The modeling approach can be extended by exploring newer deep architectures, such as transformer-based networks, stronger ensembles, and multi-modal fusion that combines MRI with PET or relevant clinical and cognitive information. Comparative experiments using a larger set of models under one consistent experimental setup would provide a clearer view of which techniques are most suitable for multi-class Alzheimer's staging. Clinical integration also requires further development, including simple and user-friendly interfaces, connection to hospital PACS and information systems, and pilot studies with radiologists and neurologists. Attention should be given to ethical, privacy, and fairness issues so that AI-assisted diagnosis is used responsibly and benefits patients in both urban and rural areas.

1.15 Conclusion

This study develops a unified framework where four transfer learning architectures GoogLeNet, DenseNet121, ResNet101, and VGG16 are implemented and evaluated for four-class MRI-based classification of AD (Non-Demented, Very Mild, and Moderate Demented) under the same dataset, preprocessing pipeline, and training strategy, using accuracy, F1-score, AUC, and mean categorical hinge loss for fair comparison. The results highlight GoogLeNet as an effective backbone, achieving about 0.98 accuracy, macro F1-score close to 0.98, AUC near 1.00, and the lowest hinge loss (around 0.05), indicating confident separation between stages. Explainable AI techniques, Grad-CAM and LIME, are integrated to visualize model decisions, with saliency maps consistently focusing on clinically relevant regions such as the hippocampus, temporal cortex, and ventricular spaces, which strengthens clinical trust by reducing the "black box" concern. The work also links these technical outcomes to the healthcare context of Bangladesh and similar resource-limited settings, suggesting a practical pathway for adapting an accurate

and interpretable GoogLeNet-based framework as a decision-support tool for earlier Alzheimer's detection after proper validation on local MRI data.

REFERENCES

[1] M. A. Deture and D. W. Dickson, "The neuropathological diagnosis of Alzheimer's disease," *Molecular Neurodegeneration 2019 14:1*, vol. 14, no. 1, pp. 1–18, Aug. 2019, doi: 10.1186/S13024-019-0333-5.

[2] A. Kumar, J. Sidhu, F. Lui, and J. W. Tsao, "Alzheimer Disease," *StatPearls*, pp. 1–27, Feb. 2024, Accessed: Nov. 11, 2025. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK499922/

[3] "2024 Alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 20, no. 5, p. 3708, May 2024, doi: 10.1002/ALZ.13809.

[4] F. Hasanzadeh, C. B. Josephson, G. Waters, D. Adedinsewo, Z. Azizi, and J. A. White, "Bias recognition and mitigation strategies in artificial intelligence healthcare applications," *npj Digital Medicine 2025 8:1*, vol. 8, no. 1, pp. 1–13, Mar. 2025, doi: 10.1038/s41746-025-01503-7.

[5] "2025 Alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 21, no. 4, p. e70235, Apr. 2025, doi: 10.1002/ALZ.70235.

[6] T. Guo, D. Zhang, Y. Zeng, T. Y. Huang, H. Xu, and Y. Zhao, "Molecular and cellular mechanisms underlying the pathogenesis of Alzheimer's disease," *Molecular Neurodegeneration 2020 15:1*, vol. 15, no. 1, pp. 1–37, Jul. 2020, doi: 10.1186/S13024-020-00391-7.

[7] M. Stoiljkovic, T. L. Horvath, and M. Hajós, "Therapy for Alzheimer's disease: Missing Targets and Functional Markers?," *Ageing Res Rev*, vol. 68, p. 101318, Jul. 2021, doi: 10.1016/J.ARR.2021.101318.

[8] G. Arsenault-Lapierre, T. X. Bui, M. Le Berre, H. Bergman, and I. Vedel, "Rural and urban differences in quality of dementia care of persons with dementia and caregivers across all domains: a systematic review," *BMC Health Serv Res*, vol. 23, no. 1, p. 102, Dec. 2023, doi: 10.1186/S12913-023-09100-8.

[9] S. Murali *et al.*, "Bringing MRI to low- and middle-income countries: Directions, challenges and potential solutions," *NMR Biomed*, vol. 37, no. 7, Jul. 2024, doi: 10.1002/NBM.4992.

[10] H. K. Kondaveeti and C. G. Simhadri, "Evaluation of deep learning models using explainable AI with qualitative and quantitative analysis for rice leaf disease detection," *Scientific Reports 2025 15:1*, vol. 15, no. 1, pp. 1–28, Aug. 2025, doi: 10.1038/s41598-025-14306-3.

[11] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults," *J Cogn Neurosci*, vol. 19, no. 9, pp. 1498–1507, Sep. 2007, doi: 10.1162/JOCN.2007.19.9.1498.

[12] D. P. Veitch *et al.*, "The Alzheimer's Disease Neuroimaging Initiative in the era of Alzheimer's disease treatment: A review of ADNI studies from 2021 to 2022," *Alzheimer's and Dementia*, vol. 20, no. 1, pp. 652–694, Jan. 2024, doi: 10.1002/ALZ.13449;WEBSITE:WEBSITE:ALZ-JOURNALS;ISSUE:ISSUE:DOI.

[13] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, "Handling imbalanced medical datasets: review of a decade of research," *Artif Intell Rev*, vol. 57, no. 10, Oct. 2024, doi: 10.1007/S10462-024-10884-2.

[14] P. Kalavathi and V. B. S. Prasath, "Methods on Skull Stripping of MRI Head Scan Images—a Review," *J Digit Imaging*, vol. 29, no. 3, p. 365, Jun. 2015, doi: 10.1007/S10278-015-9847-8.

[15] S. Roy, J. A. Butman, and D. L. Pham, "Robust skull stripping using multiple MR image contrasts insensitive to pathology," *Neuroimage*, vol. 146, pp. 132–147, Feb. 2017, doi: 10.1016/J.NEUROIMAGE.2016.11.017.

[16] M. F. Safdar, S. S. Alkobaisi, and F. T. Zahra, "A Comparative Analysis of Data Augmentation Approaches for Magnetic Resonance Imaging (MRI) Scan Images of Brain Tumor," *Acta Informatica Medica*, vol. 28, no. 1, p. 29, Mar. 2020, doi: 10.5455/AIM.2020.28.29-36.

[17] H. Li, M. Habes, D. A. Wolk, and Y. Fan, "A deep learning model for early prediction of Alzheimer's disease dementia based on hippocampal MRI," *Alzheimers Dement*, vol. 15, no. 8, p. 1059, Aug. 2019, doi: 10.1016/J.JALZ.2019.02.007.

[18] M. Khatun, M. M. Islam, H. Rahman Rifat, M. S. Bin Shahid, M. A. Talukder, and M. A. Uddin, "Hybridized Convolutional Neural Networks and Long Short-Term Memory for Improved Alzheimer's Disease Diagnosis from MRI Scans," *2023 26th International Conference on Computer and Information Technology, ICCIT 2023*, Mar. 2023, doi: 10.1109/ICCIT60459.2023.10441274.

[19] M. Taiyeb Khosroshahi *et al.*, "Explainable Artificial Intelligence in Neuroimaging of Alzheimer's Disease," *Diagnostics*, vol. 15, no. 5, p. 612, Mar. 2025, doi: 10.3390/DIAGNOSTICS15050612/S1.

[20] I. Oiza-Zapata and A. Gallardo-Antolín, "Alzheimer's Disease Detection from Speech Using Shapley Additive Explanations for Feature Selection and Enhanced Interpretability," *Electronics 2025, Vol. 14, Page 2248*, vol. 14, no. 11, p. 2248, May 2025, doi: 10.3390/ELECTRONICS14112248.

[21] N. A. Mohsin and M. H. Abdulameer, "Evaluating the Impact of 2D MRI Slice Orientation and Location on Alzheimer's Disease Diagnosis Using a Lightweight Convolutional Neural Network," *J Imaging*, vol. 11, no. 8, p. 260, Aug. 2025, doi: 10.3390/JIMAGING11080260.

[22] R. B. Sumona, J. P. Biswas, A. Shafkat, M. M. Rahman, M. O. Faruk, and Y. Majeed, "An integrated deep learning approach for enhancing brain tumor diagnosis," *Healthcare Analytics*, vol. 8, p. 100421, Dec. 2025, doi: 10.1016/J.HEALTH.2025.100421.

[23] Y. Kochura *et al.*, "Batch Size Influence on Performance of Graphic and Tensor Processing Units During Training and Inference Phases," *Advances in Intelligent Systems and Computing*, vol. 938, pp. 658–668, 2020, doi: 10.1007/978-3-030-16621-2_61.

[24] X. Larriva-Novo, M. Vega-Barbas, V. A. Villagrá, D. Rivera, M. Álvarez-Campana, and J. Berrocal, "Efficient Distributed Preprocessing Model for Machine Learning-Based Anomaly Detection over Large-Scale Cybersecurity Datasets," *Applied Sciences 2020, Vol. 10, Page 3430*, vol. 10, no. 10, p. 3430, May 2020, doi: 10.3390/APP10103430.

[25] K. Nagrecha, L. Liu, P. Delgado, and P. Padmanabhan, "InTune: Reinforcement Learning-based Data Pipeline Optimization for Deep Recommendation Models," Aug. 2023, Accessed: Nov. 11, 2025. [Online]. Available: http://arxiv.org/abs/2308.08500

[26] R. Hao, K. Namdar, L. Liu, M. A. Haider, and F. Khalvati, "A Comprehensive Study of Data Augmentation Strategies for Prostate Cancer Detection in Diffusion-Weighted MRI Using Convolutional Neural Networks," *J Digit Imaging*, vol. 34, no. 4, p. 862, Aug. 2021, doi: 10.1007/S10278-021-00478-7.

[27] S. Dardouri, "An efficient method for early Alzheimer's disease detection based on MRI images using deep convolutional neural networks," *Front Artif Intell*, vol. 8, p. 1563016, 2025, doi: 10.3389/FRAI.2025.1563016.

[28] M. Heenaye-Mamode Khan, P. Reesaul, M. M. Auzine, and A. Taylor, "Detection of Alzheimer's disease using pre-trained deep learning models through transfer learning: a review," *Artif Intell Rev*, vol. 57, no. 10, Oct. 2024, doi: 10.1007/S10462-024-10914-Z.

[29] N. Hasan, Y. Bao, A. Shawon, and Y. Huang, "DenseNet Convolutional Neural Networks Application for Predicting COVID-19 Using CT Image," *SN Comput Sci*, vol. 2, no. 5, Sep. 2021, doi: 10.1007/S42979-021-00782-7.

[30] P. B. N. Simangunsong, P. Sihombing, S. Efendi, and F. Fahmi, "DENSENET DEVELOPMENT WITH SQUEEZE-AND-EXCITATION BLOCK FOR TOMATO PLANT DISEASE CLASSIFICATION," *Eastern-European Journal of Enterprise Technologies*, vol. 2, no. 2(134), pp. 28–38, 2025, doi: 10.15587/1729-4061.2025.323176.

[31] S. A. Albelwi, "Deep Architecture based on DenseNet-121 Model for Weather Image Recognition," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 13, no. 10, p. 2022, Accessed: Nov. 11, 2025. [Online]. Available: www.ijacsa.thesai.org

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 770–778, Dec. 2016, doi: 10.1109/CVPR.2016.90.

[33] S. A. Hasanah, A. A. Pravitasari, A. S. Abdullah, I. N. Yulita, and M. H. Asnawi, "A Deep Learning Review of ResNet Architecture for Lung Disease Identification in CXR Image," *Applied Sciences 2023, Vol. 13, Page 13111*, vol. 13, no. 24, p. 13111, Dec. 2023, doi: 10.3390/APP132413111.

[34] K. Pani, K. Sha, V. Palade, B. Bouziane, and A. Alshammari, "Construction of VGG16 Convolution Neural Network (VGG16_CNN) Classifier with NestNet-Based Segmentation Paradigm for Brain Metastasis Classification," *Sensors 2022, Vol. 22, Page 8076*, vol. 22, no. 20, p. 8076, Oct. 2022, doi: 10.3390/S22208076.

[35] J. Tao, Y. Gu, J. Z. Sun, Y. Bie, and H. Wang, “Research on vgg16 convolutional neural network feature classification algorithm based on Transfer Learning,” *CISS 2021 - 2nd China International SAR Symposium*, 2021, doi: 10.23919/CISS51089.2021.9652277.

[36] S. Çelik, B. Doğanlı, M. Ü. Şaşmaz, and U. Akkucuk, “Accuracy Comparison of Machine Learning Algorithms on World Happiness Index Data,” *Mathematics 2025, Vol. 13, Page 1176*, vol. 13, no. 7, p. 1176, Apr. 2025, doi: 10.3390/MATH13071176.

[37] M. Kane, “The Precision of Measurements,” *Applied Measurement in Education*, vol. 9, no. 4, pp. 355–379, 1996, doi: 10.1207/S15324818AME0904_4.

[38] D. Martin Ward Powers, “EVALUATION: FROM PRECISION, RECALL AND F-MEASURE TO ROC, INFORMEDNESS, MARKEDNESS & CORRELATION,” *Article in Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011, doi: 10.9735/2229-3981.

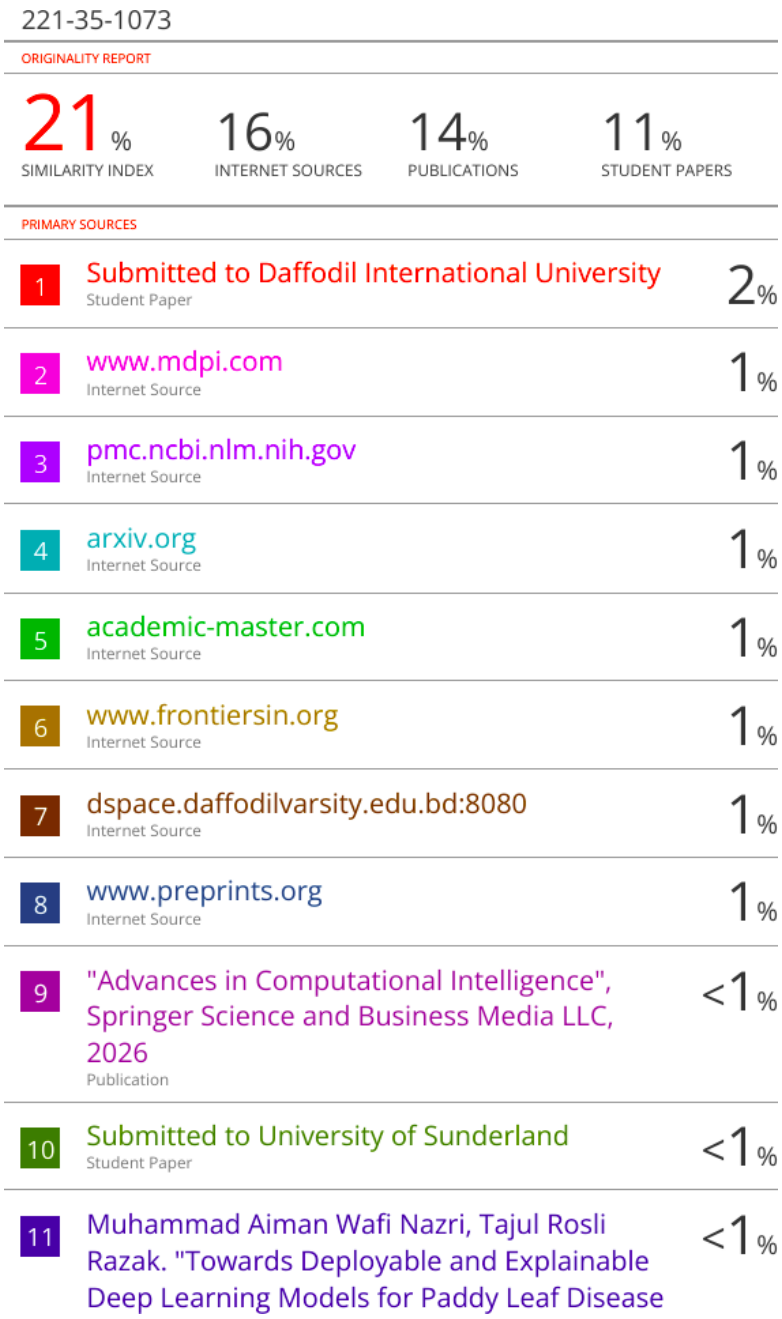
[39] J. Li, “Area under the ROC Curve has the most consistent evaluation for binary classification,” *PLoS One*, vol. 19, no. 12 December, Dec. 2024, doi: 10.1371/JOURNAL.PONE.0316019.

APPENDICES

Appendix A: Alzheimer MRI Dataset

Dataset Link: https://www.kaggle.com/datasets/uraninjo/augmented-alzheimer-mri-dataset

Plagiarism Report



221-35-1073

ORIGINALITY REPORT

21%	16%	14%	11%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	2%
2	www.mdpi.com Internet Source	1%
3	pmc.ncbi.nlm.nih.gov Internet Source	1%
4	arxiv.org Internet Source	1%
5	academic-master.com Internet Source	1%
6	www.frontiersin.org Internet Source	1%
7	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
8	www.preprints.org Internet Source	1%
9	"Advances in Computational Intelligence", Springer Science and Business Media LLC, 2026 Publication	<1%
10	Submitted to University of Sunderland Student Paper	<1%
11	Muhammad Aiman Wafi Nazri, Tajul Rosli Razak. "Towards Deployable and Explainable Deep Learning Models for Paddy Leaf Disease	<1%

Account Clearance