Ensemble Transfer Learning of Alzheimer's Disease (AD) Using MRI Scans

A major project report submitted in partial fulfillment for the award of degree of

Bachelor of Technology

in

Computer Science & Engineering

by

Shiven Aggarwal [2021a1r118]

Andleeb Afroz [2021a1r085]

Safder Abbas [2021a1r152]

Sajid Bhatt [2021a1r175]

Vansh Sharma [2022a1l005]

Under the supervision of

Dr. XXXXXXX
Assistant Professor, CSE

DEPARTMENT OF COMPUTER SCIENCES AND ENGINEERING

MODEL INSTITUTE OF ENGINEERING AND TECHNOLOGY

JAMMU, J&K, INDIA

BATCH 2021-2025

MIET/CSE/PR/.../2024-25

Model Institute of Engineering & Technology, Jammu

Certificate

This is to certify that this Major Project entitled **Ensemble Transfer Learning of Alzheimer's Disease (AD) Using MRI Scans** is a bonafide work **s**ubmitted to the Model Institute of Engineering & Technology, Jammu in partial fulfillment of the requirements for the award of the degree of "Bachelors of Technology" in Computer Science & Engineering.

(Name and sign) Guide	(Name and sign) External Examiner

College Seal

(Name and sign) Internal Examiner	(Name and sign) HOD,CSE

Certificate of Approval of Examiners

The Major Project report entitled **Ensemble Transfer Learning of Alzheimer's Disease (AD) Using MRI Scans** by **t**he team is approved for the award of Bachelors Of Technology Degree in **Computer Science & Engineering**.

Internal Examiner

External Examiner

Date:
Place: Jammu

Acknowledgement

We are deeply grateful to all those who have contributed to the successful completion of our minor project, "Ensembled Transfer Learning for Alzheimer's Disease Using MRI Scans." This project would not have been possible without the unwavering guidance, support, and encouragement we received from various individuals and institutions.

First and foremost, we would like to express our heartfelt gratitude to the Head of the Department of Computer Science and Engineering, Dr. Navin Mani Upadhyay, and our project guide, Ms. Vishalika, for their invaluable guidance, constructive feedback, and constant encouragement throughout this project. Their expertise and insights have been instrumental in shaping the direction and execution of this work.

We are also thankful to the faculty members of the Department of Computer Science and Engineering at MIET Jammu for providing us with the resources, knowledge, and inspiration necessary to undertake this project. Their support has been a cornerstone of our academic journey.

A special note of appreciation goes to our peers and colleagues, whose valuable discussions and collaborative spirit have greatly enriched our learning experience. This project has been a transformative learning experience, and we are profoundly grateful for the opportunity to contribute to innovative advancements in medical imaging and Alzheimer's disease diagnosis using deep learning techniques.

DECLARATION

We, the undersigned, hereby declare that this written submission represents our ideas and work in our own words, and where others' ideas or work have been included, we have adequately cited and referenced the original sources.

We further declare that we have adhered to all principles of academic honesty and integrity, and have not misrepresented, fabricated, or falsified any idea, data, fact, or source in this submission.

We understand that any violation of the above will be grounds for disciplinary action by the Institute and may also evoke penal action from the sources that have not been properly cited or from whom proper permission has not been obtained where required.

Team Members:
Shiven Aggarwal (2021A1R118)
Andleeb Afroz (2021A1R085)
Safder Abbas (2021A1R152)
Sajid Bhat (2021A1R175)
Vansh Sharma (2022A1L005)

Date:
Place: Jammu

Abstract

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder that significantly impacts cognitive abilities and memory, ultimately diminishing an individual's quality of life. It is the most common cause of dementia worldwide and poses a growing challenge for healthcare systems, given its increasing prevalence in aging populations. Early and accurate diagnosis is critical to improving intervention strategies and managing the disease effectively. However, existing diagnostic methods often face limitations in sensitivity and specificity, especially during the early stages of the disease. Addressing this gap, this study introduces a novel ensemble transfer learning-based approach for diagnosing AD using Magnetic Resonance Imaging (MRI) scans, leveraging cutting-edge deep learning techniques to improve diagnostic accuracy.

The proposed framework utilizes pre-trained deep learning models to extract meaningful features from T1-weighted structural MRI sequences, mitigating challenges posed by traditional diagnostic techniques and existing artificial intelligence-based solutions. By adopting a transfer learning approach, the system capitalizes on the generalizability of pre-trained models to analyze medical imaging data effectively. The experiment was conducted using data from the OASIS-3 archive, a widely recognized repository in Alzheimer's research. The dataset included MRI scans from 113 subjects, with 247 images of cognitively normal individuals and 410 images from individuals with mild cognitive impairment. These were categorized based on the Clinical Dementia Rating (CDR) scale, a robust tool for evaluating dementia severity. Preprocessing steps involved normalizing MRI images in the axial plane to ensure uniformity and enhance model performance.

Eight state-of-the-art pre-trained models, including VGG19, DenseNet201, and EfficientNetV2S, were fine-tuned for feature extraction. Ensemble techniques, such as weighted averaging and simple averaging, were employed to combine predictions from multiple models, thereby improving the overall diagnostic accuracy. The ensemble approach addresses individual model weaknesses while leveraging their strengths to produce more robust and reliable predictions. Additionally, the study explored the effects of using cropped versus uncropped MRI images, revealing insights into the optimal preprocessing strategies for AD diagnosis.

The results demonstrated significant improvements in early-stage Alzheimer's diagnosis, achieving higher diagnostic precision compared to traditional methods and existing AI-based approaches. Metrics such as accuracy, sensitivity, and specificity showed marked enhancement, underscoring the efficacy of the ensemble transfer learning framework. Notably, the qualitative differences between cropped and uncropped MRI images underscored the importance of selecting appropriate preprocessing techniques for maximizing diagnostic performance.

This study not only advances the state of the art in medical imaging for neurodegenerative disorders but also establishes a foundation for integrating such AI-driven tools into clinical workflows. The ensemble transfer learning framework offers a reliable, efficient, and objective approach to diagnosing Alzheimer's Disease, with the potential to support medical practitioners in making informed decisions.

Future directions include expanding the framework to incorporate multimodal data, such as genetic information and clinical history, and testing its performance across larger and more diverse datasets. This study marks a significant step toward leveraging artificial intelligence to address one of the most pressing challenges in neurodegenerative disease management, ultimately contributing to better patient outcomes and advancements in clinical research.

Contents

Certificate i

Certificate of Approval of Examiners ii

Acknowlegement iii

Declaration iv

Abstract vi

List of Figures ix

List of Tables x

Abbreviations xi

1 Introduction 1

1.1 Overview 2
1.2 Objective 3
1.3 Scope of the Project 5
1.4 Problem Statement 7
1.5 Related Work 10

2 Deep Learning Techniques and Challenges in AD Diagnosis 13

2.1 Deep Convolutional Neural Networks (CNNs) 13
2.2 Hybrid Deep Learning Frameworks 15

2.3 Generative Adversarial Networks (GANs) 16
2.4 Ensemble Learning Techniques 18
2.5 Challenges in Deep Learning for AD Diagnosis 20

3 Ensemble Learning with Transfer Learning for Alzheimer's Disease Classification 24
3.1 Methodology 25
3.2 Dataset Preprocessing 26
3.3 Evaluation Metrics for Alzheimer's Disease Diagnosis . . . 29
3.3.1 Accuracy 29
3.3.2 Sensitivity (Recall or True Positive Rate) 29
3.3.3 Specificity (True Negative Rate) 29
3.3.4 Area Under the Curve (AUC) 30
3.3.5 Evaluation Process 30
3.3.6 Confusion Matrix 31
3.4 Model Selection 31
3.5 Model Ensembling Technique 35

4 Results 39
4.1 Conclusion 41
4.2 Future Directions 42

Future Work 42

References 44

List of Figures

1.1 Projected AD Population and Diagnosis. 3

3.1 WorkFlow 28
3.2 Accuracy and Loss using Simple Average 37
3.3 ROC Curve using Simple Average 37
3.4 Accuracy and Loss using Weighted Average 38
3.5 ROC Curve using Weighted Average 38

List of Tables

3.1 Accuracy Scores for Alzheimer's Disease Diagnosis 32
3.2 Ensemble Method Results 36

List Of Abbreviations

AD	Alzheimer's Disease
MRI	Magnetic Resonance Imaging
CDR	Clinical Dementia Rating
CNN	Convolutional Neural Network
ReLU	Rectified Linear Unit
SVM	Support Vector Machine
GBMs	Gradient Boosting Machines
GANs	Generative Adversarial Networks
LRP	Layer-wise Relevance Propagatio
T P	True Positives
T N	True Negatives
F P	False Positive
F N	False Negative
AUC	Area Under the Curve

Chapter 1

Introduction

The Ensembled Transfer Learning for Alzheimer's Disease Diagnosis project introduces a transformative approach to addressing one of the most pressing challenges in modern healthcare: the early and accurate diagnosis of Alzheimer's Disease (AD). AD, a progressive neurodegenerative disorder, affects millions globally, with significant impacts on patients, caregivers, and healthcare systems. Early intervention and management of AD are critical for improving patient outcomes, but existing diagnostic methods often fall short in terms of accuracy, accessibility, and scalability. This project leverages advanced artificial intelligence (AI) techniques, specifically ensemble transfer learning, to create a robust and reliable diagnostic framework using Magnetic Resonance Imaging (MRI) scans.

The introduction outlines the core concept and objectives of the project, emphasizing its goal of enhancing diagnostic precision through the integration of multiple pre-trained deep learning models. By employing transfer learning, the project overcomes limitations associated with traditional machine learning methods, such as reliance on handcrafted features and difficulty in generalizing across datasets. The use of ensemble methods further strengthens the framework by combining predictions from diverse models, leading to improved accuracy and reliability.

Key objectives of the project include the development of a scalable, efficient, and objective diagnostic tool that can aid medical practitioners in identifying early stages of Alzheimer's Disease. This is achieved by utilizing state-of-the-art models like VGG19, DenseNet201, and Efficient-

NetV2S, fine-tuned for feature extraction from MRI images. Additionally, the project aims to explore the impact of preprocessing techniques, such as cropping and normalization, on diagnostic performance to identify optimal strategies for data preparation.

The introduction also highlights the broad applicability of this framework, which extends beyond Alzheimer's Disease to potential applications in diagnosing other neurodegenerative disorders. By leveraging the OASIS-3 archive, a comprehensive dataset of MRI scans and clinical information, the project ensures a strong foundation for research and development. The integration of innovative AI methodologies positions this framework as a significant advancement in medical imaging, bridging the gap between research and real-world clinical application.

This section concludes by underscoring the potential of the ensemble transfer learning framework to transform the diagnostic landscape for Alzheimer's Disease. By enabling early detection and accurate diagnosis, the project not only contributes to improving patient care but also opens new avenues for research and innovation in the field of neurodegenerative diseases. The scalability and adaptability of the proposed solution underscore its role in fostering a more efficient, accessible, and impactful approach to medical diagnostics.

Overall, the introduction serves as a comprehensive overview of the Mentor Connect project, setting the stage for a deeper exploration of the methodology, implementation, and outcomes in the subsequent sections of the report.

1.1 Overview

Alzheimer's disease is a brain disorder that slowly destroys memory and thinking skills and, eventually, the ability to carry out the simplest tasks. In most people with the disease — those with the late-onset type symptoms first appear in their mid-60s. Early-onset Alzheimer's occurs between a person's 30s and mid-60s and is very rare. Alzheimer's disease is the most common cause of dementia among older adults . The disease is named af-

ter Dr. Alois Alzheimer. In 1906, Dr. Alzheimer noticed changes in the brain tissue of a woman who had died of an unusual mental illness. Her symptoms included memory loss, language problems, and unpredictable behaviour. After she died, he examined her brain and found many abnormal clumps (now called amyloid plaques) and tangled bundles of fibres (now called neurofibrillary, or tau, tangles). These plaques and tangles in the brain are still considered some of the main features of Alzheimer's disease. Another feature is the loss of connections between nerve cells (neurons) in the brain. Neurons transmit messages between different parts of the brain, and from the brain to muscles and organs in the body. Many other complex brain changes are thought to play a role in Alzheimer's, too. This damage initially takes place in parts of the brain involved in memory, including the entorhinal cortex and hippocampus. It later affects areas in the cerebral cortex, such as those responsible for language, reasoning, and social behaviour. Eventually, many other areas of the brain are damaged. paraphrase it.

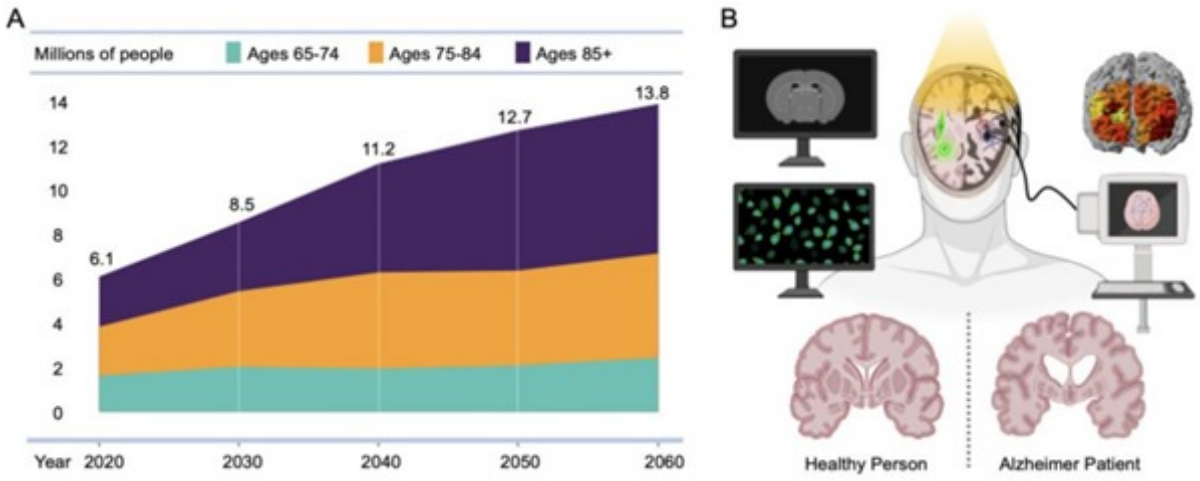

Figure 1.1: Projected AD Population and Diagnosis.

1.2 Objective

The primary objective of this project is to develop an innovative, accurate, and efficient diagnostic framework for Alzheimer's Disease (AD), addressing key limitations of existing diagnostic methods. AD, being a progressive

neurodegenerative disorder, often goes undetected in its early stages due to the limitations of traditional techniques. These methods are time-intensive, subjective, and reliant on specialized expertise, which can delay intervention. Furthermore, existing AI-based approaches, while promising, often require high computational resources or fail to deliver sufficient diagnostic accuracy for clinical reliability.

This project aims to bridge these gaps by introducing an ensemble transfer learning-based framework that significantly reduces diagnostic time, computational overhead, and inaccuracies. The specific objectives of the project are outlined below:

Timely and Accurate Diagnosis: Enable early detection of Alzheimer's Disease by leveraging MRI imaging and advanced AI techniques. The framework focuses on recognizing subtle structural changes in the brain associated with the early stages of AD, facilitating timely medical intervention.

Reduction in Computational Complexity: Design a solution that operates efficiently on moderate computational resources. By fine-tuning pre-trained deep learning models, the framework minimizes the need for large-scale computations without compromising on diagnostic precision.

Enhanced Diagnostic Precision through Ensembles: Utilize ensemble learning methods such as weighted averaging and simple averaging to combine the outputs of multiple pre-trained models. This approach ensures a higher degree of accuracy by leveraging the strengths of individual models, mitigating their weaknesses, and delivering robust predictions.

Comprehensive Data Utilization: Explore the impact of various preprocessing techniques, such as cropping and normalization of MRI images, to identify optimal strategies for extracting meaningful features. This ensures that the diagnostic model effectively utilizes the available data for improved outcomes.

Versatility and Scalability: Develop a flexible framework capable of being extended to other neurodegenerative diseases and medical imaging domains. The scalability of the model allows it to be adapted for different

datasets and use cases, making it a valuable tool for diverse clinical applications.

Overcoming Challenges of Existing Methods: Address critical shortcomings of current diagnostic tools, such as the dependency on expensive infrastructure, lack of accessibility in resource-constrained settings, and suboptimal early detection rates. The project prioritizes practical implementation to ensure its applicability in real-world healthcare scenarios.

1.3 Scope of the Project

The scope of this project encompasses the development, evaluation, and potential application of an ensemble transfer learning-based diagnostic framework for Alzheimer's Disease (AD) using MRI scans. The framework aims to address key challenges in the existing diagnostic processes, such as long diagnostic times, high computational requirements, and inaccuracies in early detection. The project focuses on improving the efficiency, accuracy, and scalability of Alzheimer's Disease diagnosis, with potential applicability in both clinical and research settings. The specific scope of this project includes the following:

Development of Diagnostic Framework: The core of the project involves designing a machine learning framework based on ensemble transfer learning. The system utilizes pre-trained deep learning models like VGG19, DenseNet201, and EfficientNetV2S, which are fine-tuned on the OASIS-3 dataset of MRI scans to extract relevant features for AD detection. This diagnostic framework is intended to provide an accurate and rapid solution for identifying early-stage Alzheimer's Disease.

Utilization of MRI Data: The project primarily focuses on using structural MRI scans, particularly T1-weighted MRI sequences, from the OASIS-3 dataset, which contains data from cognitively normal individuals and those with mild cognitive impairment. This ensures the model is trained to recognize the subtle structural changes in the brain associated with AD, thereby facilitating early-stage detection.

Preprocessing and Data Preparation:A significant aspect of the

project is exploring various preprocessing techniques such as cropping, normalization, and axial plane alignment of MRI images. The preprocessing phase ensures that the models receive high-quality input data, which is crucial for accurate feature extraction and effective diagnosis.

Ensemble Learning Integration: The project involves combining multiple models using ensemble techniques, such as weighted and simple averaging, to improve diagnostic performance. This methodology allows the strengths of each individual model to contribute to a more accurate and robust output, reducing the likelihood of false positives or false negatives.

Evaluation of Diagnostic Accuracy and Efficiency: One of the key components of this project is evaluating the performance of the ensemble transfer learning model against traditional diagnostic methods, as well as existing AI-based approaches. The model will be assessed on metrics such as accuracy, sensitivity, specificity, and computational efficiency. The goal is to demonstrate a significant improvement in early-stage AD detection and overall diagnostic accuracy.

Comparative Analysis of Cropped vs. Uncropped Images: The project includes an analysis of how different preprocessing methods, such as cropping MRI images to focus on the brain regions of interest, affect the model's diagnostic accuracy. This exploration helps identify the most effective image processing techniques for enhancing model performance.

Scalability and Flexibility for Other Applications: While the primary focus is on Alzheimer's Disease, the framework is designed to be adaptable and scalable for diagnosing other neurodegenerative diseases and conditions that may show similar structural changes in MRI scans. The flexibility of the model allows for future extension to other medical imaging applications, such as brain tumors or Parkinson's Disease.

Implementation in Clinical and Research Settings: Although this project is centered around developing an AI-based tool, the long-term goal is to make the model applicable in real-world clinical settings. By reducing the computational demands and improving diagnostic efficiency, the framework could be implemented in hospitals, research institutions,

and healthcare centers with varying levels of resources. This would help improve early diagnosis and care for patients with Alzheimer's Disease, particularly in resource-constrained environments.

Potential for Global Healthcare Impact: The project has the potential to contribute to the global fight against Alzheimer's Disease by making early and accurate diagnosis more accessible. With its efficient use of computational resources and scalable framework, the model could be deployed in regions with limited healthcare infrastructure, ensuring that patients worldwide benefit from improved diagnostic tools.

In summary, the scope of this project extends beyond the development of an AD diagnostic model to include its evaluation, comparison with existing methods, and exploration of its broader applications in medical imaging. By addressing critical limitations of current diagnostic techniques, the project aims to contribute significantly to advancements in AI-driven healthcare solutions and neurodegenerative disease management.

1.4 Problem Statement

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder that leads to severe cognitive decline and memory loss, significantly affecting the quality of life of patients and their families. The disease primarily impacts the elderly population, and with an aging global demographic, the prevalence of AD is increasing rapidly. Early and accurate diagnosis is essential for effective intervention and management, as early-stage detection allows for timely therapeutic measures that may slow disease progression. However, traditional diagnostic methods, including clinical assessments, neuropsychological tests, and neuroimaging, are time-consuming, subjective, and often lack the ability to detect AD in its earliest stages, when treatments could be most beneficial.

Current clinical practices rely heavily on the expertise of medical professionals to analyze MRI scans and other imaging data. However, even with advancements in neuroimaging technologies, the detection of subtle brain changes associated with early-stage AD remains a challenge. MRI scans

can provide valuable structural information, but interpreting these images requires expert knowledge and is subject to human variability. Furthermore, current methods are often unable to accurately differentiate between normal age-related changes and the early signs of AD, leading to misdiagnosis or delayed diagnosis, both of which can have significant consequences for patient care.

Furthermore, although artificial intelligence (AI)-based diagnostic methods have emerged as promising tools in detecting AD from neuroimaging data, these methods face several significant limitations:

High Computational Costs: Many AI models, especially deep learning-based approaches, require significant computational resources for training and inference. These models demand high-performance computing infrastructure, including specialized hardware like Graphics Processing Units (GPUs), which may not be accessible in resource-limited settings or smaller healthcare facilities. As a result, the widespread adoption of AI in clinical practices is hindered by its prohibitive cost.

Inaccurate Early Detection: Despite significant progress, current AI models still struggle with the detection of AD in its early stages. Early-stage AD often manifests in subtle brain changes that are difficult to capture, and existing models are not fine-tuned to identify these minute differences effectively. This leads to false negatives, where the disease is not detected in time to provide effective interventions.

Limited Generalization Across Datasets: A significant challenge for AI-based diagnostic models is their limited ability to generalize across different datasets. Models trained on one set of MRI scans may not perform well on scans from different populations, clinical settings, or imaging protocols. This lack of robustness compromises the model's reliability when deployed in diverse, real-world healthcare environments, leading to inconsistencies in diagnostic outcomes.

Dependence on Expert Interpretation: While AI can assist in detecting patterns in MRI images, many AI-based systems still require expert interpretation to confirm the diagnosis. This dependency introduces hu-

man bias and errors into the diagnostic process, particularly when dealing with complex or ambiguous cases.

Long Processing Times: Traditional diagnostic methods and some AI models can be time-intensive, especially when processing large datasets of high-resolution MRI images. Extended processing times can delay diagnosis and intervention, making it challenging for healthcare providers to offer timely care.

This project aims to address these challenges by developing an ensemble transfer learning-based approach for diagnosing Alzheimer's Disease using MRI scans. The goal is to leverage pre-trained deep learning models, which have been fine-tuned for different aspects of image analysis, to create a more efficient, scalable, and accurate diagnostic framework. By combining the strengths of multiple models through ensemble techniques, the project seeks to improve diagnostic accuracy, reduce computational costs, and enhance early-stage detection.

Specifically, the project aims to:

Improve Early Detection: By fine-tuning pre-trained models and integrating them in an ensemble framework, the project will increase the sensitivity of early AD detection, identifying even the most subtle signs of the disease in its early stages.

Reduce Computational Complexity: The ensemble transfer learning approach will optimize the performance of pre-trained models to reduce the computational resources required for analysis, making it more feasible for widespread clinical adoption, especially in resource-limited settings.

Increase Accuracy and Reliability: Ensemble methods will be employed to improve the diagnostic accuracy by combining the predictions from multiple models, ensuring a more robust and reliable diagnosis compared to individual models.

Enable Wider Accessibility: By designing a solution that is computationally efficient and highly accurate, the project aims to make AD diagnosis more accessible to hospitals and clinics globally, including those with limited infrastructure.

The overarching goal is to create a diagnostic framework that minimizes time delays, computational overhead, and inaccuracies associated with traditional and current AI-based methods. This innovative approach will contribute to improving the quality of care and support early diagnosis, allowing for better management of Alzheimer's Disease and potentially providing new avenues for future research and development in medical imaging and AI-driven diagnostics. Through this work, we aim to advance the field of medical AI and provide a practical solution to one of the most pressing challenges in neurodegenerative disease diagnosis today.

1.5 Related Work

The classification of Alzheimer's Disease (AD) has seen significant advancements with the integration of multimodal brain imaging data and deep learning methodologies. Particularly, convolutional neural networks (CNNs) have revolutionized diagnostic and predictive models for AD by extracting complex features from neuroimaging datasets, enabling earlier and more precise detection of the disease.

Traditionally, AD diagnosis relied on clinical assessments and neuropsychological tests supported by imaging modalities like structural magnetic resonance imaging (sMRI) or positron emission tomography (PET). However, these methods were often limited in sensitivity and specificity, especially for detecting early-stage AD. The advent of CNNs addressed these shortcomings by automating feature extraction and identifying intricate patterns in imaging data, significantly improving diagnostic accuracy.

CNNs excel at analyzing imaging data by capturing hierarchical features. Initial layers detect basic patterns like edges, while deeper layers identify complex abnormalities related to brain structure. This makes CNNs particularly effective for analyzing MRI and PET scans in AD classification. Furthermore, hybrid models combining CNNs with traditional machine learning techniques, such as support vector machines (SVM) or random forests, have demonstrated improved performance by leveraging CNNs for feature extraction and machine learning classifiers for robust

generalization.

Recent advancements include the development of multimodal frameworks integrating data from MRI, PET, functional MRI (fMRI), and electroencephalography (EEG). This approach addresses the multifaceted nature of AD by combining structural and functional insights, leading to more comprehensive and accurate diagnostic systems. Additionally, multi-scale feature fusion networks have emerged, offering the ability to analyze global patterns (e.g., brain atrophy) and local abnormalities (e.g., hippocampal shrinkage), enhancing sensitivity and specificity.

Innovative techniques like generative adversarial networks (GANs) have also been applied to AD research. GANs are employed to generate synthetic neuroimaging data, augmenting limited datasets, or improving image quality and denoising scans. These capabilities are particularly useful in addressing data scarcity and variability, common challenges in medical imaging.

Another notable approach involves stacked deep polynomial networks (SDPNs), which combine polynomial functions with deep learning for capturing non-linear relationships in imaging data. These networks have shown promise in distinguishing between healthy individuals and AD patients by effectively modeling high-dimensional features.

Despite these advancements, challenges persist. Generalization remains a significant hurdle, as models trained on specific datasets often underperform when applied to data from different clinical settings. Transfer learning, which fine-tunes pre-trained models on smaller domain-specific datasets, has been proposed to improve generalization across diverse populations.

Interpretability is another concern, as deep learning models often function as black boxes. Techniques like saliency maps and attention mechanisms are being used to highlight brain regions influencing predictions, aiding clinicians in understanding model outputs.

Data scarcity, especially for rare AD stages, poses additional difficulties. Semi-supervised and unsupervised learning techniques, which leverage un-

labeled data, have been explored to address this limitation. Moreover, integrating clinical data, such as cognitive test scores and genetic information, with imaging features has shown potential in creating holistic models for AD classification.

In summary, the application of deep learning, particularly CNNs, to AD classification has enabled significant improvements in diagnostic accuracy and early detection. However, addressing challenges related to generalization, interpretability, data scarcity, and the integration of clinical information remains critical for advancing research and clinical implementation. These efforts promise to refine AD classification further, offering new opportunities for timely interventions and improved patient outcomes.

Chapter 2

Deep Learning Techniques and Challenges in AD Diagnosis

2.1 Deep Convolutional Neural Networks (CNNs)

A Convolutional Neural Network (CNN), also known as ConvNet, is a specialized type of deep learning algorithm mainly designed for tasks that necessitate object recognition, including image classification, detection, and segmentation. CNNs are employed in a variety of practical scenarios, such as autonomous vehicles, security camera systems, and others. **The importance of CNNs** There are several reasons why CNNs are important in the modern world, as highlighted below: CNNs are distinguished from classic machine learning algorithms such as SVMs and decision trees by their ability to autonomously extract features at a large scale, bypassing the need for manual feature engineering and thereby enhancing efficiency. The convolutional layers grant CNNs their translation-invariant characteristics, empowering them to identify and extract patterns and features from data irrespective of variations in position, orientation, scale, or translation. A variety of pre-trained CNN architectures, including VGG-16, ResNet50, Inceptionv3, and EfficientNet, have demonstrated top-tier performance. These models can be adapted to new tasks with relatively little data through a process known as fine-tuning. Beyond image classification tasks, CNNs are versatile and can be applied to a range of other domains, such as natural language processing, time series analysis, and speech recognition.

Key Components of a CNN The convolutional neural network is made of four main parts. But how do CNNs Learn with those parts? They help the CNNs mimic how the human brain operates to recognize patterns and features in images:

Convolutional layers: This is the first building block of a CNN. As the name suggests, the main mathematical task performed is called convolution, which is the application of a sliding window function to a matrix of pixels representing an image. The sliding function applied to the matrix is called kernel or filter, and both can be used interchangeably. In the convolution layer, several filters of equal size are applied, and each filter is used to recognize a specific pattern from the image, such as the curving of the digits, the edges, the whole shape of the digits, and more. Put simply, in the convolution layer, we use small grids (called filters or kernels) that move over the image. Each small grid is like a mini magnifying glass that looks for specific patterns in the photo, like lines, curves, or shapes. As it moves across the photo, it creates a new grid that highlights where it found these patterns.

Rectified Linear Unit (ReLU for short): A ReLU activation function is applied after each convolution operation. This function helps the network learn non-linear relationships between the features in the image, hence making the network more robust for identifying different patterns. It also helps to mitigate the vanishing gradient problems.

Pooling layers: The goal of the pooling layer is to pull the most significant features from the convoluted matrix. This is done by applying some aggregation operations, which reduce the dimension of the feature map (convoluted matrix), hence reducing the memory used while training the network. Pooling is also relevant for mitigating overfitting.

Fully connected layers: These layers are in the last layer of the convolutional neural network, and their inputs correspond to the flattened one-dimensional matrix generated by the last pooling layer. ReLU activations functions are applied to them for non-linearity.

Convolution layers The development and implementation of artifi-

cial intelligence (AI) models for early diagnosis heavily depend on large, annotated datasets that represent a wide range of patient characteristics. However, the availability of such data for early-stage Alzheimer's diagnosis is limited. This scarcity of high-quality, annotated data hampers the ability of AI algorithms to learn and generalize accurately. Consequently, the use of AI for early detection remains constrained, and models may not perform well across diverse populations or subtle disease variations.

2.2 Hybrid Deep Learning Frameworks

Hybrid deep learning frameworks integrate the strengths of deep learning models, such as Convolutional Neural Networks (CNNs), with traditional machine learning algorithms, like support vector machines (SVMs) or random forests, to create more robust and accurate AD classification systems. These models aim to leverage the capabilities of CNNs in feature extraction while capitalizing on the generalization power of traditional machine learning classifier.

Why Hybrid Approaches Are Valuable in AD Classification

Feature Extraction with CNNs: CNNs are adept at processing large-scale imaging data, such as MRI or PET scans, and automatically learning high-level, complex features without the need for manual intervention. This ability to extract intricate spatial patterns in images makes CNNs highly effective for diagnosing neurodegenerative diseases, including AD, where subtle structural and functional changes in the brain are critical.

Classification with Traditional Machine Learning Models: While CNNs excel at feature extraction, they may not always perform optimally in directly mapping these features to class labels, especially in cases where the dataset is small or complex. Traditional machine learning models, such as SVMs, decision trees, or random forests, can complement CNNs by serving as classifiers that generalize the learned features to produce more reliable predictions. These models often perform better on datasets with fewer samples or when data variability is high.

Examples of Hybrid Models in AD Classification CNN + SVM: One common implementation is using a CNN to extract features from neuroimaging scans, which are then inputted into an SVM for final classification. The SVM is particularly useful due to its ability to find a high-dimensional hyperplane that optimally separates classes, even when feature spaces are complex and non-linear.

CNN + Random Forests: A CNN can provide a rich feature representation, which is then processed by a random forest classifier. This ensemble method can handle non-linear relationships and interactions between features, enhancing the model's performance in identifying AD indicators.

CNN + Gradient Boosting Machines (GBMs): This combination can be particularly powerful. CNNs extract complex features from brain scans, and a GBM, such as XGBoost, uses these features to make sequential corrections to classification errors, resulting in a refined and more accurate model.

2.3 Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) have introduced new possibilities for medical image analysis, including the classification of Alzheimer's Disease (AD). GANs consist of two neural networks: a generator and a discriminator, which are trained together in a competitive setting. The generator aims to create synthetic data that resembles real data, while the discriminator attempts to distinguish between real and generated data. This setup fosters the generation of high-quality, realistic data, which can be leveraged for a variety of applications in AD classification.

Concept and Importance of GANs

Data Augmentation: One of the primary applications of GANs in AD classification is the generation of synthetic brain images. This is particularly useful for augmenting small datasets, addressing the issue of data scarcity, and improving the training of deep learning models.

Data Enhancement: GANs can enhance the quality of imaging data

by generating clearer, denoised images that provide better input for classification models. This can be valuable when working with noisy or low-quality data from clinical settings.

Progression Prediction: GANs can be used to simulate the progression of AD from one stage to another, providing insights into the potential future state of a patient's brain, which can be helpful for long-term prognosis and treatment planning.

Techniques in GANs for AD

1.Conditional GANs (cGANs):

Controlled Data Generation: cGANs extend the original GAN architecture by conditioning the generator on additional information, such as disease labels or patient demographics, enabling the creation of more targeted synthetic data.

Application: In AD classification, cGANs can generate brain scans that simulate different stages of AD, helping to train models on a broader range of conditions.

2.CycleGANs:

Unsupervised Image Translation: CycleGANs are used to convert images from one domain to another without paired training data. In the context of AD, CycleGANs can translate non-AD brain images to their counterparts in a diseased state, allowing for domain adaptation and improved model generalization.

Use Case: This is especially useful for aligning data from different imaging protocols or scanner types, enhancing model training across varied datasets.

3.GAN-based Image Denoising and Quality Enhancement:

Noise Reduction: GANs can learn to remove noise from brain images, producing cleaner data for analysis. This improves the quality of input images, enabling models to better identify subtle features associated with AD.

Super-Resolution GANs (SRGANs): These can enhance the resolution of low-quality scans, making it easier for models to detect small but

crucial changes related to AD.

2.4 Ensemble Learning Techniques

Ensemble learning techniques combine multiple individual models to create a stronger and more accurate predictive system. In the context of Alzheimer's Disease (AD) classification, these techniques can help leverage the strengths of different models to improve diagnostic accuracy, robustness, and generalization. Ensemble methods integrate the outputs of several models to make more accurate predictions, reducing the likelihood of errors that might arise from using a single model.

Concept and Importance of Ensemble Learning

Bias-Variance Tradeoff: Ensemble methods balance the bias-variance tradeoff by combining multiple models, which helps mitigate the limitations of individual models. This results in an overall reduction in model variance without a significant increase in bias.

Improved Generalization: By aggregating the predictions of multiple models, ensemble methods reduce the risk of overfitting and enhance the ability to generalize to new, unseen data. This is crucial for AD classification, where variability between patient data can be significant.

Common Ensemble Learning Techniques

1.Bagging (Bootstrap Aggregating):

Concept: Bagging involves training multiple copies of the same base model on different subsets of the training data, which are generated by bootstrapping (random sampling with replacement). The final prediction is made by aggregating the outputs (e.g., majority voting for classification).

Application in AD: Bagging can be used with models like decision trees or random forests to create a more robust classifier that accounts for variations in input data. Random Forests, in particular, are popular for AD classification due to their ability to handle complex data and prevent overfitting.

Advantages: Reduces overfitting, enhances stability, and improves predictive performance by averaging out errors from individual models.

2.Boosting:

Concept: Boosting sequentially trains models, where each new model corrects the errors made by the previous ones. This technique assigns more weight to incorrectly classified instances so that the model focuses more on difficult cases in subsequent iterations.

Popular Algorithms:

AdaBoost: Adjusts weights based on the performance of previous models and combines them using weighted majority voting.

Gradient Boosting: Builds models sequentially, with each model fitting the residuals of the previous one to minimize the error function iteratively.

Application in AD: Boosting can improve the classification of AD by emphasizing hard-to-classify data points and creating a strong ensemble with enhanced accuracy.

Advantages: Typically leads to higher accuracy compared to bagging but may be prone to overfitting if not carefully tuned.

3. Stacking (Stacked Generalization):

Concept: Stacking involves training multiple different models on the same dataset and then combining their predictions using a meta-learner (e.g., logistic regression or a neural network). The meta-learner learns how to best combine the predictions of the base models to improve overall performance.

Application in AD: Different base models such as support vector machines (SVM), decision trees, and convolutional neural networks (CNNs) can be combined to leverage their unique strengths in capturing patterns from neuroimaging data. The meta-learner then synthesizes the predictions for a final diagnosis.

Advantages: Utilizes the strengths of multiple model architectures and can handle complex, high-dimensional data better than single models.

4. Voting Classifier:

Concept: A simple form of ensemble learning that combines the predictions of multiple models through majority voting (for classification) or

averaging (for regression). Each model makes an independent prediction, and the final output is based on the most common prediction or the average prediction value.

Application in AD: Different models trained on neuroimaging data can be combined, with each contributing a vote to determine whether a patient has AD or not. This approach helps stabilize predictions and reduce the chance of misclassification due to noise.

Advantages: Simple to implement and often effective in combining diverse models with varying strengths.

2.5 Challenges in Deep Learning for AD Diagnosis

Generalization is a critical aspect of deep learning models, especially in the context of Alzheimer's Disease (AD) diagnosis. It refers to the model's ability to perform well on unseen data from different sources or clinical settings. A common challenge in AD classification is that deep learning models trained on specific datasets may not perform well on new data due to differences in imaging protocols, scanner types, or patient demographics. This limits the applicability of these models in real-world clinical environments where variability in data is inevitable. Transfer Learning has emerged as a potential solution to this issue. By utilizing pre-trained models, which have already learned from large, diverse datasets, these models can be fine-tuned on smaller, domain-specific datasets to adapt and improve their performance. This approach enables models to transfer learned features across different sources, enhancing their generalization ability. Additionally, Transfer Learning reduces the need for large amounts of labeled data, which are often scarce in medical research, thus making deep learning models more practical in clinical settings.

Interpretability of Deep Learning Models

Interpretability is another essential factor for the adoption of deep learning models in clinical practice. Deep learning models, particularly convolutional neural networks (CNNs), are known for their ability to learn complex patterns in medical imaging data. However, they often operate

as "black boxes," which makes it difficult for clinicians to understand how the model arrives at its predictions. This lack of transparency poses a significant barrier to trust and clinical acceptance. To address this, several techniques have been developed to improve model interpretability. For example, saliency maps highlight the regions of the brain that most influence the model's decision, allowing researchers and clinicians to visualize the areas of the image that are most relevant for classification. Attention mechanisms enable the model to focus on important regions within the brain, providing further insight into the features driving the model's predictions. Layer-wise Relevance Propagation (LRP) is another technique that decomposes the output of the model to attribute importance to different parts of the input image, making it easier to identify which features contribute the most to the final decision. Improving interpretability is crucial to bridge the gap between AI models and clinical practice, making it easier for healthcare professionals to trust and effectively use AI-based diagnostic tools.

Data Scarcity and Augmentation Techniques

A significant challenge in AD diagnosis using deep learning models is the scarcity of high-quality, annotated data. Collecting and labeling medical imaging data is a time-consuming and expensive process, and available datasets are often small or imbalanced, particularly for rare stages of AD. This scarcity of data limits the effectiveness of deep learning models, as they typically require large, diverse datasets to generalize well. Data augmentation techniques are commonly used to address this issue by artificially increasing the size of the training dataset. These techniques include geometric transformations such as rotation, translation, and flipping, which help models become invariant to changes in image orientation and positioning. Intensity variations, such as adjusting brightness or contrast, enable the model to learn to be robust to variations in image acquisition. Additionally, synthetic data generation methods, like using Generative Adversarial Networks (GANs), can create synthetic medical images that mimic real-world data, enhancing the dataset and addressing data imbalance.

Noise injection, which involves adding random noise to training data, can also make models more resilient to real-world imperfections. By employing these techniques, researchers can overcome the limitations of data scarcity and improve the model's ability to generalize across different datasets.

Integration of Clinical Data

While imaging data plays a pivotal role in AD diagnosis, integrating clinical data alongside neuroimaging features can provide a more comprehensive and informative foundation for diagnosis. Clinical data, such as cognitive test scores, patient medical history, genetic information, and demographic details, can add valuable context to neuroimaging data. This multimodal approach improves the interpretability and predictive accuracy of deep learning models, making them more representative of the complex nature of AD. Integrating clinical and imaging data can be achieved through feature fusion, where clinical features are combined with imaging features at various stages of the model, enhancing the model's understanding of how different data types interact. Multimodal deep learning models, which process both imaging and clinical data, can provide richer representations and improve diagnostic performance. The integration of clinical data also contributes to model interpretability by providing additional context that helps clinicians understand the decision-making process. This combined approach aligns well with the goal of making AI-based tools more transparent and clinically viable, improving diagnosis, treatment plans, and patient outcomes.

Challenges in Deep Learning Models for AD Diagnosis

Despite the advancements in deep learning for Alzheimer's Disease diagnosis, several challenges remain. One of the main issues is data scarcity. The limited availability of high-quality, annotated datasets, particularly for rare AD stages or specific subtypes, hinders the performance of deep learning models. Additionally, these models often require significant computational resources, which can make them impractical in resource-limited settings. Overfitting is another challenge, as deep learning models trained on small or imbalanced datasets may learn noise instead of generalizable

features. Deep learning models also struggle with generalization, as they may not perform well on new datasets that differ from the training data, such as those from different scanners or patient populations. Finally, interpretability remains a barrier to clinical adoption, as deep learning models are often seen as black boxes, which reduces trust and hinders their widespread use. Addressing these challenges is crucial for making deep learning models more effective, reliable, and applicable in real-world clinical environments.

Chapter 3

Ensemble Learning with Transfer Learning for Alzheimer's Disease Classification

Ensemble learning is a technique that combines the predictions of multiple models to enhance accuracy, robustness, and generalization. When applied to transfer learning, it leverages the strengths of pre-trained models to improve classification performance, particularly in complex tasks such as Alzheimer's disease (AD) classification from MRI scans. Combining judgments from multiple models often yields superior outcomes compared to using a single model. This approach can effectively extract subtle patterns from data and enhance generalization, making it highly suitable for medical imaging tasks.

Types of Ensemble Transfer Learning Approaches

1. Simple Averaging In the simple averaging approach, the predictions of all models in the ensemble are treated equally. Each model provides its output (typically a probability score or class label), and the final prediction is calculated as the arithmetic mean of these outputs. This method assumes that all models contribute equally to the decision-making process and is effective when the models have comparable performance.

For instance, consider three transfer learning models—ResNet, DenseNet, and InceptionNet—trained on the same dataset. If each model predicts a probability of 0.7, 0.6, and 0.8 for a specific class, the final ensemble prediction would be the average of these probabilities:

Final Prediction = (0.7 + 0.6 + 0.8) / 3 = 0.7 The simplicity of this method makes it computationally efficient and easy to implement, but it may not perform optimally if there are significant performance variations among the models.

2. Weighted Averaging Weighted averaging assigns different weights to individual models based on their performance metrics, such as validation accuracy, precision, or F1-score. Models that perform better on the validation set are given greater influence in the final prediction. This approach is particularly useful when the ensemble consists of models with varying levels of accuracy or specialization.

For example, suppose the weights assigned to ResNet, DenseNet, and InceptionNet are 0.5, 0.3, and 0.2, respectively, and their predictions for a specific class are 0.7, 0.6, and 0.8. The final prediction would be a weighted average:

Final Prediction = (0.7 × 0.5) + (0.6 × 0.3) + (0.8 × 0.2) = 0.35 + 0.18 + 0.16 = 0.69

Weighted averaging ensures that models with superior predictive power contribute more to the ensemble, thereby enhancing the overall accuracy and robustness.

3.1 Methodology

The general task in AD diagnosis can be framed as a mathematical problem where we aim to map a set of biomarkers, including MRI features, to diagnostic outcomes. Let X represent the set of biomarkers, which could include neuroimaging data, genetic markers, and cerebrospinal fluid (CSF) proteins, and let Y represent the diagnostic labels, where Y = 0 , 1 , with 0 for healthy individuals and 1 for those diagnosed with Alzheimer's. The dataset D consists of N samples, each containing a pair di = (xi,yi) , where x i is the biomarker data for the i-th individual, and y i is the corresponding diagnostic label.

The goal is to develop a predictive model that can accurately map X to Y, predicting whether an individual has Alzheimer's based on the pro-

vided biomarker data. However, training deep learning models on medical datasets, especially those related to Alzheimer's, can be challenging due to the limited size and complexity of the data. To address this, we propose an innovative ensemble learning approach that integrates transfer learning models, aiming to improve the classification performance and robustness of AD detection from MRI scans.

Softmax Function and Multi-Class Classification To improve classification accuracy, the final output layer of our model employs the Softmax function. The Softmax function is used to compute the probability of each class (e.g., cognitively normal vs. Alzheimer's) for a given input. The mathematical form of the Softmax function is:

$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$

where:

$\sigma(z)_i$ represents the probability of class i

e^{z_i} is the exponential of the logit for class i

$\sum_{j=1}^{K} e^{z_j}$ is the sum of the exponentials of all class logits, and

K is the number of classes.

In Alzheimer's classification, the classes typically correspond to the status of the individual: cognitively normal or diagnosed with Alzheimer's Disease. The Softmax function converts raw model predictions (logits) into probabilities, allowing the model to determine the likelihood of each class, facilitating more accurate and confident classifications.

3.2 Dataset Preprocessing

This study utilized data from 113 subjects drawn from the OASIS-3 dataset, a comprehensive resource that includes longitudinal multimodal neuroimaging, clinical, and cognitive data focusing on normal aging and Alzheimer's Disease (AD). The MRI scans in this dataset are available in NIfTI format, and the dataset comprises both cognitively healthy individuals and those diagnosed with mild cognitive impairment (MCI). Specifically, 247 MRI scans were collected from cognitively normal subjects, while 410 scans were from patients diagnosed with MCI. The dataset serves as a valuable

resource for studying the early stages of dementia, particularly in differentiating between normal aging and the onset of Alzheimer's Disease.

For this study, T1-weighted structural MRI sequences were processed in the axial plane of the brain, providing a detailed view of the brain's internal structures. The Clinical Dementia Rating (CDR) scale was used to classify the subjects' dementia status. A CDR score of 0 indicates a cognitively healthy individual, while a score between 0.5 and 1.5 corresponds to mild cognitive impairment, a critical stage in the early detection of Alzheimer's Disease. This system of classification allows for a nuanced understanding of the cognitive trajectories of individuals, helping to inform the development of predictive models for dementia-related disorders.

The initial step in processing the MRI data involved converting the raw NIfTI files into PNG format, which resulted in 255 individual slices per subject from the axial view of the brain. Each subject's MRI scan was divided into multiple slices to capture a comprehensive representation of the brain's structure. These slices were particularly valuable for their ability to reveal critical regions involved in neurodegenerative diseases, such as the hippocampus, which is deeply embedded within the temporal lobe. Additionally, certain areas of the brain in these images showed notable shrinkage, a hallmark of early-stage dementia, particularly in patients with mild cognitive impairment.

To ensure accuracy in the analysis, the central slices of the brain were prioritized during preprocessing. These middle slices, which contain more brain tissue and less interference from bone structures, provide a clearer depiction of the brain's regions of interest. Since brain tissue loss is a key biomarker in dementia, particularly in Alzheimer's Disease, these central slices proved to be more useful for model training, providing more informative data for the neural network. In contrast, slices at the edges of the brain, which are dominated by bone and skull structures, are less informative and can even be misleading in the classification process.

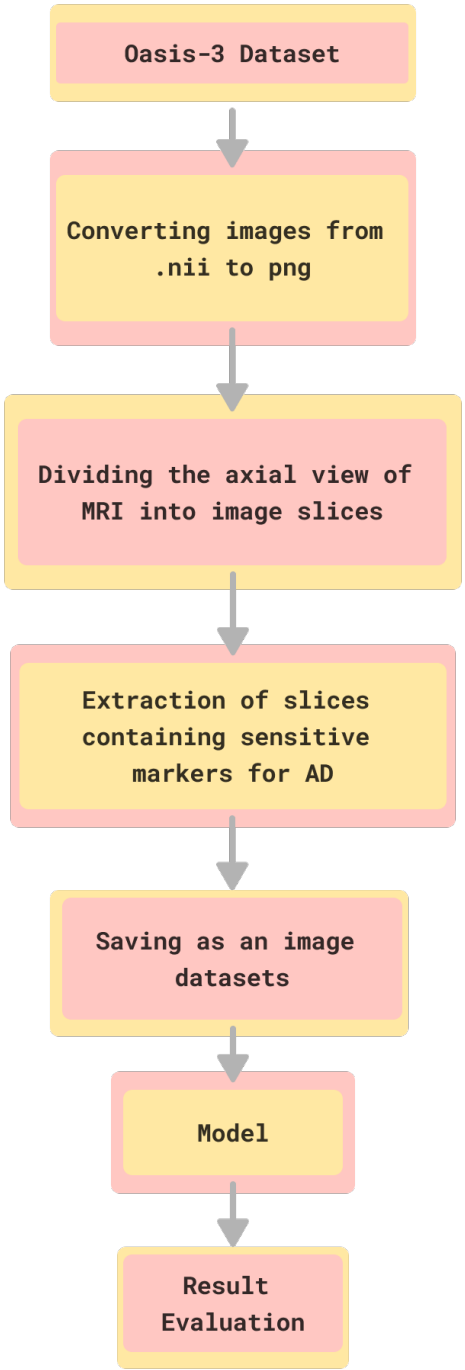

Figure 3.1: WorkFlow

3.3 Evaluation Metrics for Alzheimer's Disease Diagnosis

In the context of evaluating deep learning models for classification tasks, such as Alzheimer's disease diagnosis using MRI scans, the following metrics provide a comprehensive understanding of model performance:

3.3.1 Accuracy

Definition: The ratio of correctly predicted instances (both True Positives and True Negatives) to the total number of instances.

$$\text{Accuracy}(\%) = \frac{TP + TN}{TP + TN + FP + FN} \times 100$$

Importance: While accuracy is a simple metric, it may not always be the best indicator in imbalanced datasets (where one class is overrepresented). In cases where false positives or false negatives are critical, accuracy alone might be insufficient.

3.3.2 Sensitivity (Recall or True Positive Rate)

Definition: The proportion of actual positive instances correctly identified by the model. It reflects the ability of the model to detect positive cases, such as correctly identifying Alzheimer's patients.

$$\text{Sensitivity}(\%) = \frac{TP}{TP + FN} \times 100$$

Importance: A higher sensitivity value means fewer positive cases are missed (i.e., fewer False Negatives). This is particularly critical in medical diagnoses where missing a positive instance (e.g., diagnosing Alzheimer's) could have significant consequences.

3.3.3 Specificity (True Negative Rate)

Definition: The proportion of actual negative instances correctly identified by the model. It measures the model's ability to reject negative cases

correctly, such as identifying healthy individuals.

$$\text{Specificity}(\%) = \frac{TN}{TN + FP} \times 100$$

Importance: Specificity ensures that the model is not misclassifying negative cases as positive, which is particularly valuable when avoiding unnecessary treatments or interventions.

3.3.4 Area Under the Curve (AUC)

Definition: AUC is derived from the Receiver Operating Characteristic (ROC) curve, which plots the True Positive Rate (Sensitivity) against the False Positive Rate (1 - Specificity) at various threshold levels.

Importance: AUC offers a more comprehensive measure of model performance by evaluating its ability to distinguish between positive and negative classes across all decision thresholds. A higher AUC indicates that the model performs well in differentiating between the classes, even if the decision threshold changes.

3.3.5 Evaluation Process

Dataset: The dataset consists of 657 images, with each model evaluated using a 5-fold cross-validation technique. This approach splits the data into 5 subsets, training the model on 4 subsets and testing on the remaining one, ensuring a more generalized performance evaluation.

Training Parameters:

- **Epochs:** 30 epochs were used for model training to ensure adequate convergence of the models.
- **Batch Size:** A batch size of 16 was selected to balance between memory constraints and model performance.
- **Learning Rate:** A fixed learning rate of 0.001 was applied, and the Adam optimizer was used for efficient training and convergence.

3.3.6 Confusion Matrix

For a model evaluated using these metrics, the confusion matrix might show:

Predicted Positive	**Predicted Negative**
Actual Positive	TP (True Positives)
FN (False Negatives)	
Actual Negative	FP (False Positives)
TN (True Negatives)	

Using the confusion matrix, the accuracy, sensitivity, specificity, and AUC score for each model can be calculated, offering insights into areas like:

- How well the model detects positive instances (sensitivity).
- How well it avoids misclassifying negative instances as positive (specificity).
- Overall model performance in distinguishing between classes (AUC).

3.4 Model Selection

This section presents a comprehensive evaluation of several pre-trained deep convolutional neural networks (CNNs) on the OASIS dataset, with results summarized in Table.

The models assessed include VGG19, NASNetLarge, MobileNet, DenseNet201, ResNet152, EfficientNetV2S, InceptionV3, and Xception. Among these, VGG19 emerged as the top performer, achieving an impressive accuracy of 97.7%. In contrast, the NASNetLarge model demonstrated the lowest performance with an accuracy of 91.2%. The other models in the evaluation—MobileNet, DenseNet201, ResNet152, EfficientNetV2S, InceptionV3, and Xception—recorded accuracy scores of 92.2%, 96.1%, 94.6%, 95.2%, 93.6%, and 93.1%, respectively, highlighting the generally high performance across the tested architectures.

Table 3.1: Accuracy Scores for Alzheimer's Disease Diagnosis

Models	**Accuracy**	**Specificity**	**Sensitivity**
Before Cropping			
MobileNet	0.922	0.927	0.913
VGG19	0.977	0.949	0.961
DenseNet201	0.961	0.965	0.958
ResNet152	0.946	0.942	0.911
EfficientNetV2S	0.952	0.966	0.943
InceptionV3	0.936	0.939	0.923
NASNetLarge	0.912	0.927	0.913
After Cropping			
MobileNet	0.915	0.892	0.867
VGG19	0.982	0.973	0.993
DenseNet201	0.974	0.967	0.975
ResNet152	0.966	0.957	0.968
EfficientNetV2S	0.975	0.988	0.969
InceptionV3	0.952	0.947	0.962
NASNetLarge	0.926	0.899	0.901
Xception	0.959	0.944	0.966

Notably, the evaluation process also incorporated an enhancement technique, where the accuracy of the models was improved through a cropping strategy. This technique aims to focus the model's attention on the most relevant features within the input data, reducing noise and potentially improving classification performance. After applying the cropping technique, DenseNet201 was observed to surpass VGG19 in terms of the Area Under the Curve (AUC) score, despite its lower overall accuracy. This observation underscores an important nuance in the assessment of model performance: accuracy and AUC score do not always correlate in a straightforward manner.

The discrepancy between accuracy and AUC performance can be attributed to the fundamental differences in what each metric evaluates. Accuracy is a more direct measure of a model's ability to correctly classify instances, taking into account both the rank order of predictions and the decision threshold for classification. However, AUC, which is derived from the Receiver Operating Characteristic (ROC) curve, evaluates a model's ability to discriminate between classes across all possible thresholds, regardless of the threshold set for final classification. As such, a model may

exhibit high accuracy yet perform poorly on the ROC curve if it is not well-calibrated for distinguishing between positive and negative classes. On the other hand, a model with a higher AUC might show better performance in terms of ranking the predictions, even if its accuracy is somewhat lower.

This distinction is particularly critical in tasks where distinguishing between classes is more important than simply maximizing the number of correct classifications. For instance, in medical imaging or diagnostic applications, the ability of a model to consistently rank positive instances higher than negative ones (even if a few predictions are misclassified) can be more valuable than achieving high accuracy, as it may lead to better decision-making in real-world scenarios.

In summary, while VGG19 demonstrated the highest accuracy in this evaluation, DenseNet201 showed a superior AUC score, indicating better performance in ranking the predicted probabilities. This serves as a reminder that accuracy alone may not fully capture the quality of a model's predictions, particularly in scenarios where class imbalance or decision thresholds play a significant role. The inclusion of multiple evaluation metrics, such as AUC, is thus crucial for a more comprehensive understanding of a model's strengths and weaknesses in complex tasks. The cropping technique employed in this study involved extracting relevant image slices by identifying the largest contour, specifically the region within the skull where the hippocampus resides. As illustrated in Figure **??**, the extreme points of the brain were localized, allowing for a more focused analysis by excluding irrelevant areas. This targeted approach reduces computational complexity by omitting pixels that do not contribute meaningful feature information or contain extraneous space, which could otherwise hinder the performance of the classification model. The unnecessary background space often introduces noise into the image, which can negatively impact model accuracy. To address this challenge, we applied thresholding or Canny edge detection to identify and isolate the continuous boundaries of the brain, particularly the regions with uniform color or intensity. This step enabled precise localization of the image's top, left, bottom, and right

extreme points, ensuring that the model focuses on the most relevant features.

Following the image cropping, the same set of models—trained with the same hyperparameters as in the baseline (without cropping)—was re-evaluated. The results, summarized in Table **??**, demonstrate a marked improvement in both accuracy and AUC scores across all eight models. Notably, VGG19 achieved the highest accuracy of 98.2%, further solidifying its status as a leading architecture in this context. MobileNet, on the other hand, recorded the lowest accuracy score. The other models performed as follows: DenseNet201 achieved an accuracy of 97.4%, ResNet152 scored 96.6%, EfficientNetV2S reached 97.5%, InceptionV3 secured 95.2%, NASNetLarge obtained 92.6%, and Xception rounded out the results with an accuracy of 95.9%.

In terms of AUC, VGG19 again outperformed the others, achieving an outstanding score of 98.6%. This is a strong indication of VGG19's ability to not only correctly classify instances but also effectively rank the predicted probabilities, particularly in distinguishing between the positive and negative classes across various thresholds. In contrast, NASNetLarge attained the lowest AUC score of 91.1%, suggesting a relatively weaker ability to differentiate between classes when considering the model's overall performance across varying decision thresholds.

The observed performance gains with the cropping technique can be attributed to the reduction in irrelevant background noise and the enhanced focus on the critical regions of interest, which are crucial for the accurate identification of features such as the hippocampus. This approach underscores the importance of preprocessing techniques that enhance data quality, enabling deep learning models to perform at their best. The improvements in both accuracy and AUC scores highlight the efficacy of the cropping strategy in optimizing model performance, particularly for tasks involving intricate structures like brain imaging.

In summary, the application of the cropping technique not only improved the efficiency of the models by focusing on relevant brain regions

but also led to significant gains in both classification accuracy and AUC scores. These findings emphasize the importance of thoughtful image preprocessing in enhancing the performance of deep learning models in medical imaging tasks, ultimately providing more robust and reliable results.

3.5 Model Ensembling Technique

The training of the individual models was followed by ensemble strategies aimed at improving predictive performance. Specifically, we explored two distinct ensemble techniques: simple averaging and weighted averaging. Both ensemble methods were implemented under identical experimental conditions to ensure fairness in the comparison. This included the use of a learning rate of 0.0001, the Adam optimizer for optimization, and a training regimen of 100 epochs. The images used in this study were processed using a cropping technique to enhance the quality and relevance of the input data.

To evaluate the efficacy of the ensemble methods, we selected three high-performing models based on their individual accuracies reported in Table 3.2: VGG19, DenseNet201, and EfficientNetV2S. These models were designated as M1, M2, and M3, respectively. Following this, we conducted a series of experiments that involved combining these models in various configurations. Specifically, we evaluated the following combinations: M1 + M2 + M3, M1 + M2, M2 + M3, and M1 + M3. This experimental setup allowed us to assess the impact of different ensemble pairings and their performance under both simple and weighted averaging methods.

The results revealed that the combination of models M2 and M3, when averaged using the simple averaging method, yielded the highest performance, achieving an impressive accuracy of 98.9

On the other hand, the ensemble comprising all three models—M1 (VGG19), M2 (DenseNet201), and M3 (EfficientNetV2S)—achieved the lowest accuracy of 94.1

In addition to accuracy, we also assessed the models based on their Area Under the Curve (AUC), a key metric for evaluating classification models

in terms of their ability to discriminate between classes. The ensemble of M1 + M2, when combined using the simple averaging method, achieved the highest AUC score, indicating that this model pairing was particularly adept at distinguishing between the positive and negative classes. Conversely, the ensemble of M1 + M2 + M3, using the weighted averaging technique, resulted in the lowest AUC score, reinforcing the notion that the inclusion of additional models and the use of inappropriate weighting can hinder performance, particularly in terms of discriminatory power.

In conclusion, our experiments provide valuable insights into the effectiveness of various ensemble strategies for improving model performance. The results clearly demonstrate that simple averaging, particularly with a subset of models like M2 and M3, can lead to superior accuracy and AUC scores compared to more complex combinations or weighted methods. This study highlights the importance of careful model selection and ensemble strategy when seeking to optimize machine learning models for classification tasks.

Table 3.2: Ensemble Method Results

Ensemble Method	**Models**	**Accuracy**	**Specificity**	**Sensitivity**
Simple Average	M1 + M2 + M3	0.971	0.960	0.973
	M1 + M2	0.969	0.958	0.983
	M1 + M3	0.977	0.989	0.969
	M2 + M3	0.989	0.967	0.978
Weighted Average	M1 + M2 + M3	0.941	0.942	0.922
	M1 + M2	0.976	0.961	0.984
	M1 + M3	0.969	0.964	0.936
	M2 + M3	0.981	0.970	0.925

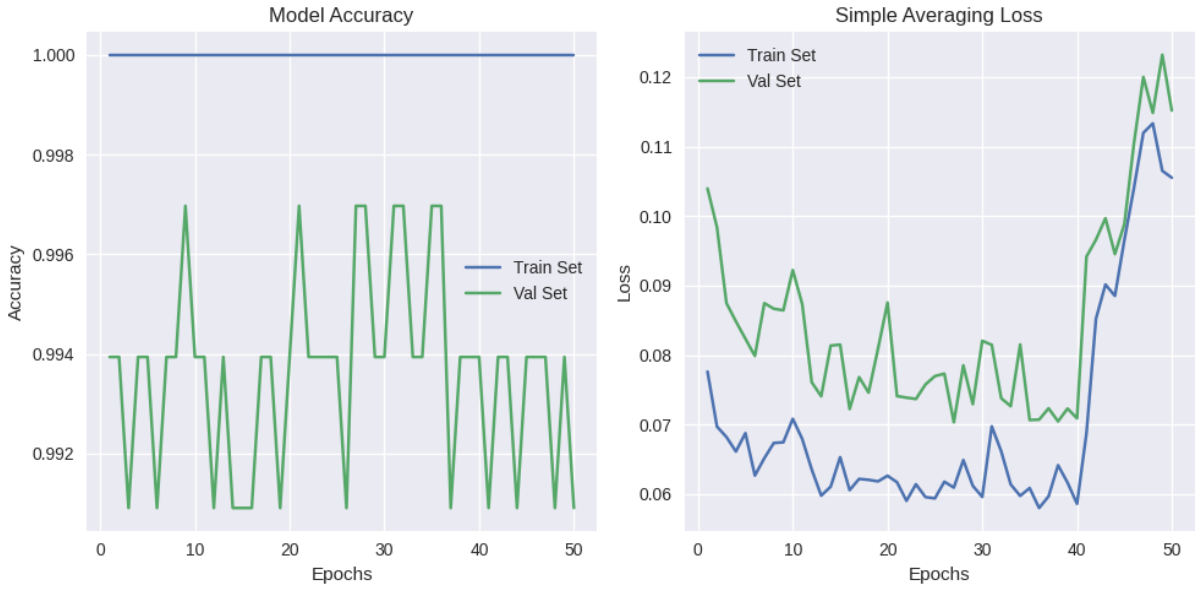

Figure 3.2: Accuracy and Loss using Simple Average

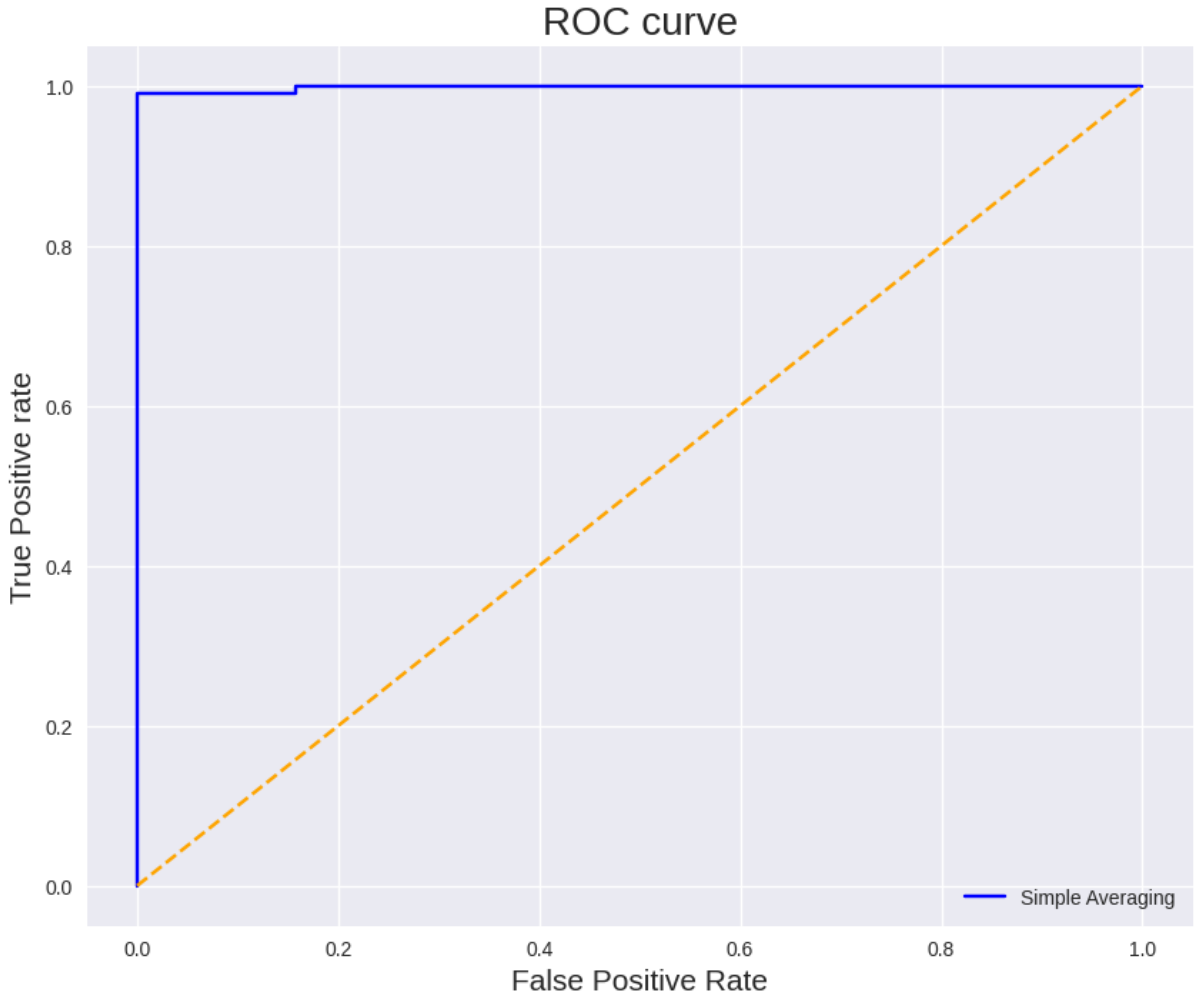

Figure 3.3: ROC Curve using Simple Average

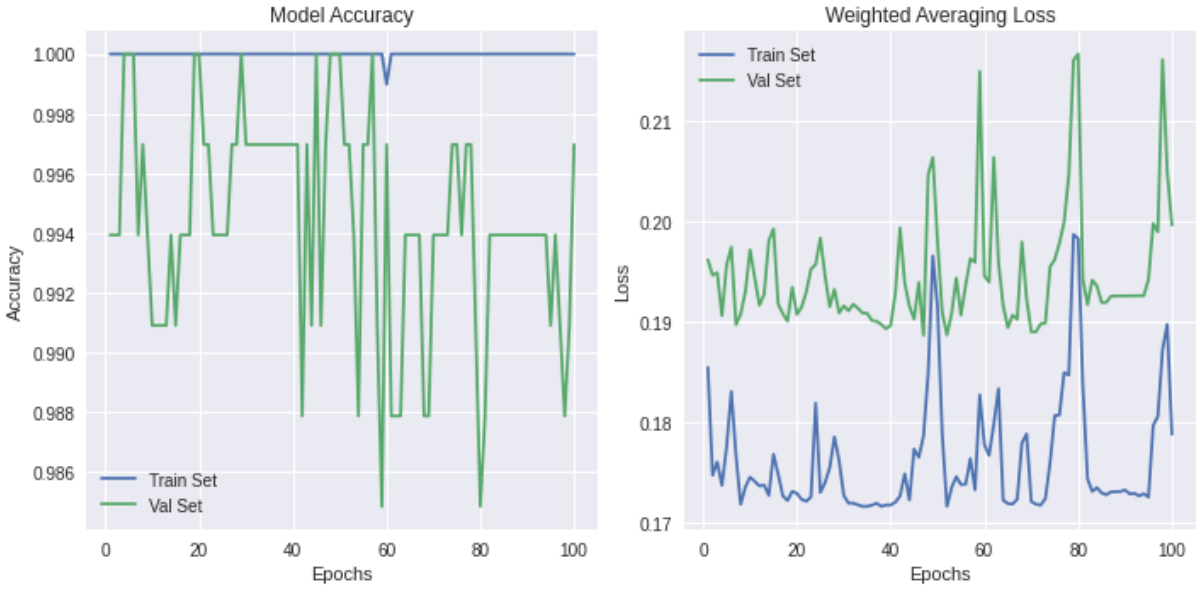

Figure 3.4: Accuracy and Loss using Weighted Average

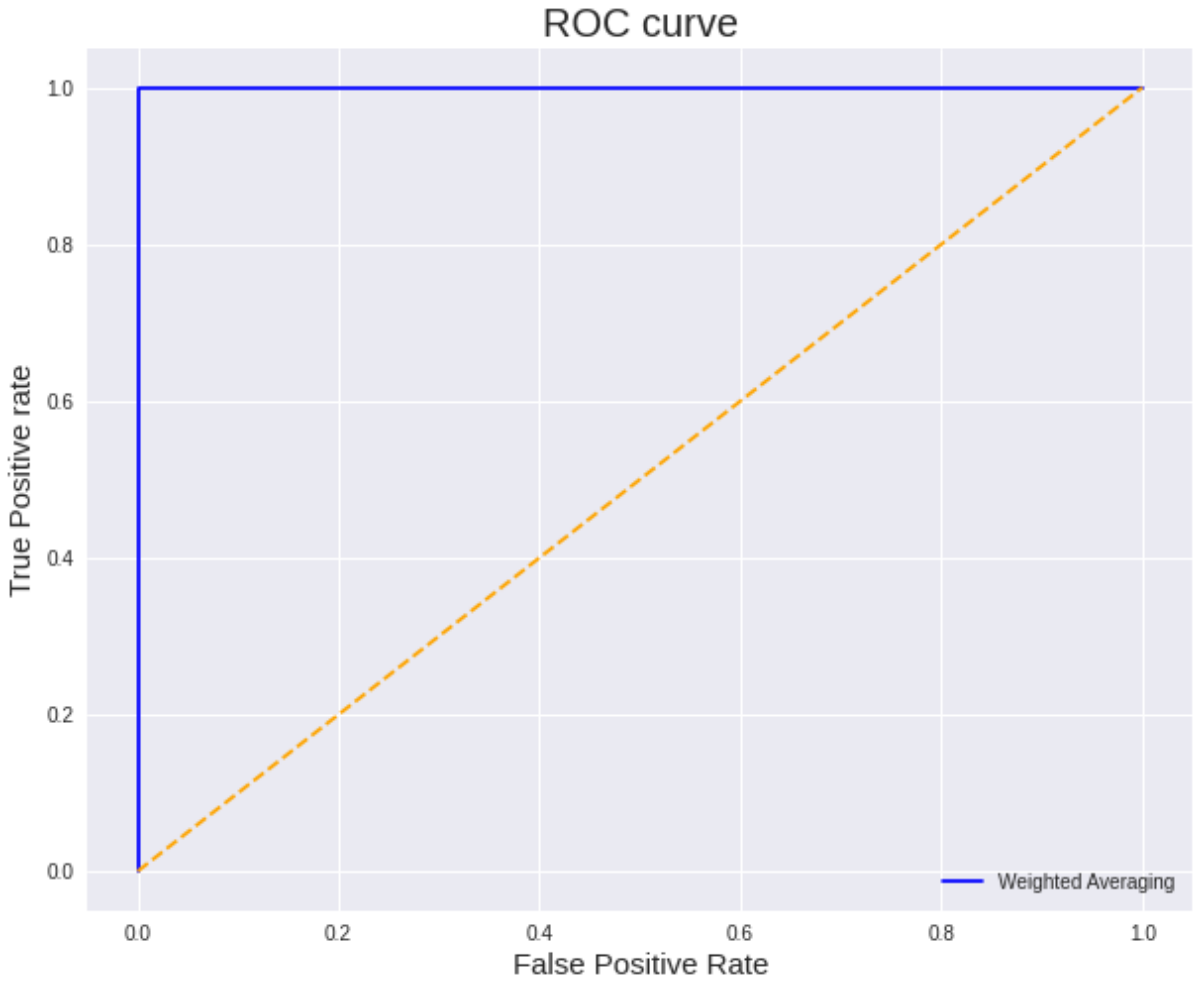

Figure 3.5: ROC Curve using Weighted Average

Chapter 4

Results

The cropping technique significantly impacted the performance of CNNs when applied to MRI images from the OASIS dataset. Cropping, which involves selecting specific regions of interest (ROIs), helped focus the model's attention on areas such as the hippocampus or cortical regions affected by Alzheimer's disease. Simple center cropping resulted in a loss of valuable peripheral information, while more sophisticated cropping methods, such as random cropping or pre-processing segmentation, improved model performance by retaining essential context. The highest accuracy of 98.2% was achieved by VGG19 following cropping.

Ensemble learning strategies were implemented to combine the predictions from VGG19, DenseNet201, and EfficientNetV2S. Simple averaging of DenseNet201 and EfficientNetV2S yielded the highest accuracy of 98.9%. However, the ensemble of all three models with weighted averaging resulted in the lowest accuracy of 94.1%, highlighting the negative impact of inappropriate weighting. The ensemble of VGG19 and DenseNet201 achieved the highest AUC score, demonstrating superior discriminatory power, while the ensemble of all models with weighted averaging showed the lowest AUC, further emphasizing the detrimental effect of improper weighting.

Preprocessing and data augmentation techniques were also applied to optimize the performance of CNNs. Normalization adjusted pixel intensities to the range [0, 1], and denoising techniques were used to reduce noise in MRI images. Resizing the images to 224x224 pixels

ensured compatibility with the CNN input dimensions. Data augmentation, including random rotations, flips, zooms, and elastic deformations, improved model robustness and prevented overfitting.

Conclusion and Future Work

4.1 Conclusion

In this study, we proposed an ensemble-based transfer learning framework aimed at enhancing the accuracy and robustness of Alzheimer's disease classification from MRI images. By combining simple and weighted averaging ensemble techniques, our approach effectively leveraged the strengths of multiple models, allowing for improved feature extraction and better generalization to unseen data. The application of data augmentation strategies further ensured that the model could handle the inherent variability of MRI images, contributing to its robustness.

Our results demonstrated that both ensemble methods performed comparably well, highlighting the adaptability of ensemble learning techniques for medical image analysis. Additionally, the evaluation of both uncropped and cropped MRI images emphasized the versatility of our framework, enabling it to adapt to different input configurations.

Looking to the future, we plan to expand our framework by incorporating more transfer learning models, exploring advanced ensembling techniques such as stacking and boosting, and evaluating its performance across diverse Alzheimer's disease databases. We also aim to integrate adversarial networks for generating synthetic MRI images, which could further augment our training data and improve the model's ability to detect early-stage Alzheimer's disease.

In conclusion, our ensemble-based transfer learning framework demonstrates significant potential for advancing Alzheimer's disease diagnosis. With continued refinement, it holds the promise of offering more accurate, reliable, and early detection capabilities, ultimately benefiting clinical prac-

tices and patient outcomes.

4.2 Future Directions

Future research directions in evaluating CNNs for Alzheimer's diagnosis using the OASIS dataset include exploring hybrid models combining CNNs with recurrent neural networks (RNNs) or attention mechanisms. Integrating multi-modal data, such as PET scans, genetic information, and clinical data, could enhance the diagnostic capabilities of models. Developing Explainable AI (XAI) tools would improve the interpretability of model decisions, which is crucial in clinical settings. Implementing automated ROI detection could further enhance the cropping process and focus model attention on critical areas. Additionally, leveraging pre-trained models from medical datasets like the UK Biobank could improve performance, while optimizing models for real-time processing on edge devices would support faster diagnoses. Standardizing preprocessing protocols for MRI imaging could ensure consistency and reproducibility across studies.

Bibliography

[1] Mingxing Tan and Quoc V. Le, *EfficientNet: Rethinking model scaling for convolutional neural networks*, Proceedings of the 36th International Conference on Machine Learning, 97, 6105-6114, 2019. Available at: https://arxiv.org/abs/1905.11946.

[2] François Chollet, *Xception: Deep learning with depthwise separable convolutions*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1251-1258, 2017. Available at: https://arxiv.org/abs/1610.02357.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, *Deep residual learning for image recognition*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770-778, 2016. Available at: https://arxiv.org/abs/1512.03385.

[4] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathan Shlens, and Zbigniew Wojna, *Rethinking the inception architecture for computer vision*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2818-2826, 2016. Available at: https://arxiv.org/abs/1512.00567.

[5] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger, *Densely connected convolutional networks*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4700-4708, 2017. Available at: https://arxiv.org/abs/1608.06993.

[6] Md. Simul Hasan Talukder, Md. Shawon Rahman, and Shama Naz, *An improved model for diabetic retinopathy detection using transfer learn-*

ing and ensemble learning, arXiv preprint arXiv:2308.05178, 2023. Available at: https://doi.org/10.48550/arXiv.2308.05178.

[7] Timothy Salimans, Andrej Karpathy, and Xi Chen, *A comprehensive review of deep learning in medical image analysis*, IEEE Transactions on Medical Imaging, 35(5), 1227-1242, 2016. DOI: 10.1109/TMI.2016.2521510.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, *Imagenet classification with deep convolutional neural networks*, Proceedings of the 25th International Conference on Neural Information Processing Systems, 1097-1105, 2012. Available at: https://arxiv.org/abs/2012.02722.