

Abhinand V

Abhinand_MajorReportMay20_2026

 Major project

Document Details

Submission ID

trn:oid:::27535:139908661

Submission Date

May 21, 2026, 9:31 PM GMT+5:30

Download Date

May 21, 2026, 9:34 PM GMT+5:30

File Name

Abhinand_MajorReportMay20_2026.pdf

File Size

2.3 MB

42 Pages

9,549 Words

54,986 Characters

8% Overall Similarity



The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text
- Small Matches (less than 8 words)

Match Groups

-  **66 Not Cited or Quoted 8%**
Matches with neither in-text citation nor quotation marks
-  **0 Missing Quotations 0%**
Matches that are still very similar to source material
-  **0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 7%  Internet sources
- 5%  Publications
- 0%  Submitted works (Student Papers)

Match Groups

- 66 Not Cited or Quoted 8%**
Matches with neither in-text citation nor quotation marks
- 0 Missing Quotations 0%**
Matches that are still very similar to source material
- 0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 7% Internet sources
- 5% Publications
- 0% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	dspace.dtu.ac.in:8080	1%
2	Internet	www.mdpi.com	<1%
3	Internet	link.springer.com	<1%
4	Publication	Luo, Ruikang. "Exploration of 3D CNN Based Deep Learning Models for Real-Time ...	<1%
5	Internet	opendata.uni-halle.de	<1%
6	Internet	epub.fh-joanneum.at	<1%
7	Internet	repository.essex.ac.uk	<1%
8	Internet	adni.loni.usc.edu	<1%
9	Internet	dokumen.pub	<1%
10	Internet	www.coursehero.com	<1%

11	Publication	Gharib, Waleed Sayed Ahmed Fathy. "Interpretable Predictive Modeling for Patie...	<1%
12	Internet	liu.diva-portal.org	<1%
13	Internet	www.cg.tuwien.ac.at	<1%
14	Publication	Samuel Ocen, Lawrence Muchemi, Michaelina Almaz Yohannis. "Optimized CNN E...	<1%
15	Internet	academicrepository.khas.edu.tr	<1%
16	Internet	files.core.ac.uk	<1%
17	Internet	sites.duke.edu	<1%
18	Internet	www.ctad.fr	<1%
19	Internet	public-pages-files-2025.frontiersin.org	<1%
20	Internet	www.researchgate.net	<1%
21	Publication	Amorim, Manuel João Gomes Alves. "Exploring the Efficiency and Interpretability ...	<1%
22	Internet	mdpi-res.com	<1%
23	Publication	Chen, Qixu. "On End-User Error Handling in Interactive Queries and Tables", Hon...	<1%
24	Publication	Kaushal Kishor Bhatt. "DETECTION OF DEMENTIA VIA HYBRID CONVOLUTIONAL N...	<1%

25	Internet	isjem.com	<1%
26	Publication	Kumar, Dheeraj. "Automated Detection and Classification of Cholelithiasis from U...	<1%
27	Internet	5dok.org	<1%
28	Internet	ofsoundof.github.io	<1%
29	Publication	James E. Maddux, Barbara A. Winstead. "Psychopathology - Foundations for a Con...	<1%
30	Publication	Rikuto Yoshizawa, Kohei Yamamoto, Tomoaki Ohtsuki. "Investigation of Data Lea...	<1%
31	Publication	Surbhi Dhir, Ruchika Bala, Nidhi Goel, Arun Sharma. "Improved Transfer Learning...	<1%
32	Internet	iieta.org	<1%
33	Publication	Abhilasha Sharma, Vishwas Rathi, Anupam Biswas, Anil Singh, Omer Rana. "Multi...	<1%
34	Publication	Ghulfam Hussain. "Prostate Cancer Detection in Bi-parametric MRI Using Deep Le...	<1%
35	Publication	Hu, Xiangyu. "Multi-Perspective Feature Learning for Facial Expression Recogniti...	<1%
36	Publication	Jha, Anshul. "Neural Networks for Intelligent Healthcare: Transfer Learning and ...	<1%
37	Publication	Shalli Rani, Ayush Dogra, Ashu Taneja. "Smart Computing and Communication fo...	<1%
38	Publication	Sumindar Kaur Saini, Vasudha Bansal, Ravinder Kaur, Mamta Juneja. "ColpoNet fo...	<1%

39	Publication	Suneeta Satpathy, Álvaro Rocha, Sachi Nandan Mohanty, Tanupriya Choudhury. "...	<1%
40	Publication	Vivek Sunnapwar, Hariram Chavan, Namrata F. Ansari. "Advances in Science and ...	<1%
41	Internet	bc.pollub.pl	<1%
42	Internet	comsis.org	<1%
43	Internet	dspace.daffodilvarsity.edu.bd:8080	<1%
44	Internet	huyhieupham.github.io	<1%
45	Internet	persimune.github.io	<1%
46	Internet	repository.publisso.de	<1%
47	Internet	research.chalmers.se	<1%
48	Internet	rsisinternational.org	<1%

Leakage-Aware Multi-Slice MRI-Based Deep Learning Framework for Dementia Stage Classification

Thesis Submitted

**in Partial Fulfillment of the Requirements for the
Degree of**

MASTER OF TECHNOLOGY

by

ABHINAND V

24/DSC/15

abhinandv_24dsc15@dtu.ac.in

Under the supervision of

Dr. Divyashikha Sethia

Associate Professor



DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

**Shahbad Daulatpur, Main Bawana Road,
Delhi-110042, India**

MAY, 2026

DECLARATION

I hereby declare that the Report work entitled **Leakage-Aware Multi-Slice MRI-Based Deep Learning Framework for Dementia Stage Classification** which is being submitted to Delhi Technological University, in partial fulfillment of requirements for the award of the degree of Master Of Technology (Software Engineering) is a bonafide report of M.Tech Major Project carried out by me. The material contained in the report has not been submitted to any university or institution for the award of any degree.

Abhinand V
24/DSC/15

CERTIFICATE

This is to certify that Project Report entitled **Leakage-Aware Multi-Slice MRI-Based Deep Learning Framework for Dementia Stage Classification** submitted by Abhinand V (24/DSC/15) for partial fulfillment of the requirement for the award of degree Master Of Technology (DSC604) is a record of the candidate work carried out by him under my supervision.

Dr. Divyashikha Sethia

Project Guide

Associate Professor

Department of Software Engineering

Delhi Technological University

ABSTRACT

Dementia is a cognitive disorder characterized by a persistent reduction in memory, reasoning, and functional skills . Accurate and early diagnosis plays a crucial role in devising a successful treatment program . This report explores a deep learning approach for dementia classification based on multi-class MRI images utilizing the OASIS-2 dataset .

In the initial stage, different CNN models, i.e., VGG16 , ResNet50 , DenseNet121 , and InceptionV3 , are trained on 2D MRI slices. While these models exhibit high accuracy, a careful examination suggests that leakage issues had caused inflated accuracy results due to erroneous slicing techniques that resulted in shared slices between training and validation sets from the same patients .

The current study introduces a leakage-aware evaluation technique based on a data split according to patient level. An additional multi-slice data fusion process is also developed to analyze MRI slices during inference time and extract more anatomical details from brain regions.

In the ultimate model, MRI data are fused with ten clinical variables (MMSE, nWBV, Age, CDR, eTIV, ASF, SES, EDUC, Sex, and MR Delay) using multimodal EfficientNetV2B0 architecture. Class imbalance issues are addressed by reducing dementia categories to three, namely Non-Demented, Mild-Demented, and Demented. The developed model attains a 70.48% accuracy rate, which is a fair indication of the model's generalization capability, unlike the inflated baseline accuracies. The model performs satisfactorily in identifying Non-Demented and Demented individuals; however, it struggles with Mild-Demented diagnoses.

In conclusion, this report illustrates the significance of applying a thorough evaluation technique for clinical deep learning models . Combining leakage-free training and multi-slice multimodal fusion improves dementia classification accuracy.

ACKNOWLEDGEMENT

I would like to express my gratitude and appreciation to all those who supported me in completing this report.

A special thanks to my mentor and project guide, Dr. Divyashikha Sethia, for her invaluable guidance and support throughout this work.

(Abhinand V)

Contents

	1 Introduction	3
	1.1 Background	3
	1.2 Problem Statement	4
	1.3 Proposed Solution	5
	1.4 Objectives	5
	1.5 Thesis Overview	6
	2 Technical Background	7
	2.1 Datasets for Dementia Research	7
	2.1.1 OASIS-2 Dataset	7
	2.1.2 ADNI Dataset	8
	2.1.3 Comparison of OASIS-2 and ADNI	9
	2.2 Deep Learning in Dementia Prediction	9
	2.3 Convolutional Neural Networks (CNNs)	9
	2.4 Transfer Learning	10
	2.5 CNN Architectures	10
	2.5.1 VGG16	10
	2.5.2 ResNet50	10
	2.5.3 DenseNet121	11
	2.5.4 InceptionV3	11
	2.5.5 EfficientNetV2B0	11
	2.6 Optimization and Regularization	12
	2.7 Model Interpretability	12
	2.7.1 Grad-CAM	12
	2.7.2 SHAP Analysis	12
	2.8 Evaluation Metrics	13
	3 Related Works	14
	3.1 Classical Machine Learning Approaches	14
	3.2 Deep Learning and CNN-Based Approaches	15
	3.3 Ensemble and Multimodal Fusion Methods	15
	3.4 Summary and Research Gap	16
	4 Baseline Classification for Dementia	18
	4.1 Background	18
	4.2 Objectives	18
	4.3 Methodology	19
	4.3.1 Data Preprocessing	19
	4.3.2 Model Development	20

4.3.3	Evaluation Strategy	20
4.4	Results and Analysis	20
4.4.1	Model Performance	20
4.4.2	Analysis	21
4.4.3	Discussion	21
4.4.4	Limitations	22
5	Leakage-Aware Multi-Slice Fusion for Dementia Classification	23
5.1	Background	23
5.2	Objectives	23
5.3	Methodology	24
5.3.1	Identifying and Eliminating Data Leakage	24
5.3.2	Multi-Slice Fusion Strategy	25
5.3.3	Fusion Model Architecture	26
5.3.4	Training Strategy	26
5.4	Results and Analysis	26
5.4.1	Model Performance	26
5.4.2	Analysis	27
5.4.3	Grad-CAM Interpretability	28
5.4.4	SHAP Clinical Feature Importance	29
5.4.5	Discussion	30
5.4.6	Limitations	30
6	Conclusion and Future Work	31
6.1	Conclusion	31
6.2	Future Work	31

 15

 13

List of Tables

2.1	Subject distribution in OASIS-2 [1].	7
2.2	Comparison of OASIS-2 and ADNI datasets.	9
2.3	Evaluation Metrics and Their Formulas [2]	13
3.1	Comparison of Related Works Using Various Datasets	16
4.1	Comparison of CNN and ML Model Performance on OASIS-2 MRI Dataset (Slice-Level Split)	21
5.1	Multi-Slice Fusion Model Performance (Subject-Level Split, 3 Classes)	27

List of Figures

2.1	Simplified architecture of VGG16 [3].	10
2.2	Simplified architecture of ResNet50 with residual skip connections [4].	11
2.3	Simplified architecture of DenseNet121 with dense connectivity [5]. .	11
2.4	Simplified architecture of InceptionV3 with parallel multi-scale branches [6].	11
2.5	Simplified architecture of EfficientNetV2B0 with compound scaling [7].	12
4.1	End-to-end flow of the baseline CNN methodology.	19
5.1	End-to-end flow of the leakage-aware multi-slice fusion methodology. .	24
5.2	Per-class accuracy of the multi-slice fusion model.	27
5.3	Grad-CAM [8] visualization for a dementia case.	28
5.4	Grad-CAM [8] visualization for a mild dementia case.	28
5.5	Per-class feature importance using SHAP values [9].	29
5.6	Global feature importance across all classes using SHAP [9].	29

Chapter 1

Introduction

Dementia is a progressive neurocognitive disorder characterized by memory impairment, loss of reasoning, and disturbances in daily functioning [10]. More than 40 million people worldwide are affected, and this number is projected to nearly triple by 2050 [11]. Alzheimer's disease, which accounts for 70–80% of all dementia types, is the leading cause [12]. In addition to aging and genetic vulnerability, other risk factors like high blood pressure, obesity, poor education level, and social isolation are also contributing greatly towards the onset of dementia [13].

One of the most significant instruments used in dementia studies is neuroimaging as it allows for measuring the structural and functional properties of the brain objectively [14]. Among all neuroimaging techniques currently in use, MRI proves to be the most popular for exploring early structural changes in the brain related to the course of dementia [15]. Neuroimaging data collections like OASIS-2 [1] and ADNI [16, 17] include vast MRI datasets and a wealth of clinical information, which can help to build predictive models.

In addition, deep learning techniques, specifically CNNs, have been shown to be remarkably successful in learning spatial features from MRI images [18, 19]. Evaluation of models used in deep learning based on MRI poses a significant problem. Several slice-based techniques end up incorporating data leakage inadvertently because MRI slices from the same patient end up in both training and validation datasets [20]. This causes the results to be overestimated, and the need to solve this issue arises.

1.1 Background

Definition of Dementia Dementia can be defined as a type of syndrome involving a neurocognitive condition whereby an individual experiences a gradual reduction in mental skills such as remembering things, reasoning and doing daily activities [10]. It is a common condition found all around the world which has become a serious health issue facing the 21st century [11, 12]. The largest causative agent of dementia

is **Alzheimer's disease**, contributing 70-80% of all cases of dementia. In addition to genetics and aging, there are many other risk factors that increase the onset and progression rate. These include high blood pressure, obesity, limited education, lack of exercise, and social seclusion.

However, despite its heavy impact globally, dementia continues to be under-diagnosed and under-treated, especially in developing nations. Dementia is currently diagnosed using a process that includes both clinical interviews and **tests, including the Mini-Mental State Examination (MMSE) and the Clinical Dementia Rating (CDR)**. This is not only an extensive and costly procedure but also demands the input of specialists, hence the urgent need for automatic diagnosis aids [14].

Neuroimaging studies have gained momentum in recent times to become a key part of dementia investigation and assessment. Structural MRI imaging is not invasive and enables the detection of alterations in brain structures—such as hippocampal volume reduction, cortex thickness reduction, and increase in ventricular volume—that occur prior to or alongside cognition decline [15]. Neuroimaging databases like OASIS-2 [1] and ADNI [16, 17] offer longitudinal MRI and corresponding clinical information at several stages of the disease, thus allowing for data-driven modeling of the disease course.

Deep learning techniques, specifically CNNs, have helped immensely in analyzing neuroimaging data by discovering sophisticated spatial features automatically using MRI images as inputs without requiring feature design [21, 18]. The main problem, which is frequently ignored, in such cases pertains to model evaluation. Many existing studies perform data splitting at the level of individual MRI slices rather than at the level of subjects, causing slices **from the same patient to appear in both training and test sets**. This data leakage [20] leads to artificially inflated accuracy estimates and models that fail to generalize to truly unseen patients [22]. Addressing this methodological flaw is a central motivation of the present work.

1.2 Problem Statement

Dementia diagnosis is still challenging due to the dependence of diagnosis based on clinical assessments as well as expensive neuroimaging methods [13]. While classical MRI and PET scans provide an exact diagnosis, these approaches involve considerable costs, rendering them unsuitable for large-scale deployment [14]. Additionally, most of the existing machine learning approaches utilize exclusively clinical variables or imaging data independently without their proper combination, which prevents appropriate generalization [23]. Most importantly, existing slice-wise analysis of MRI using deep learning models fails due to data leakage arising from train/test set splitting [20, 22].

1.3 Proposed Solution

The proposed solution is divided into two approaches. The first approach builds a baseline by employing standard CNN transfer learning architectures such as VGG16, ResNet50, DenseNet121, and InceptionV3 on 2D slices of the MRI image dataset using slice-level evaluation [21]. This approach specifically employs the commonly used methodology in prior studies in an effort to showcase its fundamental weakness — namely, data leakage due to slice-level splits [20, 22].

The second approach provides a solution by utilizing a leakage-aware multimodal multi-slice fusion model. The use of subject-level partitioning with *GroupShuffleSplit* guarantees there are no overlaps of subjects' slices between different train-test splits [20]. Six anatomically consistent axial MRI slices are processed with an efficient network backbone, EfficientNetV2B0 [7], the representations of which are aggregated with ten clinical variables in a joint classification head. Class imbalance is mitigated by collapsing four stages of dementia into three: Non-Demented, Mild-Demented, and Demented.

Model interpretability is ensured by Grad-CAM [8] for visualizing spatial attention and SHAP [9] for analyzing the importance of clinical features, confirming that predictions are consistent with known neurobiological markers of dementia [13].

1.4 Objectives

This research involves two complementary paths. First, a baseline is created based on existing CNN algorithms applied to single MRI scans with slice-wise assessment to gain insight into the potential upper bound of performance under a conventional yet misguided method. Second, an innovative approach for a leakage-aware multi-slice fusion is proposed, correcting the existing evaluation protocol.

Baseline Classification for Dementia (Slice-Level)

1. Train VGG16, ResNet50, DenseNet121, and InceptionV3 based on 2D slices generated from OASIS-2 MRI volumes [1] using transfer learning [24].
2. Process 3D NIfTI images into normalized 224×224 annotated slices for use as CNN inputs [18].
3. Assess the performance of different models through accuracy, precision, recall, and F1-score [2].
4. Demonstrate that performing data splitting at the slice level results in data leakage [20] and produces biased performance estimates [22].

Leakage-Aware Multi-Slice Fusion for Dementia Classification

1. Apply subject-level splitting using *GroupShuffleSplit* to prevent data leakage [20].
2. Fuse features from six adjacent axial slices (indices 62–72) processed by EfficientNetV2B0 [7] to capture spatially distributed neurodegeneration.
3. Integrate ten clinical variables (MMSE, nWBV, Age, CDR, eTIV, ASF, SES, EDUC, Sex, MR Delay) with MRI features in a unified multimodal architecture [14].
4. Validate model interpretability using Grad-CAM [8] for spatial attention and SHAP [9] for clinical feature importance.

1.5 Thesis Overview

The rest of this thesis is presented as follows. Chapter 2 introduces the technical background covering deep learning fundamentals, CNN architectures, benchmark datasets, and evaluation metrics. Chapter 3 provides a review of related work involving machine learning and deep learning methods for the classification of dementia. The results of training traditional CNNs using the slice-level split technique, which demonstrates data leakage, are described in Chapter 4. Chapter 5 describes the proposed leakage-aware multi-slice fusion architecture, in which the multimodal model is built employing EfficientNetV2B0 along with ten clinical variables. The findings of both studies are discussed in Chapter 6, which also provides conclusions about the significance of the evaluation protocol and future work directions.

Chapter 2

Technical Background

This chapter discusses the technical concepts behind the framework proposed in this thesis. Specifically, this chapter talks about neuroimaging data in dementia, basics of deep learning, CNNs, optimization, and evaluation. Knowledge of these technical aspects is important to comprehend the methodology used in Chapters 4 and 5.

2.1 Datasets for Dementia Research

There are two main resources for dementia neuroimaging that are often used for dementia study: OASIS-2 and ADNI. Both resources offer data of patients having structural brain MRI and comprehensive clinical information.

2.1.1 OASIS-2 Dataset

Open Access Series of Imaging Studies, Second Edition (OASIS-2) [1] is a longitudinal dataset based on T1-weighted MRI scans of 150 subjects, aged between 60 and 96 years, with repeated visits for up to five occasions. Being longitudinal in nature, such a dataset allows one to monitor changes in brain structure throughout different stages of development of dementia. Apart from MRI scans, the dataset includes some other information, namely CDR, MMSE, EDUC, SES, gender, eTIV, nWBV, Atlas Scaling Factor (ASF), and MR delay between the visits.

Class	Number of Subjects
Non-Demented	72
Demented	64
Converted	14

Table 2.1: Subject distribution in OASIS-2 [1].

Subjects that comprise the non-demented class are those without evident cognitive deficiency when scanned. Subjects in the demented class are clinically diagnosed

as having dementia. Those in the converted class are subjects whose classification starts from the non-demented to demented when revisited in their subsequent sessions [1]. The class labels are based on CDR scores and categorized according to four levels of dementia severity in the baseline studies (Non-Demented, Mild-Demented, Moderate-Demented, and Highly-Demented) then reduced to three categories in the fusion study due to class imbalance problem.

2.1.2 ADNI Dataset

Alzheimer's Disease Neuroimaging Initiative (ADNI) [16, 17] is one of the biggest longitudinal neuroimaging studies conducted at multiple sites across the globe. Founded in 2004 and being sponsored by National Institute on Aging, National Institute of Biomedical Imaging and Bioengineering, Food and Drug Administration, several drug manufacturing companies, as well as non-profit organizations, the initiative has recruited over 2,000 volunteers into four stages of the project including ADNI-1, ADNI-GO, ADNI-2, and ADNI-3.

Among the types of data collected by ADNI are the following:

- *Tissue morphometry images* obtained through structural magnetic resonance imaging (T1-weighted, T2-weighted, and FLAIR).
- *Functional imaging* using fMRI, FDG PET, and amyloid PET.
- *CSF biomarkers*, such as amyloid- β and tau proteins.
- *Genomic markers* like APOE genotyping.
- *Neuropsychology data* (MMSE, ADAS-cog, CDR, and RAVLT).
- *Demographic data* like age, gender, education in years and ethnicity.

Classification of ADNI subjects into four diagnostic groups occurs, which include CN (cognitively normal), SMC (significant memory concern), MCI (mild cognitive impairment), and AD (Alzheimer's disease). There is further subdivision into EMCI and LMCI in MCI. These subdivisions are made during the later stages of the ADNI study so that the progression towards AD can be studied in detail.

However, while the multisite nature of ADNI, with its hundreds of clinical centers that use uniform imaging modalities, may add generalizability compared to other datasets such as OASIS-2, which are limited to a single site, it poses a greater challenge to working with the dataset [23]. Harmonization is necessary due to inter-site differences in scanner equipment, protocol, and even image quality. Despite these issues, ADNI's large size, variety in diagnostics, and longitudinal data tracking have made it the standard dataset used for machine learning algorithms for computational Alzheimer's studies. The algorithms created using OASIS-2 are generally extensible to ADNI.

2.1.3 Comparison of OASIS-2 and ADNI

Table 2.2: Comparison of OASIS-2 and ADNI datasets.

Property	OASIS-2	ADNI
Number of subjects	150	2,000+
Imaging modalities	Structural MRI (T1)	MRI, fMRI, PET, CSF
Genetic data	No	Yes (APOE genotype)
Number of sites	1	60+
Diagnostic categories	4 (CDR-based)	4-6 (CN, SMC, EMCI, LMCI, MCI, AD)
Longitudinal	Yes	Yes
Publicly available	Yes (free)	Yes (application required)
Primary use	Dementia staging, MRI classification	AD detection, MCI conversion prediction

2.2 Deep Learning in Dementia Prediction

The process of feature extraction does not need to be done by human intervention; it can be accomplished through deep learning algorithms that automatically enable the model to capture features in imaging datasets [19, 18]. In the study of dementia, the first indication of neurostructural abnormalities may manifest as changes in tissue intensity, thickness, or morphology [13]. The application of CNNs to MRI images allows for recognition of specific patterns of atrophy within different regions of the brain, such as the parietal cortex, temporal lobe, and hippocampus [15, 21].

2.3 Convolutional Neural Networks (CNNs)

The CNN model architecture is specifically tailored for image analysis [19]. It extracts low-level features (textures, edges) and combines them progressively towards more complex features relevant for disease categorization using hierarchical architecture [18].

- **Convolutional layers** utilize filter functions to analyze the MRI slice for anatomical texture patterns, intensity differences, and structural borders.
- **Pooling layers** downsize the dimensionality of feature maps, decreasing computational complexity while retaining critical data [19].
- **Fully connected layers** combine learned features from convolutional units and classify them to define the boundaries between dementia classes.
- **Softmax activation layers** map the activations to probabilities for each class of dementia [18].

2.4 Transfer Learning

The majority of medical imaging data sets, such as OASIS-2 [1], are fairly small in comparison with larger-scale data sets utilized in computer vision. Training a CNN from scratch with insufficient data tends to cause overfitting [25]. Transfer learning can solve this problem by taking advantage of the pretrained models trained on ImageNet [24]. The pretrained CNN is capable of extracting features efficiently; therefore, only the last layers have to be fine-tuned for classification of dementia severity stages [25].

2.5 CNN Architectures

The following is an explanation of the five CNN models used in this report, with the figures showing the high-level structure of each model.

2.5.1 VGG16

The VGG16 model [3] has been developed by Simonyan and Zisserman. It utilizes a simple yet effective deep architecture consisting of 13 layers of convolutional operations and 3 layers of fully connected operations, employing filters of size 3×3 with stride = 1 and padding = "same". Since the filters' dimensions remain constant throughout the model, the depth of the model, that is, 16 weight layers, becomes the major determining factor for its representation capabilities. Max-pooling layers are placed after certain layers of convolution operations to reduce their resolution by half. With 138 million parameters, VGG16 becomes the deepest model used in this report.

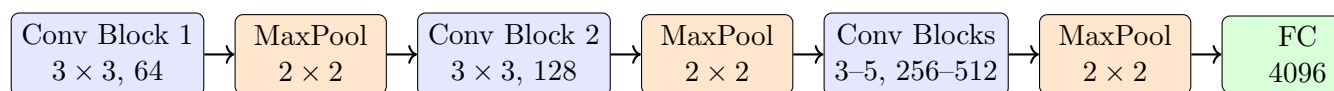


Figure 2.1: Simplified architecture of VGG16 [3].

2.5.2 ResNet50

The ResNet50 model [4] utilized residual (shortcut) connections to mitigate the issue of vanishing gradients that occurred in very deep models. The output in each residual block consists of adding a residual function $\mathcal{F}(\mathbf{x})$ to its input $\mathbf{x}$ such that the output becomes $\mathcal{F}(\mathbf{x}) + \mathbf{x}$. The gradient can thus flow straight through the model without obstruction, facilitating the training of deep models with over 50 layers. ResNet50 employs bottleneck blocks with 1×1 , 3×3 , 1×1 convolutions to achieve greater efficiency, having around 25 million parameters.

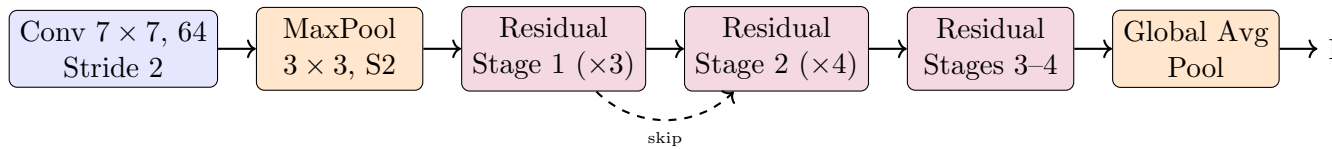


Figure 2.2: Simplified architecture of ResNet50 with residual skip connections [4].

2.5.3 DenseNet121

The concept of skip connections is further expanded in DenseNet121 [5], where each layer is connected to every other layer that comes after it in the same dense block. Specifically, the output of layer ℓ is produced based on the concatenation of feature maps of all previous layers $\{x_0, x_1, \dots, x_{\ell-1}\}$. DenseNet121 consists of four dense blocks, each followed by a transition layer (batch normalization + 1×1 conv + average pooling), with around 8 million parameters.

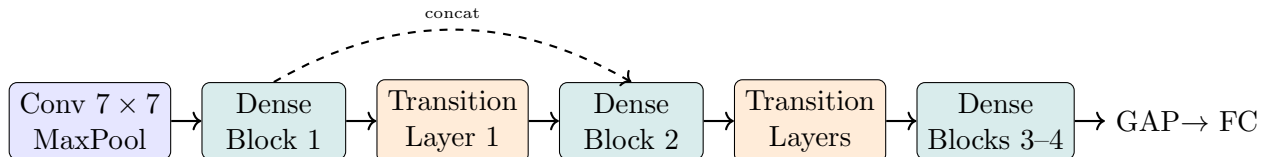


Figure 2.3: Simplified architecture of DenseNet121 with dense connectivity [5].

2.5.4 InceptionV3

InceptionV3 [6] uses parallel multi-scale convolution filters in each Inception layer that allows the model to identify features at different spatial scales in parallel fashion. InceptionV3 model makes use of parallel 1×1 , 3×3 , and 5×5 convolutions. In addition, factorization techniques, where a 5×5 filter is replaced by two stacks of 3×3 filters and asymmetric filter technique, where an $n \times n$ filter is replaced with a $1 \times n$ followed by an $n \times 1$ filter, are used to make computation cheaper.

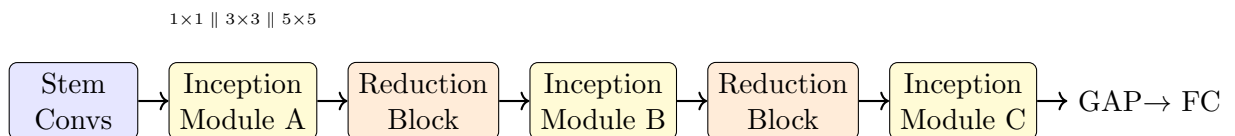


Figure 2.4: Simplified architecture of InceptionV3 with parallel multi-scale branches [6].

2.5.5 EfficientNetV2B0

EfficientNetV2B0 [7] adopts a technique called compound scaling that can jointly adjust network depth, width, and input image size with just one compound factor ϕ . In contrast to adjusting the dimensions separately, compound scaling guarantees

that all three dimensions scale together, resulting in a substantial improvement in accuracy versus efficiency. EfficientNetV2B0 attains a greater accuracy rate on ImageNet data than VGG16, but requires only 7.2 million model parameters while VGG16 uses 138 million parameters, thereby rendering it more appropriate for the OASIS-2 data set.

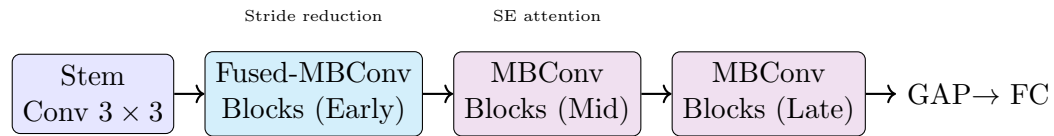


Figure 2.5: Simplified architecture of EfficientNetV2B0 with compound scaling [7].

2.6 Optimization and Regularization

The training procedure utilizes the Adam optimizer [26] with a learning rate of 0.001. The sparse categorical cross-entropy function is used for the baseline models. For the fusion model, the categorical cross-entropy function with a label smoothing factor of 0.1 [27] is employed to avoid overly confident predictions. The following regularization techniques are utilized: data augmentation with rotation, zooming, and brightness alterations [28], dropout layers [29], and early stopping.

2.7 Model Interpretability

2.7.1 Grad-CAM

The Gradient-Weighted Class Activation Mapping (Grad-CAM) approach [8] involves a method that creates class-discriminative heatmaps from any convolutions neural network, without needing any architecture change or retraining of the network. The process entails calculating the gradient of the class score with regard to the feature maps in the last convolution layer, which are used to generate a heatmap with the aid of the gradients. The resultant heatmap is overlaid on the input MRI slice to provide information on the parts that have contributed most in generating the prediction of the network.

2.7.2 SHAP Analysis

SHAP (SHapley Additive exPlanations) [9] offers an approach for calculating the importance of a feature based on game theory. The SHAP value is defined as the average marginal impact that the feature has when added to any subset of features. Within the analysis of clinical features, the SHAP values determine how much the

ten features associated with dementia (MMSE, nWBV, Age, CDR, eTIV, ASF, SES, EDUC, Sex, and MR Delay) impact the model's output for each class of dementia.

2.8 Evaluation Metrics

These metrics are derived using the confusion matrix to evaluate classifier performance [2]. In this case, TP is defined as true positive, TN as true negative, FP as false positive, and FN as false negative.

Table 2.3: Evaluation Metrics and Their Formulas [2]

Metric	Formula	Description
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	Percentage of all samples classified correctly according to their actual class.
Precision	$\frac{TP}{TP + FP}$	Out of all samples predicted to be demented, what percentage is actually demented?
Recall	$\frac{TP}{TP + FN}$	Out of all samples which are actually demented, what percentage is detected by the model?
F1-Score	$\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$	A metric that balances precision and recall, especially helpful when classes are imbalanced [2].

When combined, these metrics provide a comprehensive analysis of how the model performs, particularly taking into consideration issues related to class imbalance within the OASIS-2 data set [1]. The F1-Score stands out among these metrics since it penalizes models that perform well in the majority class but poorly in minority classes.

Chapter 3

Related Works

This chapter presents a brief overview of literature concerning the application of deep learning and machine learning algorithms for predicting and classifying dementia. Particular importance is given to the methodology adopted, datasets used, and important conclusions reached, with specific focus on the OASIS and ADNI databases. The related studies have been classified under three categories: classical machine learning, deep learning and CNN, and ensemble/multimodal fusion techniques.

3.1 Classical Machine Learning Approaches

The early efforts in computational classification of dementia were mainly based on manual engineering of features from clinical records, which were then used in conventional machine learning models to prove that structured information about patients' demographics and cognition can effectively predict their dementia severity.

A machine learning framework for dementia prediction based on the OASIS dataset was developed by Dhakal et al. [12]. Machine learning algorithms such as AdaBoost, Extra Trees, Gradient Boosting, Logistic Regression, and Support Vector Machine (SVM) were considered in the paper. The best results in terms of accuracy (96.77%) were obtained with SVM when all features were utilized, while chi-square and LASSO feature selection techniques improved the feature set. However, the high figures do not make up for the absence of any subject-level split in the study.

In another study based on the ADNI dataset [16, 17], modifiable risk factors such as education level, hypertension, and lack of physical activity were used as features with Naïve Bayes, Random Forest, Decision Tree, and Logistic Regression. The binary classification between healthy and abnormal subjects produced an accuracy rate of up to 92% [30]. Although the application of modifiable risk factors is relevant from a clinical perspective, the lack of unstructured data (neuroimaging data) makes the methodology less practical for early diagnosis.

The predictive capabilities of OASIS-2 longitudinal MRI data were investigated by Alshamlan et al. [31] with SVM, Random Forest, and Logistic Regression al-

gorithms. Feature selection was performed using mutual information, resulting in high accuracies (over 90%) and emphasizing the significance of longitudinal MRI metadata in classification tasks.

In a study by Shastry et al. [32], a hybrid ensemble learning algorithm involving Random Forest and Logistic Regression techniques was utilized.

3.2 Deep Learning and CNN-Based Approaches

The introduction of deep learning techniques has made major strides in the area of dementia classification through the capability to automatically extract features from MRI scans without any need for handcrafting them.

Saratxaga et al. [15] introduced both 2D and 3D CNN classifiers, named BrainNet2D and BrainNet3D respectively, for detecting Alzheimer's disease from OASIS-1 and OASIS-2 MRIs. With transfer learning using pre-trained CNNs, accuracies exceeding 85% were achieved. It was observed that, in small training samples, the performance of the 2D convolution models was better than that of 3D convolution models. This may be because more parameters are needed for 3D convolution networks, which might not be necessary for small sample sizes.

VGG16 [3], ResNet50 [4], DenseNet121 [5], and InceptionV3 [6] were employed by Kubi et al. [21] on the OASIS-2 dataset of MRI images to predict the severity stages of dementia, achieving up to 90% accuracy in the process. Skull stripping and intensity normalization represented preprocessing techniques. Nonetheless, this study made use of a train-test splitting scheme based on slices, resulting in data leakage as shown in this report.

Javeed et al. conducted a systematic review [23], which included 75 articles from 2011 to 2022, including both ADNI and OASIS datasets. CNN-based methods, ensemble approaches, and SVMs were found to be highly effective. The most significant issues mentioned include dataset variability, absence of standardization in testing, and lack of model transparency, which are all addressed in this report.

3.3 Ensemble and Multimodal Fusion Methods

There have been various researches that incorporate imaging data along with clinical information or have used ensemble techniques to increase the robustness of the results by getting the best-reported values in the literature.

For instance, Periya Nayaki et al. [33] have implemented various machine learning models such as Voting Classifier and XGBoost-Random Forest (XGBRF) on OASIS data using oversampling via SMOTE to deal with imbalanced classes. The model Voting Classifier had attained an accuracy of 98.17% and an F1-score of 97.60%,

which is the best reported in this study. Nevertheless, these results are based only on clinical attributes without any MRI representation.

The convergence of multiple imaging techniques and clinical streams has been especially successful in the context of ADNI research, which includes PET, fMRI, CSF markers, and genetics in addition to structural MRI scans. Several studies conducted using ADNI data have proven that combining amyloid-PET and tau-PET imaging with neuropsychological test results considerably enhances the predictive performance for MCI-to-AD progression beyond imaging data alone [23, 14]. The multimodal systems used in these studies provide inspiration for the clinical features integration technique used in the proposed fusion architecture described in Chapter 5.

It has become clear from the research performed by Yagis et al. [20] that the use of proper evaluation procedures is key to obtaining accurate results. They formally proved that slicing-level cross-validation leads to substantial data leakage in MRI image classification. Experiments conducted by Yagis et al. [20] demonstrate that the models trained and evaluated using slicing-level cross-validation show up to 10–20 percent higher accuracy than their actual generalization capability.

Table 3.1: Comparison of Related Works Using Various Datasets

Authors (Year)	Dataset	Approach	Models Used	Accuracy (%)	F1-Score (%)
Dhakal et al. (2023)[12]	OASIS	Feature selection + ML	SVM, RF, LR	96.77	93.70
Bin-Hezam & Ward (2019)[30]	ADNI	Modifiable risk factors	LR, RF, Naïve Bayes	92.00	–
Javeed et al. (2023)[23]	ADNI, OASIS	Systematic review	CNNs, RF, SVM, ANN	Varied	–
Shastri et al. (2023)[32]	Public	Hybrid ensemble	Logistic RF Boosting	Improved	–
Saratxaga et al. (2021)[15]	OASIS-1,2	2D/3D CNNs + TL	BrainNet2D, BrainNet3D	85.00	–
Kubi et al. (2025)[21]	OASIS-2	CNN dementia staging	VGG16, ResNet50, DenseNet121, InceptionV3	90.00	87.00
Periya Nayaki et al. (2024)[33]	OASIS	Ensemble ML + SMOTE	Voting Classifier, XGBRF	98.17	97.60
Alshamlan et al. (2024)[31]	OASIS-2	Temporal MRI features	SVM, RF, LR	90.00	–
Yagis et al. (2021)[20]	OASIS, ADNI	Leakage analysis	CNNs (slice-level eval)	Inflated	–

3.4 Summary and Research Gap

There are still some limitations found in the current literature review: Most deep learning models with very high accuracy on the OASIS-2 or other small-scale data

sets did not use subject-level splitting for data evaluation, which can make the model susceptible to data leakage. In addition, most studies only used imaging data or clinical data to train the model but did not combine the two types of data. This report attempts to address both issues with multimodal and multislice fusion.

Chapter 4

Baseline Classification for Dementia

In this chapter, the training of the classical CNNs on the 2D MRI slices is achieved by slice-level data splitting. This is explained in Section 4.3, and experimental results are shown in Section 4.4. The approach is almost equivalent to those employed in other recent literature works [21, 15]; but a specific issue with the experimental evaluation process becomes evident, and this motivates the advanced method introduced in Chapter 5.

4.1 Background

It has become common practice to employ slice classification in CNN models, which have been trained using transfer learning based on the ImageNet [24] database. The efficiency with which the features from the image can be learned using deep neural network structures such as VGG16, ResNet50, DenseNet121, and InceptionV3 after they are trained is quite remarkable [21, 15]. The commonly used database for storage of longitudinal MR images of dementia patients is the OASIS-2 [1] database. One problem, however, that tends to get overlooked while employing this process is data leakage [20, 22].

4.2 Objectives

1. *Establishing the Baseline of Performance Using Slice-Based Data:* Training and validating four different CNN models (VGG16, ResNet50, DenseNet121, and InceptionV3) on the OASIS-2 axial slice-level images using pretrained weights of ImageNet is done according to the protocol used by current researchers working with slice-based data [24].
2. *Evaluating and Comparing the Four Different Models:* Evaluations of the per-

formances of the four CNN models will be done according to accuracy, precision, recall, and F1-scores [2], considering the train-validation-test splits made at the slice level, hence establishing the performance baseline for four stages of dementia.

3. *Identify and Demonstrate Data Leakage:* Carry out a study of the process of slice-based splitting in order to prove that an individual can end up being a part of both the test and training sets.

4.3 Methodology

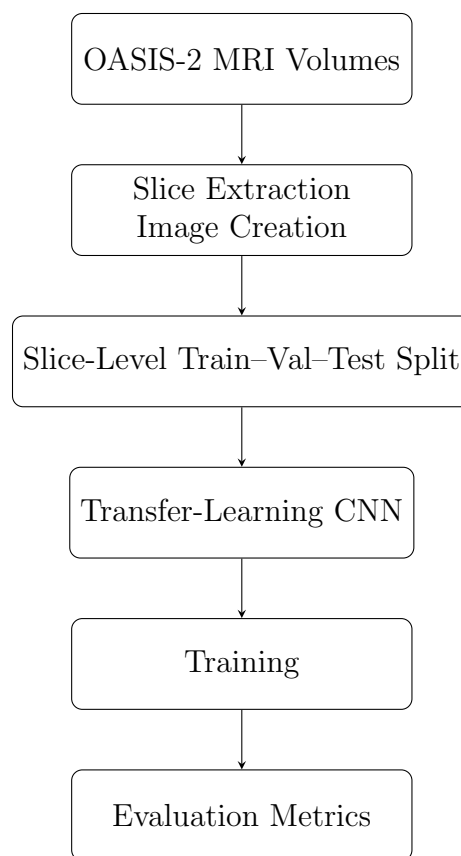


Figure 4.1: End-to-end flow of the baseline CNN methodology.

4.3.1 Data Preprocessing

The NIfTI files are received as complete 3D scans and must be converted to 2D slices before being input into 2D CNNs. The preprocessing pipeline consists of loading each scan, aligning it spatially, and extracting axial slices using fixed depth indices to ensure consistent coverage across patients and exclude variations due to head position or orientation [18].

Images are normalized and resized to 224×224 pixels to match the input requirements of the selected CNN architectures [3]. The splitting of images into training, validation, and test sets is performed at the slice level — meaning each 2D image is treated as an independent sample regardless of its source subject. This introduces data leakage, as discussed in Section 5.3 [20].

4.3.2 Model Development

The transfer learning technique is used to initialize the four CNN architectures with weights pre-trained on the ImageNet dataset [24]. The convolutional layers' parameters stay frozen, while the newly attached classifier is trained using OASIS-2 [25]. This classifier involves three layers: Global Average Pooling, a dense transform layer, and a softmax classifier for the four dementia classes.

For training purposes, a sparse categorical cross-entropy loss function along with the Adam optimizer having a constant learning rate equal to 0.001 is utilized. Geometric and intensity augmentations are applied to increase the variety of training data. Early stopping technique is applied to avoid overfitting of the model.

4.3.3 Evaluation Strategy

Accuracy, precision, recall, and F1-Score are among the measures used for the performance analysis that have been computed using the confusion matrix [2]. Given that evaluation is done at the slice level, the same leakage is evident in the results.

4.4 Results and Analysis

4.4.1 Model Performance

Table 4.1 presents the classification accuracy and corresponding metrics for the four CNN models on four dementia severity classes, alongside comparable results from the literature.

Table 4.1: Comparison of CNN and ML Model Performance on OASIS-2 MRI Dataset (Slice-Level Split)

Study / Our Work	Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
This Work	VGG16[3]	85.00	85.00	85.00	84.00
This Work	ResNet50[4]	78.00	79.00	78.00	84.00
This Work	DenseNet121[5]	56.00	46.00	56.00	42.00
This Work	InceptionV3[6]	62.00	53.00	62.00	54.00
Kubi et al. (2025)[21]	DenseNet121	90.00	85.00	60.00	87.00
Kubi et al. (2025)[21]	VGG16	86.00	79.00	79.00	70.00
Kubi et al. (2025)[21]	ResNet50	65.00	82.00	72.00	78.00
Kubi et al. (2025)[21]	InceptionV3	57.00	52.00	52.00	53.00
Dhakal et al. (2023)[12]	SVM	96.77	100.00	87.93	93.70
Periya et al. (2024)[33]	Voting Classifier	98.17	98.30	97.00	97.60

4.4.2 Analysis

- *VGG16* — **Best Performing Baseline Model:** The best performance baseline model is VGG16, which produced the highest baseline accuracy rate of 85%. This is expected because it is well-known for its strong feature extraction ability and consistent training on image data [21].
- *ResNet50* — *Average Performance Model:* The model with decent performance was ResNet50, which performed with 78% accuracy owing to its use of residual connection, thus allowing for feature reuse in this deep neural network model trained on a smaller dataset.
- *DenseNet121 and InceptionV3:* It is likely that overfitting or lack of feature generalization is responsible for the poor performance of DenseNet121 and InceptionV3 models, with accuracies of 56% and 62%, respectively, due to their use on the small OASIS-2 dataset [25].
- *Comparison to Literature:* The literature also shows that ensemble learning models utilizing clinical data outperformed models based solely on imaging [33, 12], demonstrating the advantages of multimodal information.

4.4.3 Discussion

The use of transfer learning was found to be effective in detecting the required spatial patterns from the MRI slices, saving training time while improving generalization ability over training from scratch [25]. The preprocessing stage successfully converted 3D Nifti volumes into labeled 2D slices for CNNs [18].

On the other hand, the inflated performance metrics can be attributed to data leakage [20]. Splitting slices randomly from individual 2D slices without considering the patient identity meant that slices of the same patients were present in both training and testing datasets. This implies that the model had access to certain anatomical patterns in the training dataset that were also present in the test dataset, thus violating the independence assumption of ML [22].

4.4.4 Limitations

- *Slice Level Data Splitting Leakage:* The primary limitation [20, 22]. The slices belonging to the same patient can occur in both train and test sets, leading to highly optimistic results. This flaw is addressed in Chapter 5.
- *Limited Dataset Size:* Due to the restricted number of patients used in OASIS-2 [1], there could be an increase in the overfitting chance [25].
- *Imbalanced Data:* Because of the insufficient samples of four dementia classes, the multi-class classification problem becomes hard to solve accurately [33].
- *Representation of 2D Slices:* In lieu of using volumes of 3D data, using just 2D slices removes all the spatial context associated with the brain [15].
- *Single-Slice Classifier:* Dependence on a single slice does not allow for the understanding of the distributed multi-region neurodegeneration associated with dementia.

These constraints, particularly data leakage, class imbalance, and single-slice representation, directly inspire the enhanced methodology introduced in Chapter 5.

Chapter 5

Leakage-Aware Multi-Slice Fusion for Dementia Classification

This chapter discusses the main contribution to this thesis work: a leakage-aware, multimodal, multi-slice fusion framework for dementia classification. This method is discussed in Section 5.3, which clearly deals with the three main weaknesses of Chapter 4: data leakage, single slice operation, and the lack of patient information.

5.1 Background

The limitations of slice-based CNNs concerning data leakage, representation of only one slice at a time, and lack of clinical information necessitate an improved evaluation framework [20, 22]. Partitioning data at the level of the subject is important for accurate evaluation of MRI models [20]. Dementia-induced neurodegeneration is a multiregion phenomenon that includes the hippocampus, temporal lobes, and cortex [13], thus making multi-slice representation intrinsically better than single slice representation [14].

5.2 Objectives

1. *Evaluation Without Leakage:* Use subject-wise split by means of *GroupShuffleSplit*, thus assuring that all slices from each individual subject will be placed either in train, validation, or test sets, respectively, and merge features from six axially aligned slices (indices 62, 64, 66, 68, 70, 72), obtained from a single EfficientNetV2B0 [7, 20, 22].
2. *Multimodal Fusion:* Merge the results of brain MRI image recognition with ten clinical features (MMSE, nWBV, Age, CDR, eTIV, ASF, SES, EDUC, Sex, MR Delay) in a single classification layer.

3. *Interpretability and Validation:* Validate spatial attention patterns through Grad-CAM [8] and clinical feature importance through SHAP analysis [9], validating the fact that the model predictions agree with the established neurobiological indicators of dementia [13].

5.3 Methodology

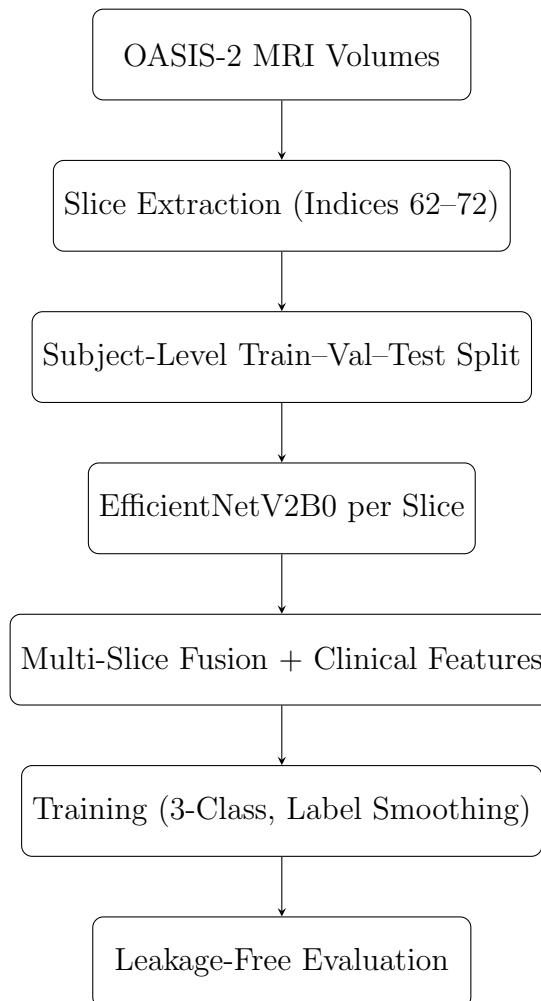


Figure 5.1: End-to-end flow of the leakage-aware multi-slice fusion methodology.

5.3.1 Identifying and Eliminating Data Leakage

1. The Problem: Slice-Level Leakage

With respect to the baseline approach, the slices are viewed as independent samples. However, by nature, slices within the same individual are highly similar. The random splitting of slices leads to:

- *Subject leakage*: Since slices within the same MRI scan are almost anatomically identical, assigning them randomly across the train and test sets results in the model being able to identify the patients themselves, not the patterns of their illnesses [20].
- *Violating the i.i.d. assumption*: The independent nature of the data required for proper machine learning evaluation is violated, leading to biased performance assessment [22].
- *Highly inflated values*: The accuracy, precision, recall, and F1-score are excessively high [2].
- *Failure to generalize*: The model with slice-level leakage does not perform well on individuals that have not been previously seen by the algorithm [22].

2. Subject-Level Partitioning

Every subject in OASIS-2 gets assigned a unique identification number, and all the slices that belong to the same subject will be considered as a whole while splitting. The method *GroupShuffleSplit* is performed using the subject ID as a group, which ensures that there will not be any overlap in subjects between the partitions [20]. This ensures the model learns the structures associated with the disease itself [13].

3. Stratified Distribution

Apart from splitting at the level of subjects, stratification sampling is done to ensure that there is an even balance in the ratio of classes for each partition [33], thus avoiding any bias due to classes.

5.3.2 Multi-Slice Fusion Strategy

1. Motivation and Slice Selection

Neurodegeneration associated with dementia involves multiple brain areas, which include the hippocampus, temporal lobes, and cortex [13, 15]. One slice cannot cover all the structural changes that happen [14]. Six slices, whose indices include 62, 64, 66, 68, 70, and 72, are chosen based on their anatomical importance in identifying neurodegeneration due to dementia [21].

2. Multimodal Clinical Feature Integration

Ten clinical characteristics from the OASIS-2 demographics dataset have been used: Age, EDUC, SES, MMSE, CDR, eTIV, nWBV, ASF, Sex, and MR Delay [1]. These characteristics are then standardized via *StandardScaler*, followed by passing them

through an independent clinical path using Dense(128) and Dense(64), where the output is concatenated with the output of the MRI feature network [14].

5.3.3 Fusion Model Architecture

1. Feature Extraction and Fusion

Each one of the selected six slices is then individually run through a pre-trained EfficientNetV2B0 backbone with frozen convolutional layers [7]. EfficientNetV2B0 is selected over VGG16 since it demonstrates better accuracy on ImageNet with only 7.2 million parameters, unlike VGG16, which has 138 million parameters, greatly minimizing the possibility of overfitting to the smaller OASIS-2 dataset[25]. These slices generate feature vectors that encode structure and intensity features related to the brain.

2. Classification Head

Both the MRI features and the clinical features are passed through the fully connected classifier using dropout. The output layer has the softmax activation function to produce the probability score for each of the three classes. These three classes are Non-Demented, Mild-Demented, and Demented.

5.3.4 Training Strategy

The network is trained using the Adam optimizer [26] with a constant learning rate of 0.001. For training, categorical cross-entropy loss with label smoothing of 0.1 [27] is used. It helps smooth one-hot target distributions. Data augmentation [28], dropout [29], and early stopping methods are used for further regularizing the training process. Subject-independent data splits are used during training such that each subject appears in only one split.

5.4 Results and Analysis

5.4.1 Model Performance

The leakage-aware multi-slice fusion model obtained an overall accuracy of 70.48% in the classification of subjects into three categories of dementia.

Table 5.1: Multi-Slice Fusion Model Performance (Subject-Level Split, 3 Classes)

Class	Precision	Recall	F1-Score	Support
Non-Demented	0.8262	0.8007	0.8132	1104
Mild-Demented	0.5455	0.5570	0.5511	474
Demented	0.7500	1.0000	0.8571	72
Accuracy	0.7048			1650

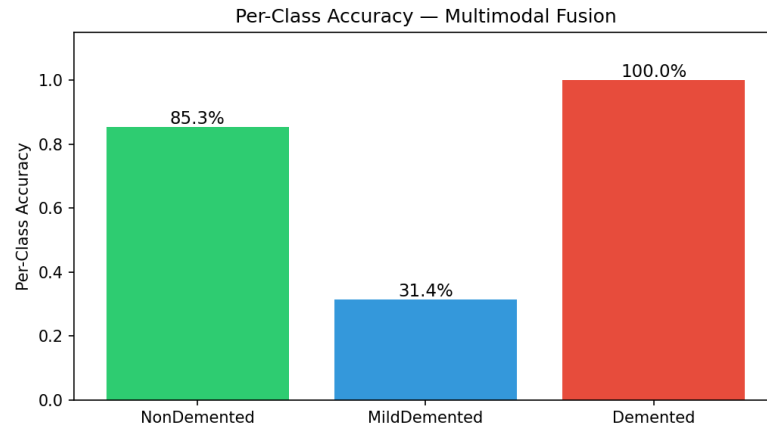


Figure 5.2: Per-class accuracy of the multi-slice fusion model.

5.4.2 Analysis

The performance is satisfactory for the **Non-Demented** class where Precision=0.826 and Recall=0.801 (F1=0.813) and indicates how well the model works with the most common class [2]. The model obtains the best result for the **Demented** class as it perfectly recognizes all instances of this class (Recall=1.000) and provides Precision=0.750 (F1=0.857). However, there are some cases of misclassification caused by data noise. The **Mild-Demented** class is the least recognized, with the lowest F1-score of 0.551, which corresponds to the natural difficulty of detecting early-stage diseases when brain changes are not pronounced and have some similarity to both normal aging and dementia [13]. Lastly, the reduction from 85% accuracy of the baseline VGG16 model (slice-level splitting) to 70.48% (subject-level splitting) proves that the former value was artificially increased due to data leakage [20].

5.4.3 Grad-CAM Interpretability

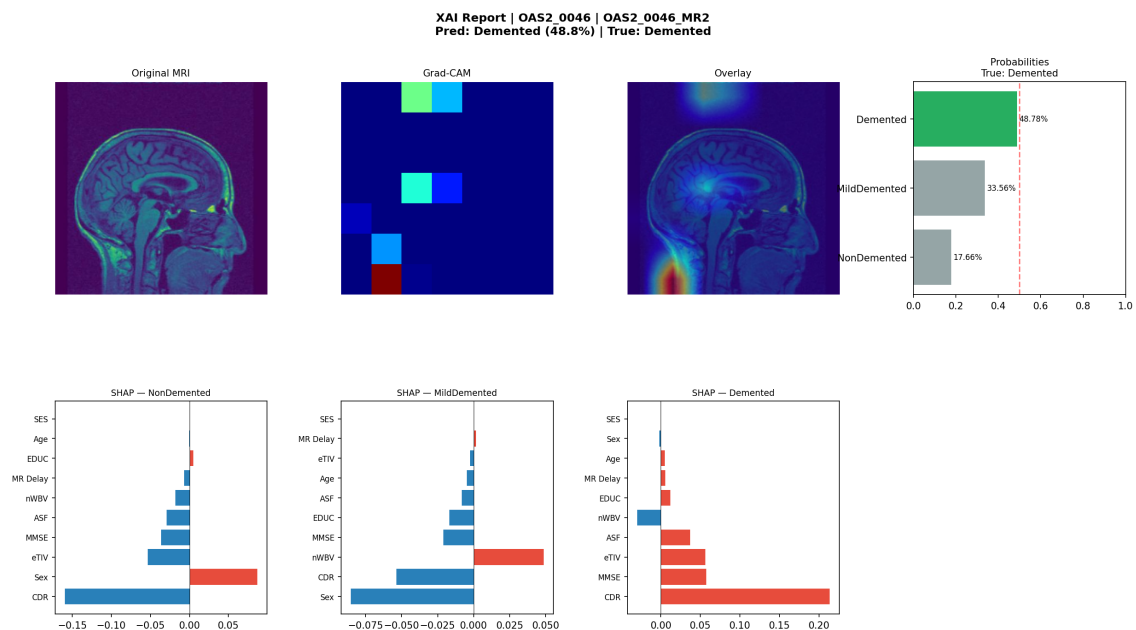


Figure 5.3: Grad-CAM [8] visualization for a dementia case.

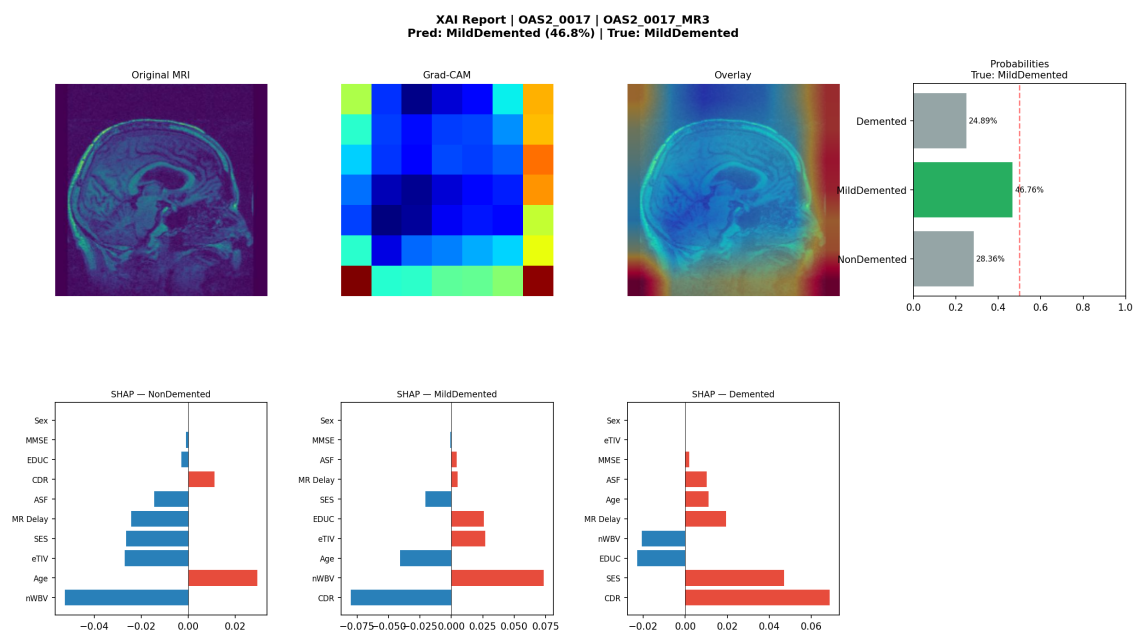


Figure 5.4: Grad-CAM [8] visualization for a mild dementia case.

Grad-CAM analysis highlights anatomically relevant areas associated with dementia-related structural changes, including temporal and parietal regions and areas surrounding the ventricles. These findings are consistent with known patterns of neurodegeneration in Alzheimer's disease [13], supporting the biological validity of the model's learned representations.

5.4.4 SHAP Clinical Feature Importance

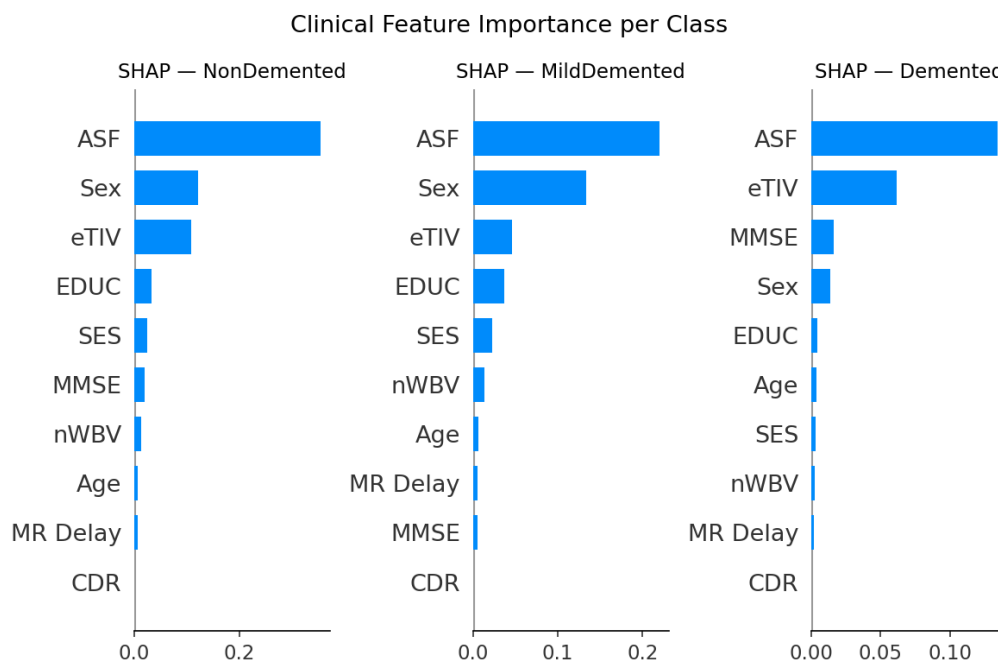


Figure 5.5: Per-class feature importance using SHAP values [9].

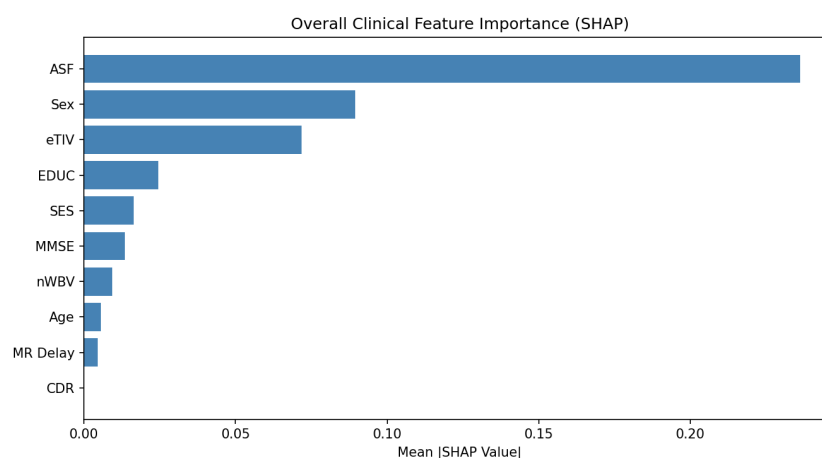


Figure 5.6: Global feature importance across all classes using SHAP [9].

The most critical variable in the model is the **ASF (Atlas Scaling Factor)**, which implies that differences between brain volumes and an atlas image have the strongest link with dementia severity [1]. Another two critical factors in the model are **Sex** and **eTIV**. This indicates that demographic characteristics and the total intracranial volume significantly impact predictions made by the model [12]. It is interesting to note that **MMSE**, which directly assesses cognitive abilities, plays a minor role compared to brain volumes, indicating that the model utilizes more anatomical

markers rather than cognitive performance [9]. Variables like **EDUC**, **SES**, and **MR Delay** have insignificant roles.

5.4.5 Discussion

Two major findings can be derived from the above observations. Firstly, subject-wise splitting is essential for evaluating models reliably for deep learning applications in MRI [20], as failure to do so means that models will learn patient-specific patterns instead of features that discriminate between diseases, yielding overly optimistic performance assessments [22]. Secondly, multi-slice fusion greatly enhances model robustness through recognition of distributed neurodegenerative features across different brain areas [14].

Integration of metadata allows the use of existing biomarkers in the model for inference. The model's dependence on volumetric biomarkers (ASF, eTIV, nWBV), which is confirmed using SHAP, corresponds to an existing clinical perspective on the development of the disease [13].

5.4.6 Limitations

However, there are several drawbacks even after implementing the improvements outlined above. The use of **2D multi-slice rather than 3D volumetric imaging** implies that the exact spatial relationship between the points in the whole brain volume is not taken into account [15]. The small scale of the utilized OASIS-2 database [1] can result in overfitting due to insufficient data for training [25]. Class imbalance still occurs even after collapsing the four groups down to three with the Demented group comprising only 72 test cases. The leakage-aware fusion model along with parallel clinical pathway implies greater computational power usage compared to simpler baselines [14].

Nonetheless, leakage-aware fusion proves to be a valuable improvement in developing clinically relevant classifiers of dementia, which will be addressed further in Chapter 6.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

This study explored two complementary approaches to MRI-based deep learning classification of dementia using the OASIS-2 dataset [1]. In the baseline classification study (Chapter 4), transfer learning models — VGG16, ResNet50, DenseNet121, and InceptionV3 — demonstrated their ability to extract meaningful spatial features from individual 2D MRI slices under a slice-level splitting protocol. Among these, VGG16 showed the most stable results, achieving 85% accuracy. However, as was later found out, split on slices led to leakage of information [20], as images of the same patient were used for both training and testing. As a result, reported measures of performance were overinflated and could not serve as a trustworthy estimate [22].

These issues have been addressed in the leakage-aware multimodal multi-slice fusion model (Chapter 5) with the help of a leakage-free evaluation technique and subject-wise splits together with an image classifier based on EfficientNetV2B0 [7].

To deal with the issue of class imbalance, four stages were reorganized into three classes. Then, a model taking into account 10 clinical factors together with 6 anatomically consistent slices was constructed, resulting in 70.48% accuracy which is worse than the baseline, but is an accurate indicator of generalization [20].

SHAP and Grad-CAM analyses [8, 9] reveal that the predictions of the model are biologically plausible, as brain volumetric parameters, such as ASF, turn out to be the strongest predictors of the disease [13]. This shows once again that leakage-free techniques of evaluation are as crucial as the model architecture in medical image analysis tasks [18].

6.2 Future Work

There are several avenues that could be pursued to expand upon this research. Progressing from the **2D slice-based architecture to 3D volume architecture** by

employing 3D CNNs or vision transformers on the MRI scans will allow the model to detect more complicated spatial relationships within the brain as a whole [15]. The addition of **other imaging modalities** in conjunction with structural MRI and clinical data such as PET scans, functional MRI, and CSF biomarkers may improve the classification performance [14]. More advanced balancing algorithms can include focal loss, weighted cross-entropy, or GAN-based MRI generation techniques to alleviate the issue of imbalanced classes [28]. Using attention mechanisms to replace the concatenation process will allow the model to assign weights to the most significant slices and clinical features per instance. Lastly, testing on **large-scale and more diverse databases** such as ADNI [16, 17] would determine the scalability of the model across multiple demographics, and **clinical trials** would be needed to validate the framework.

Bibliography

- [1] D. S. Marcus *et al.*, “Open Access Series Of Imaging Studies (OASIS-2): Longitudinal MRI Data In Nondemented And Demented Older Adults,” 2010. [Online]. Available: <https://www.oasis-brains.org>
- [2] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.
- [3] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [4] K. He *et al.*, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [5] G. Huang *et al.*, “Densely connected convolutional networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4700–4708.
- [6] C. Szegedy *et al.*, “Rethinking the inception architecture for computer vision,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826.
- [7] M. Tan and Q. Le, “EfficientNetV2: Smaller models and faster training,” in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 139. PMLR, 2021, pp. 10 096–10 106. [Online]. Available: <https://proceedings.mlr.press/v139/tan21a.html>
- [8] R. R. Selvaraju *et al.*, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.
- [9] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30. Curran Associates, Inc., 2017, pp. 4765–4774.

- [10] World Health Organization, "Dementia," *WHO Fact Sheets*, 2023. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/dementia>
- [11] M. Prince *et al.*, "World Alzheimer report 2015: The global impact of dementia," *Alzheimer's Disease International*, 2015.
- [12] S. Dhakal *et al.*, "Dementia Prediction Using Machine Learning," *Procedia Computer Science*, vol. 219, pp. 1297–1308, 2023.
- [13] C. R. Jack *et al.*, "NIA-AA research framework: Toward a biological definition of Alzheimer's disease," *Alzheimer's & Dementia*, vol. 14, no. 4, pp. 535–562, 2018.
- [14] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual Review of Biomedical Engineering*, vol. 19, pp. 221–248, 2017.
- [15] C. L. Saratxaga *et al.*, "MRI Deep Learning-Based Solution For Alzheimer's Disease Prediction," *Journal Of Personalized Medicine*, vol. 11, no. 9, p. 902, 2021.
- [16] C. R. Jack *et al.*, "The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods," *Journal of Magnetic Resonance Imaging*, vol. 27, no. 4, pp. 685–691, 2008.
- [17] M. W. Weiner *et al.*, "The Alzheimer's Disease Neuroimaging Initiative 3: Continued innovation for clinical trial improvement," *Alzheimer's & Dementia*, vol. 13, no. 5, pp. 561–571, 2017.
- [18] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.
- [19] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [20] E. Yagis *et al.*, "Effect of data leakage in brain MRI classification using 2D convolutional neural networks," *Scientific Reports*, vol. 11, p. 22544, 2021.
- [21] N. N. B. A. Kubi and S. Nazir, "Dementia Prediction With Multimodal Clinical And Imaging Data," *International Journal Of Information Technology*, vol. 17, pp. 5–16, 2025.
- [22] S. Kaufman *et al.*, "Leakage in data mining: Formulation, detection, and avoidance," *ACM Transactions on Knowledge Discovery from Data*, vol. 6, no. 4, pp. 1–21, 2012.

- [23] A. Javeed *et al.*, “Machine Learning For Dementia Prediction: A Systematic Review And Future Research Directions,” *Journal Of Medical Systems*, vol. 47, no. 1, p. 17, 2023.
- [24] S. J. Pan and Q. Yang, “A survey on transfer learning,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.
- [25] N. Tajbakhsh *et al.*, “Convolutional neural networks for medical image analysis: Full training or fine tuning?” *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.
- [26] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [27] R. Müller *et al.*, “When does label smoothing help?” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.
- [28] C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” *Journal of Big Data*, vol. 6, no. 1, p. 60, 2019.
- [29] N. Srivastava *et al.*, “Dropout: A simple way to prevent neural networks from overfitting,” *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [30] R. Bin-Hezam and T. E. Ward, “A Machine Learning Approach Towards Detecting Dementia Based On Its Modifiable Risk Factors,” *International Journal Of Advanced Computer Science And Applications*, vol. 10, no. 8, 2019.
- [31] H. Alshamlan *et al.*, “Improving Alzheimer’s Disease Prediction With Different Machine Learning Approaches And Feature Selection Techniques,” *Diagnostics*, vol. 14, no. 19, p. 2237, 2024.
- [32] K. A. Shastry and S. A. Sattar, “Logistic Random Forest Boosting Technique For Alzheimer’s Diagnosis,” *International Journal Of Information Technology*, vol. 15, no. 3, pp. 1719–1731, 2023.
- [33] A. P. Nayaki *et al.*, “Alzheimer’s Disease Detection Using Machine Learning Models,” in *2024 Third International Conference On Electrical, Electronics, Information And Communication Technologies (ICEEICT)*. IEEE, 2024, pp. 1–8.