

Alzheimer's disease multiclass detection through deep learning models and post-processing heuristics

Rani Ghassan AL Rahbani^a, Anastasia Ioannou^b, Tao Wang^c

^aEuropean University Cyprus, Department of Computer Science and Engineering, Nicosia, Cyprus, rr231862@students.euc.ac.cy

^bEuropean University Cyprus, Department of Computer Science and Engineering, Nicosia, Cyprus, an.ioannou@euc.ac.cy

^cCollege Computer and Data Science and International Digital Economy College, Minjiang University, Fuzhou, China

Abstract

Alzheimer's disease (AD) poses a significant challenge globally, impacting millions with its progressive memory loss and cognitive decline. Despite lacking a cure, early detection and intervention can mitigate its effects and improve patients' quality of life. Recent advancements in AD research have leveraged deep learning algorithms applied to brain MRI images, showing promising results in predicting its stages. However, sustainable techniques necessitate further exploration. This paper presents a novel approach integrating two CNN algorithms, ResNet and EfficientNet, along with a post-processing algorithm, to enhance AD diagnosis. Empirical analyses are conducted on two public datasets, ADNI and OASIS, to develop and evaluate the proposed technique. The utilized method capitalizes on the complementary nature of the two CNN models and the tailored post-processing step, employing a weighted averaging ensemble learning technique, achieve superior predictive performance. The uniqueness of the proposed approach lies in its integration of multiple CNN architectures and the inclusion of a specialized post-processing algorithm. By explicitly addressing the limitations of existing methods and showcasing notable accuracies of 98.59% for EfficientNet, 94.59% for ResNet, and 98.97% with post-processing on the first dataset, and 97.25% for EfficientNet, 99.36% for ResNet, and 99.41% with post-processing on the second dataset, the presented work contributes to advancing AD diagnosis and underscores the potential of deep learning in healthcare applications.

Keywords: Alzheimer, Machine Learning, Deep Learning, CNN, EfficientNet, ResNet, Post Processing

1. Introduction

Alzheimer's disease is a neurological condition that progressively impairs memory, thinking, and behaviour, particularly among the elderly people. Early detection of AD is crucial for timely intervention and treatment. As per the report, the projected growth of the U.S. population affected by AD anticipates an increase from 6.5 million to 13.8 million by the year 2060 under current medical circumstances [13] and a global burden of approximately 55 million individuals are affected by dementia, with over 60% of cases found in middle- and low-income countries.

In recent years, deep learning models, particularly convolutional neural networks (CNNs) and transformer-based models, have shown promise in automatic detection and classification of AD using neuroimaging data, such as magnetic resonance imaging (MRI) scans [17] [18]. However, there are still significant gaps in the existing models that need to be addressed for more accurate and reliable diagnosis.

CNN-based models have been widely used for AD detection, leveraging their ability to extract spatial features from neuroimaging data. Several studies have proposed innovative CNN architectures, such as 3D CNN frameworks with multi-level features [54], adaptive hybrid attention networks [47], and deep feature fusion networks [46], to improve the accuracy of AD classification. However, despite these advance-

ments, CNN-based models often struggle with capturing long-range dependencies and global context information from MRI scans, which are essential for accurate diagnosis, especially in the early stages of AD.

On the other hand, transformer-based models, such as Swin Transformer [45] and dimension-centric proximate attention networks [24], have gained attention for their ability to capture long-range dependencies and global context information more effectively than traditional CNNs. These models leverage self-attention mechanisms to analyze relationships between different regions of the brain in MRI scans, thereby improving the accuracy of AD classification. However, transformer-based models have their own limitations, including computational complexity and the need for large amounts of training data.

Despite the advancements in both CNN-based and transformer-based models for AD detection, there is still a need for more robust and interpretable models that can effectively integrate spatial and contextual information from neuroimaging data. Additionally, existing studies often focus on individual aspects of AD diagnosis, such as classification or staging, rather than providing comprehensive solutions that address the entire diagnostic pipeline. Therefore, there is a gap in the literature for holistic approaches that combine the strengths of CNN-based and transformer-based models to improve the accuracy, reliability, and interpretability of AD diagnosis.

Addressing the limitations of existing approaches, this paper proposes a novel and efficient ensemble approach for the automatic detection and classification of AD using MRI images. Our approach integrates two powerful CNN algorithms, EfficientNet and ResNet, along with a post-processing ensemble learning algorithm, combining both CNN algorithms and using a weighted averaging technique, to leverage their respective strengths in spatial feature extraction and long-range dependency modeling, improving the accuracy and the performance of the model.

Additionally, we contribute to the field by conducting extensive experiments on two publicly available datasets, Alzheimer’s Disease Neuroimaging Initiative (ADNI) and Open Access Series of Imaging Studies (OASIS), containing MRI brain images. Through rigorous experimentation, we demonstrate the superiority or parity of our method compared to state-of-the-art techniques in supervised medical image classification tasks across both datasets. By addressing the gaps in existing CNN-based and transformer-based models, our proposed framework offers a holistic approach that combines spatial feature extraction and long-range dependency modeling, thus improving the accuracy, reliability, and interpretability of AD diagnosis. We have also performed a comparison of ensemble and individual deep learning models.

Furthermore, we provide insights into the potential implications of our findings for clinical practice and discuss future research directions in the field of AD diagnosis. By synthesizing key studies and referencing notable research contributions, we aim to contextualize our contributions within the broader landscape of AD detection using deep learning models and neuroimaging data.

The structure of the remaining sections of this paper is outlined as follows: In Section 2, recent research on supervised learning, medical image classification, and various brain diseases along with comparative analysis is reviewed. Section 3 encompasses all the preliminaries and algorithms employed in this study. The method proposed in this paper is detailed in Section 4. Section 5 delves into the public datasets utilized for the development and evaluation of this paper model. It also presents and analyzes the experimental evaluation results. A discussion of the proposed method is presented in Section 6. Finally, Section 7 concludes the paper with closing remarks.

2. Related Work

Clinical data, image analysis, and associated artificial intelligence (AI) models may hold significant potential for positively impacting people’s lives in a relatively short time span ([29]). Indeed, image classification using machine learning and deep learning for medical images such as MRI, PET, and CT scans has proven highly beneficial, enabling new possibilities in medical image analysis.

2.1. Medical Image Classification

As machine learning and deep learning have evolved, breakthrough discoveries have been made in medical image classification, a critical task in medical image analysis. Supervised

learning methods, in particular, have played a significant role in training algorithms to analyze medical images, including X-rays, MRI scans, and CT scans, accurately categorizing them into various classes. This aids in the detection and diagnosis of diseases and abnormalities. In this paper, the proposed approach is validated and recent literature reviews are summarized as follows.

2.1.1. Brain tumors

The authors in [26] suggested two deep learning models. One can perform binary classification with the classes being normal and abnormal for brain cancer. The other is a multiclass classifier with the classes being pituitary, meningioma, and glioma brain cancers. The dataset used is a public dataset of MRI scans, and the models achieved an accuracy of 99.5% for binary classification and 98.7% for multiclass classification.

On the other hand, to identify brain cancers in MRI images, the authors of [32] suggest a deep learning framework based on a CNN. The dataset consists of 3264 MRI images, and the model achieved an accuracy of 93.3%.

Furthermore, [37] offers a thorough overview of deep learning applications for the study of brain tumors. The authors discuss the many deep learning models that have been applied to classify, predict, and segment brain tumors, as well as listing some of the unsolved problems in this field.

Meanwhile, [33] reviews the most current deep and federated learning-based approaches for diagnosing brain tumors. The authors discuss the deep learning models utilized for identifying and categorizing brain tumors, as well as the difficulties in applying these models in practical contexts. They also explore how federated learning could help with some of these issues.

Additionally, [31] offers a thorough analysis of deep learning techniques for segmenting brain tumors. The authors discuss the various deep learning models employed for brain tumor segmentation, along with their difficulties and drawbacks. They also highlight some of the positive potentials for this field of study.

Finally, [34] enhanced previously disclosed tumor classification findings by employing a generic neural network (NN) approach rather than specific processing techniques. Using CNN, the model achieved an accuracy of 90.26%.

2.1.2. Alzheimer

The authors in [30], utilize MRI images for diagnosing AD using deep learning. The method involves extracting information from MRI scans using a CNN.

Furthermore, in [20], the authors employed two strategies. The first technique utilizes 2D and 3D convolution-based basic CNN architectures to process structural brain scans in 2D and 3D T1-weighted MRI modalities from ADNI dataset. The second approach involves using the VGG19 [41] model and other previously trained medical image classification models by applying the transfer learning principle on the same dataset. These techniques were highly effective, achieving AD stage classification accuracies of 93.61% and 95.17% for 2D and 3D scans,

respectively. The pre-trained VGG19 model was further improved, achieving an accuracy of 97% for multiclass classification.

Moreover, in [8], the authors developed a unique and improved Computer-Aided Diagnosis (CAD) system based on a CNN capable of distinguishing between individuals with normal cognitive function and those with AD. The suggested method was evaluated using 18FDG-PET scans from 855 individuals, including 220 AD patients and 220 normal control subjects from the ADNI dataset. The results indicated that the proposed CAD system achieved accuracy, sensitivity, and specificity values of 96%, 96%, and 94%, respectively.

On another note, in [16], the authors presented a novel unsupervised technique that makes use of the ADNI dataset and unsupervised CNN algorithms. The outcomes were encouraging. The method's accuracy when analysing single-slice data was 95.52% for separating AD from mild cognitive impairment (MCI) and 90.63% for separating MCI from normal cognition. But when three orthogonal panels (TOP) of MRI images were used as the data set, the accuracy peaked at 97.01% for AD versus MCI and 92.6% for MCI versus Normal Cognition. Interestingly, the technique combines a supervised classification step with a Support Vector Machine (SVM) and unsupervised feature learning with CNNs, specifically PCANet. As a result, even though the feature extraction process is unsupervised, the entire technique consists of both supervised and unsupervised steps.

Furthermore, the authors in [10] trained four different models using the ADNI dataset: Deep Neural Networks (DNN) achieved 99.2% accuracy, CNN achieved 99.9% accuracy, Deep Autoencoder (DA) achieved 91.95% accuracy, and Deep Boltzmann Machine (DBM) achieved 95.35% accuracy.

At this juncture, the work of authors is delved into in [39], who propose a versatile method utilizing structural MRI and machine learning (ML) for diagnosing AD and moderate cognitive impairment (MCI). They leveraged data from 570 participants in the ADNI dataset and 531 subjects from the OASIS project database to train and test the classifiers. Various classifiers were evaluated and integrated through voting to make judgments. Additionally, they assessed the classifiers' potential for clinical application, their applicability across datasets and techniques (IR-SPGR and MPRAGE), the impact of incorporating graph theory metrics on diagnostic performance, and the relative importance of different brain regions. The "healthy controls (HC) vs. AD" classifier, when trained and evaluated on the combined ADNI and OASIS datasets, achieved a balanced accuracy (BAC) of 90.6% and a Matthew's correlation coefficient (MCC) of 0.811. Similarly, the "HC vs. MCI vs. AD" classifier, trained and evaluated on the ADNI dataset, obtained a MCC of 0.438 and a BAC of 62.1% (with a 33.3% by-chance limit). Notably, hippocampal traits emerged as the most significant contributors to categorization judgments (approximately 25–45%), followed by temporal (roughly 13%), cingulate, and frontal areas (each around 8–13%). The classifiers exhibited robust cross-dataset and cross-protocol generalization. Furthermore, the addition of graph theory metrics did not improve classification performance.

Furthermore, in [50], the authors conducted an extensive cross-dataset evaluation of machine learning models for AD identification. They evaluated models such as Random Forest, Support Vector Machine, Decision Tree, XGBoost, Voting Classifier, AdaBoost, and Gradient Boosting using OASIS dataset. The findings revealed that models incorporating feature selection techniques outperformed those without. Notably, models such as Random Forest, Voting, and Extra Tree demonstrated good accuracy, precision, recall, and F1-Score, ranging from 88% to 93%, when applied to longitudinal datasets. Conversely, models trained without feature selection exhibited lower accuracy, ranging from 40% to 92%. Furthermore, feature-selected models consistently outperformed non-feature-selected models across datasets, with accuracy scores for Random Forest, Gradient Boosting, and XGBoost ranging from 67% to 72%. The importance of feature selection in enhancing the accuracy of AD detection was underscored by variations in precision, recall, and F1-Score metrics among models and datasets.

Moreover, in [48], the authors propose an early diagnosis approach for AD using machine learning, employing a variety of algorithms. The GaussianNB probabilistic technique, which assumes feature independence and utilizes a Gaussian distribution for continuous data, achieved a remarkable accuracy of 96%. Decision Tree, utilizing a tree-like structure of decisions, achieved an accuracy of 89.33% in result prediction. The Random Forest ensemble of decision trees, aimed at improving prediction and mitigating overfitting, also demonstrated noteworthy accuracy of 96%. XGBoost, known for creating gradient-boosted decision trees efficiently and scalably, achieved an accuracy score of 93.33%. Notably, the Voting Classifier, an ensemble approach combining predictions from multiple algorithms, achieved the highest validation accuracy of 96%, highlighting its effectiveness. Lastly, GradientBoost achieved a 92% accuracy rate by optimizing a loss function using a group of weak prediction models. These outcomes underscore the efficacy of the ensemble-based Voting Classifier, which exhibited the highest accuracy of 96% in AD prediction using the OASIS dataset. The study underscores the significant promise of machine learning algorithms in facilitating early diagnosis of AD.

In [25], the authors utilize the OASIS dataset to explore early-stage AD prediction. The study employs a range of machine learning methods, including Random Forest, Decision Tree, Support Vector Machine (SVM), Gradient Boosting, and Voting classifiers, achieving an impressive validation average accuracy of 83%. Performance metrics used for evaluation include F1-score, Precision, Recall, and Accuracy. Decision Tree, with data division based on feature cutoff values, achieves an accuracy of 80.46%, with precision, recall, and F1-score values ranging from 78% to 80%. Notably, Random Forest, an ensemble of decision trees aimed at reducing overfitting, achieves an accuracy of 86.92% with 85% precision at 81% recall and a corresponding F1-score of 80%. Support Vector Machine, utilizing hyperplanes to classify data points, achieves an accuracy of 81.67%, with precision, recall, and F1-score values at 77%, 70%, and 79% respectively. XGBoost, employing

gradient boosting for enhanced speed and performance, attains a high accuracy of 85.92%, with precision and recall both at 85%. Similarly, the Voting Classifier achieves an accuracy of 85.12%, with precision and recall both at 83%. The study concludes that the Voting Classifier, Random Forest, and XGBoost are the most effective models for predicting AD in its early stages, verified by cross-validation to ensure reliability. In addition to traditional machine learning algorithms, the study incorporates deep learning techniques. Specifically, deep convolutional autoencoders are used to extract high-level features from the imaging data. These features are then used in conjunction with machine learning algorithms (SVM, Random Forest, Decision Tree, XGBoost, Voting Classifiers) to enhance the analysis of AD. The deep convolutional autoencoder architecture consists of multiple convolutional and deconvolutional layers designed to capture the intricate patterns in the imaging data, which are critical for accurate early-stage AD prediction. The models are trained using publicly accessible datasets from Kaggle and OASIS in these experiments. The use of these public datasets ensures a degree of reproducibility, as other researchers can access the same data and replicate the experiments. However, the paper acknowledges the inherent complexity of early AD prediction, particularly for accurate and timely diagnosis (potentially aiding clinicians and lowering mortality rates). To address the complexity of the problem, the study emphasizes careful feature selection and data preparation. It uses advanced tools and techniques to remove unnecessary features and include pertinent metrics like education and MMSE (Mini-Mental State Examination) scores. These steps are crucial for achieving high accuracy and reliable performance in predicting early-stage AD. Overall, the integration of deep learning with traditional machine learning models, along with the use of public datasets, enhances the robustness and reproducibility of the study's findings, providing valuable insights into early-stage AD prediction.

This study [11] considered data from the National Alzheimer's Coordinating Center with a host of clinical variables including prior medical conditions, neuropsychological tests, and neuroimaging tests. In this paper, a deep ensemble learning framework was leveraged for enhancing the predictive accuracy. This framework employs a triple-layer architecture: in the first layer, sparse auto-encoders reduce features and then a deep belief network ranks various garnered predictions with the aid of an additional layer termed the voting layer, before finally being optimized by the neural networks in the optimization layer into the final prediction. The proposed method outperformed others in ensemble methods in terms of classification accuracy by about 4%, making it a potential tool that could help in the diagnosis and management of Alzheimer's disease in general practice.

The researchers in [27] drew 2125 brain scans against the ADNI database, which were categorized by the Alzheimer's disease, mild cognitive impairment, and normal cognitive function of subjects. In this study, an ensemble model was proposed to increase the accuracy of classification, which integrates extreme gradient boosting, decision trees, support vector machines with polynomial kernel techniques. This is an exam-

ple of a Master-Slave architecture with grid-based tuning done for better optimization and cross-validation for the robust evaluation of performance. The results from the ensemble model were good: an accuracy of 89.77 percent before optimization and a further improvement in accuracy to 95.75 percent after the optimization of parameters, thereby outperforming all other machine learning models tested in this study.

This paper [15] focused on the examination of Alzheimer's disease with the help of an OASIS dataset that colleague image and MRI reports. This paper implements ensemble learning, which involves the combination of machine learning algorithms: Random Forest, SVM, XGBoost, Adaboost, Decision Trees, Voting Classifier, and Gradient Boosting. In this study, an Artificial Neural Network model was developed with features such as a dense layer, dropout rate, and specific activation functions, after which its performance was evaluated. The known metrics used in evaluating it were accuracy, precision, and sensitivity. The ANN had the greatest test accuracy of 91.96%, thereby outperforming gradient boosting at 85.7% and voting classifier methods at 83.04%; this indicates improved AD diagnosis capability.

In this research [18], MRI files from the ADNI dataset and a local dataset from Firoozgar hospital were used to train and validate a model with regard to Alzheimer's brain disease detection. The MRI images used were T2-weighted and axial-view. They have developed in this line a weighted probability-based ensemble method that integrates six different CNN classifiers: DenseNet201, DenseNet169, DenseNet121, ResNet50, Inception-Resnet V2, and VGG19. The formula that mathematically determines the output of the ensemble model is $O_j = \sum_{i=1}^6 w_i \times \alpha_j^i$, where O_j is the sum of weighted probabilities for class j , w_i is the weight of the i -th classifier, and α_j^i is the probability value of class j in the i -th classifier. The paper achieved good results in differentiating between LMCI and AD: ensemble approach delivered 98.57% for NC vs. AD, 96.37% for NC vs. EMCI, 94.22% for EMCI vs. LMCI, and 99.83% for LMCI vs. AD. Validation on the classification of the three categories yielded 88.46% on the local dataset. Even though the individual class CNN models were not performing very well, this ensemble method provided promising potential. The authors confirm that more comprehensively sized validation datasets would be needed to further establish the generalizability of results.

Lastly, authors in [22] examined how well brain imaging data could be used to distinguish between subjects with AD, mild cognitive impairment (MCI), and cognitively normal (CN) subjects using an Adaptive Deep Belief Network (Adaptive DBN) model. The study made use of MRI and 18F-FDG-PET scans from the ADNI archive. The main technique was optimising the network structure by layering Adaptive RBMs with a neuron generation-annihilation algorithm, and subsequently creating a DBN by stacking these RBMs. To improve classification power and facilitate group learning, the DBN also adopted a teacher-student based learning strategy. The outcomes showed promise, especially when it came to differentiating between AD and CN. In this classification task, the Adaptive DBN achieved a test set accuracy of 96.7% for MRI images and 98.8% for

PET images. The test set accuracy for MCI vs. AD using MRI scans was 98.3%, even though accuracy for the MCI vs. CN and MCI vs. AD classifications wasn't stated explicitly for the training set. In terms of early AD and MCI detection, the Adaptive DBN model performed better overall than other CNN (Convolutional Neural Network) models, indicating its potential for precise brain pathology classification

2.1.3. Huntington's Disease

Authors in [53] explore the possibility of categorizing patient illness severity based on each footstep's pressure data through the use of deep learning algorithms. Tests utilizing the Unified Huntington's Disease Rating Scale (UHDRS) Motor Subscale showed that the use of VGG16 and related modules resulted in a classification accuracy of 89%.

On the other hand, the authors of [35] employed multidimensional pattern evaluation approaches to a number of derived voxel-based and segmented region-based datasets to see if information about illness state could be decoded from MRI images. Utilizing support vector machines (SVM) and linear discriminant analysis (LDA), it was discovered that several fundamental, emission-weighted, and functional MRI measurements could accurately distinguish pre-Huntington's disease (HD) and controls with up to 76% accuracy.

2.1.4. Amyotrophic Lateral Sclerosis (ALS)

The goal of [49] was to predict the survival of ALS patients using clinical traits and MRI data with deep learning. The authors classified 135 ALS patients as short, medium, or long survivors. The accuracy of the deep learning model employing data from brain morphology was 62.5%, clinical features was 62.5%, and MRI accuracy was 68.8%. The accuracy rose to 84.4% when the three models were combined.

2.1.5. Traumatic Brain Injury (TBI)

In the case of TBI in both humans and small animals, the authors in [36] offered a CNN-based brain extraction system for skull stripping based on multi-contrast MR imaging. Skull stripping accuracy improved significantly in experiments using MR images of mice and humans. The model was tested on 19 human patients with mild to severe TBI and scored 97.19% accuracy. Furthermore, the same model was applied to 16 images of normal mice, scoring 94.86% accuracy, and to 10 mice brains with TBI, scoring 95.43% accuracy.

2.1.6. Comparative Analysis

As this study is dealing with CNN models, highlighting the difference between CNN and other methods such as Vision Transformer (ViT), diffusion models, and Recurrent Neural Network (RNN) is a must.

In fact, In [44], [54], and [18], the authors emphasize the significance of CNNs in image processing and AD identification. For the purpose of extracting features and identifying spatial hierarchies and local patterns, CNNs are crucial. But they can

miss crucial information, especially when working with complex neuroimaging data, and they have trouble capturing wide-ranging, long-range dependencies across all brain volumes. Despite the difficulties, the authors in [54] and [18] revealed an accuracy of 94.61% and 88.46%, respectively.

Moreover, as there are usually few annotated datasets in the field of medical imaging, CNNs' need on huge datasets poses a substantial challenge. This emphasizes the necessity of coming up with innovative solutions to the data shortage problem. The limitations of CNNs in image-based disease diagnosis are analysed by the authors in [47] and [43], who note that model performance may be hampered by the models' reliance on domain expertise for feature selection. They achieved accuracies of 98.53% and 82.2%, in turn. Notwithstanding these difficulties, CNNs are still crucial for enhancing visual comprehension and advancing a variety of applications.

On the other hand, things are both challenging and exciting in the case of transformer-based solutions, such the Vision Transformer (ViT). Transformers (like ViT) are promising in many areas, but they pose a significant challenge to medical imaging because of their voracious appetite for data [44]. Furthermore, despite transformers' exceptional ability to capture long-range relationships, a significant issue with their interpretability remains, making it challenging for doctors to trust and understand the choices made by the models [44]. However, transformer-based models show that they can express global dependencies and can handle both sequential and non-sequential input intelligently, thanks to the self-attention mechanism [54]. Nonetheless, careful design and training are needed due to their broad use in natural language processing tasks and the intricate requirements of adapting them to 3D medical data [54]. A novel deep feature fusion network that integrates EfficientNet-B01, Global Context Network (GCN), Hybrid Multi-Focus Attention Block (HMAB), and Group Shuffle Depth-wise Convolution (GSDW) is proposed by the authors in [47] and [46]. This technique combines the advantages of transformers and CNNs to effectively incorporate spatial data and global contextual information to enhance MCI classification performance. It offers a potential remedy for the issues pertaining to the interpretation of sMRI data in neurological conditions, by achieving accuracies of 98.53% and 77.2%, respectively. In [45], the writers conduct a comprehensive examination of vision transformers (VTs) and their hybrid counterparts, attaining a precision of 79.8%. Vision transformers, with their self-attention processes, have emerged as a potential replacement for CNNs. Their proficiency is in recognising interrelated elements in pictures. Their capacity to represent local correlations might be limited, in contrast to CNNs, which would affect generalisation. Scientists have created Hybrid Vision Transformers (HVTs), which blend convolutional and self-attentional processes, in an effort to overcome issue. These CNN-Transformer designs offer exceptional performance for various vision applications by combining the advantages of local and global picture representations. The research also provides a taxonomy of modern HVT designs, outlining positional embeddings, convolutional components, attention mechanisms, and multi-scale processing to aid in the comprehension of hybrid model landscapes. The pa-

per suggests more investigation and study in this emerging field of architecture by emphasizing the usefulness of hybrid vision transformers in a variety of computer vision applications. It also emphasizes how vital HVTs are in establishing a connection between local and global picture representations.

Moreover, diffusion models for AD diagnosis are thoroughly examined, and the results show a number of creative strategies and techniques [55; 14; 21; 17; 40]. [55] explores the spatiotemporal dynamics of tau protein misfolding and suggests a reaction-diffusion model supported by Physics-Informed Neural Networks (PINNs) and symbolic regression. Their model, which simulates diffusion dynamics using a graph representation, captures the distribution of tau proteins in the brain using data from the ADNI. Similarly, to improve the accuracy of illness categorization, [14] present a dynamically changed hypergraph diffusion model for AD diagnosis, coordinating semi-supervised hypergraph learning. Using an alternate optimization strategy, their solution combines labelled and unlabeled datasets, promoting diffusion processes that are smoothly driven by the hypergraph p-Laplacian, reaching an accuracy of 92.11%. Furthermore, [21] highlight the critical role that Diffusion Tensor Imaging (DTI) plays in the early detection of AD, clarifying the usefulness of DTI scalar metrics, such as mean diffusivity (MD) and fractional anisotropy (FA), in identifying white matter alterations that are suggestive of cognitive decline. Their multi-modality MRI fusion method, that achieved an accuracy of 97%, combines structural MRI and DTI data to provide a reliable way to classify AD phases. The importance of DTI-derived features, especially in superficial white matter (SWM), is further explained by [17], highlighting the usefulness of DTI reconstruction techniques like q-space diffeomorphic reconstruction (QSDR) as possible biomarkers for AD diagnosis. The authors attained an accuracy of 95.8%. Last but not least, [40] offer a thorough analysis of diffusion models in AD biomarker research, highlighting the significance of DTI in defining brain anomalies at different phases of the illness. Even with its effectiveness, DTI should be used in conjunction with additional imaging modalities to provide a thorough diagnosis of AD. When taken as a whole, these research highlight how crucial diffusion models are to understanding AD pathophysiology and improving diagnostic precision.

Finally, numerous studies investigate the potential of recurrent neural networks (RNNs), in particular the Long Short-Term Memory (LSTM) architecture, in capturing temporal dynamics and sequential patterns present in patient data with regard to AD diagnosis and prediction [12; 19; 9; 52; 51]. Using LSTM-based RNNs, [12] provide an automated prediction framework that forecasts the course of AD by examining patient biomarkers at various time points. Using the complex patterns of disease progression, the model, trained on the ADNI dataset, shows greater accuracy in distinguishing between AD and mild cognitive impairment (MCI). The importance of LSTMs in AD prediction is also highlighted by [19], who credit their success to their ability to handle time-series data, which is essential for early disease detection. LSTMs demonstrate their capacity to capture temporal dynamics suggestive of disease development by including patient data from prior visits into the prediction

process. To diagnose AD through EEG data categorization, [9] expand on the use of LSTM-based RNNs by utilising the architecture’s ability to process sequential inputs and store information for long stretches of time. Their research highlights how important LSTM units are for reducing problems such as gradient vanishing and explosion, which improves the model’s ability to learn from long-term relationships in EEG signals. Additionally, [52] examine how RNNs and LSTMs may be integrated into speech analysis for the purpose of identifying AD, emphasising how well these models can capture linguistic and acoustic patterns linked to the illness. Researchers gain significant improvements in classification accuracy by combining these architectures with other neural networks and attention mechanisms, highlighting the promise of deep learning models in speech-based diagnostics. [51], on the other hand, depart from RNN and LSTM models and concentrate on a multi-task deep learning (MTDL) framework that uses a deer hunting optimization (DHO) technique to optimize CNN model hyperparameters for AD classification and segmentation of the hippocampus. Although their study does not directly address RNNs or LSTMs, it does add to the larger body of work on deep learning-based AD prediction models. All of these studies highlight how flexible and effective RNNs are in capturing the temporal dynamics and sequential patterns that are essential for diagnosing and predicting AD, especially the LSTM variations.

3. Background & Preliminaries

3.1. Supervised Learning

Supervised learning, a subset of machine learning and artificial intelligence, is distinguished by its approach to training algorithms to categorize data correctly using labeled datasets. The model adjusts its weights as input data is provided, refining its adaptation during the cross-validation process. Supervised learning plays a crucial role in enabling organizations to develop scalable solutions for various real-world challenges [7].

3.2. Deep Learning

Deep learning is a machine learning method that instructs algorithms to learn by mimicking human cognitive processes. Driverless cars utilize deep learning as a crucial technology to recognize stop signs and distinguish individuals from objects. It is essential for enabling voice commands on consumer electronics, including tablets, smartphones, and TVs. Recently, deep learning has garnered significant interest, and for good reason—it is producing outcomes that were previously unattainable.

In deep learning, a computerized model learns to carry out categorization tasks by analyzing images, text, or sound. Deep learning models can achieve remarkable precision, sometimes even surpassing human capability. Training these models requires a substantial collection of labeled data and sophisticated multi-layered neural network designs [6]. Deep learning relies heavily on CNN, specialized architectures designed to process data with grid-like topology, such as images and audio. Image recognition tasks have been transformed by CNNs, which

automatically learn hierarchical representations from raw pixel data.

3.3. Convolutional Neural Network (CNN)

As stated in the Section 3.2, CNN is a type of neural network composed of neurons primarily used for image identification and processing, owing to its ability to detect similarities in images. CNNs are specifically designed to efficiently process and analyze images through a series of convolutional and pooling layers. These layers extract relevant features from the input images, capturing hierarchical patterns and spatial relationships. Typically, deep learning models use CNNs with more than just a few layers. The first notable deep model is AlexNet (2012), which has 7 layers, as opposed to earlier "shallow" models with only 2-3 layers. Moreover, advancements in CNN architectures, such as residual connections and attention mechanisms, have further improved their performance and interpretability. While CNNs are powerful tools, their training process requires a large amount of labeled data. Fig. 1 illustrates how CNNs work in image classification.

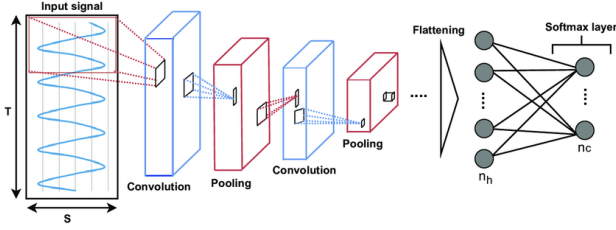


Fig. 1: An example of a CNN model that has c output classes comprised of a softmax layer, convolutional layers, pooling layers, and h dense layers. After deep features are extracted, convolutional and pooling layers handle the input data before sending it to dense layers [23].

3.3.1. Activation Functions

In deep neural networks, the result of the transfer function is passed through an activation function at each node of the network to identify non-linear relationships between the inputs and transform them into more meaningful outputs. Activation functions are utilized to introduce non-linearity into the network. Each activation function has its own unique properties and is suitable for certain use cases.

Different activation functions have different mathematical formulas. Below the two main activation functions that were employed for the purpose of this paper are discussed.

1. Rectified Linear Unit (ReLU)

Next, let us discuss the activation function utilized for the research purposes. In deep neural networks, ReLU (Rectified Linear Unit) is a non-linear activation function. It addresses the vanishing gradient issue, enabling models to be trained more rapidly and perform more effectively. The ReLU formula is as follows:

$$g(z) = \max(0, z) \quad (1)$$

The ReLU activation function is differentiable at all values except at zero. For values greater than zero, the function outputs the input. However, for negative values, the

function returns zero. Thus, the ReLU activation function effectively replaces negative values with zero, while preserving positive values unchanged. This behavior can be expressed as follows:

$$g(z) = \max(0, z) = \frac{z + |z|}{2} = \begin{cases} 0, & \text{if } z \leq 0 \\ 1, & \text{if } z > 0 \end{cases} \quad (2)$$

Due to its capability in addressing the issue of vanishing gradients, ReLU is commonly employed in all convolutional layers except the final convolutional layer, which serves as the output layer providing the final prediction. Fig. 2 depicts the ReLU Function Graph that shows the output of ReLU activation function plotted against its input.

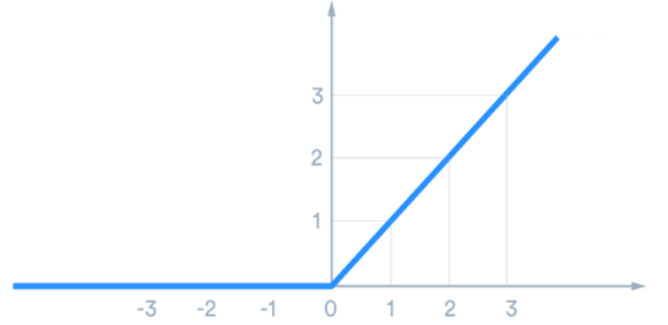


Fig. 2: ReLU Function Graph

2. SoftMax

A vector of K real numbers is transformed into a distribution of K potential outcomes using the SoftMax function. It is applied in multinomial logistic regression and is a generalization of the logistic function to many dimensions. According to Luce's choice axiom, the SoftMax function is frequently employed as the final activation function of a neural network to normalize the output of the network to a probability distribution across expected output classes. The SoftMax function normalizes a vector z of K real numbers into a probability distribution with K probabilities corresponding to the exponentials of the input values. It accepts this vector as an input. In other words, certain vector components before applying SoftMax could be negative or higher than one, and they might not add up to 1. However, after using SoftMax, every factor will be in the range $(0,1)$, and each of them will add up to 1, thus they can be interpreted as probabilities. Additionally, higher probability will result from greater input components. The SoftMax formula is as follows.

$$g(z) = \text{softmax}(z_k) = \frac{e^{z_k}}{\sum_{j=1}^K e^{z_j}} \quad (3)$$

for $k = 1, \dots, K$ and $z = (z_1, \dots, z_K) \in \mathbb{R}^K$.

As mentioned before, the SoftMax activation function is utilized as the final activation function of a neural network, typically in scenarios where classification involves multiple classes. Generally, the Softmax function "softens" the differences between the scores, ensuring a more balanced distribution of probabilities. In this case, with a total of four classes, this paper will be employing SoftMax as the activation function in the output layer for all deep learning algorithms. Fig. 3 depicts the graph of the SoftMax function, that shows the output probability for each class in a classification problem.

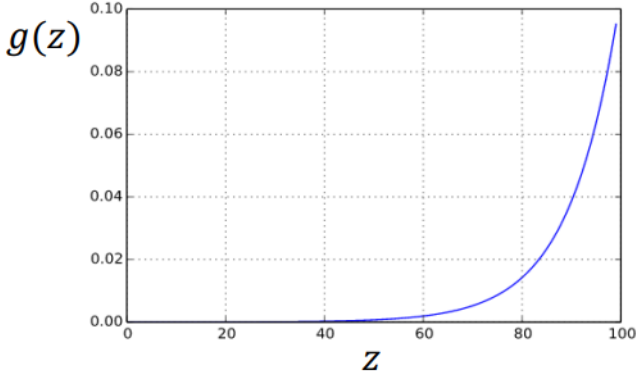


Fig. 3: SoftMax function graph

3.3.2. Loss Function

Let us now discuss the Loss Function, which is a mathematical function that calculates the discrepancy between a neural network's predicted output and the ground-truth output. It is utilized to train the neural network by adjusting its weights and biases in order to minimize this discrepancy. The type of problem being addressed, such as regression or classification, determines the appropriate loss function to be used.

A commonly used loss function in machine learning is categorical cross-entropy, which quantifies the discrepancy between the predicted and actual probability distributions. It is frequently employed in multi-class classification scenarios where the output may belong to more than one class. The categorical cross-entropy loss function of the distribution p relative to a distribution y over a given set is given as:

$$L(y, p) = - \sum_{i=1}^c y_i \log(p_i), \quad \text{for } c \text{ classes,} \quad (4)$$

where, y_i is the truth label and p_i is the SoftMax probability for the i^{th} class. In this paper, as there are a total of four classes, categorical cross-entropy will be used as loss function.

3.3.3. Optimizers

Optimizers are methods used in deep learning to adjust the model's parameters during training, aiming to minimize a loss function. By iteratively updating weights and biases, they enable neural networks to learn from input data.

Several optimizers are available, with one of the most commonly used being the Adam optimizer, which adapts the learning rates for each parameter. It combines the advantages of AdaGrad and RMSProp. Indeed, Adam is widely recognized in deep learning for its ability to achieve good results quickly. Fig. 4, taken from [28], illustrates the effectiveness of the Adam optimizer compared to other optimizers. The following formula represents the Adam Optimizer:

$$\theta_{t+1} = \theta_t - \frac{\alpha \cdot \hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \quad (5)$$

where, θ_{t+1} represents the parameters after the update, θ_t represents the current parameters before the update, α is the learning rate, a hyperparameter that determines the size of the step taken towards the minimum of the loss function, \hat{m}_t is the bias-corrected first-moment(mean) vector estimate of the gradients(the slope of the loss function) at iteration t , \hat{v}_t is the bias-corrected second-moment(uncentered variance) vector estimate of the gradients at iteration t and ϵ is a small scalar added to prevent division by zero and maintain numerical stability.

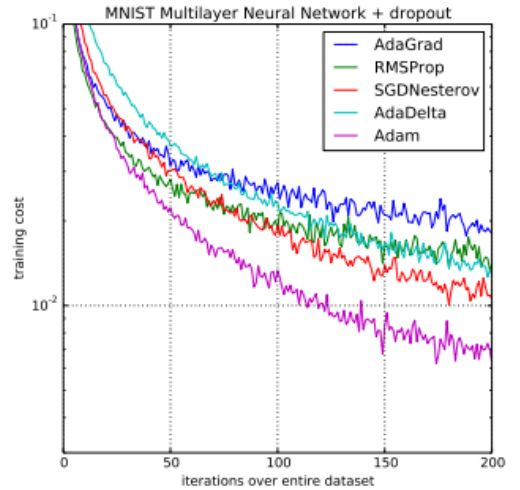


Fig. 4: Comparison of Adam to other optimization algorithms training a multi-layer perceptron taken from [28]

3.3.4. EfficientNet

EfficientNet is a CNN design and scaling technique that employs an additive coefficient to uniformly scale all dimensions of depth, breadth, and resolution. The EfficientNet scaling technique consistently increases network breadth, depth, and resolution using a set of predefined scaling coefficients, contrasting with the conventional practice of arbitrarily scaling these variables [42].

The multi-objective neural architecture employed by EfficientNet-B0, a mobile-size baseline network, aims to achieve accuracy and FLOPS objectives. The model's architecture is depicted in Fig. 5 and was influenced by Mnas-Net.

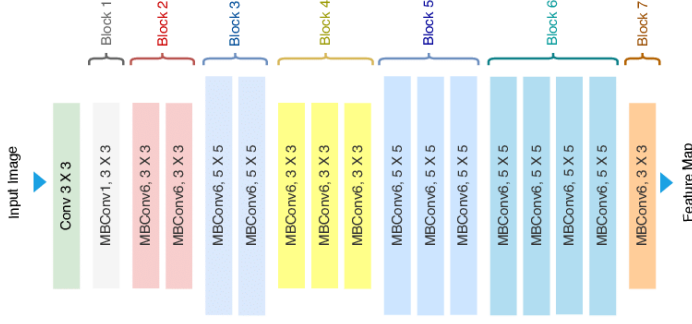


Fig. 5: EfficientNet-B0 Architecture [2]

3.3.5. ResNet

A CNN design known as Residual Network (ResNet) overcomes the "vanishing gradient" issue and enables the construction of networks with hundreds or even thousands of convolutional layers, surpassing simpler systems [5]. By utilizing the layer inputs as a guide, the weight layers within ResNet learn residual functions. Bypassing connections facilitate identity mappings, which are added to the layer outputs [38]. The ResNet architecture adheres to two fundamental design principles: firstly, each layer contains the same number of filters; secondly, to preserve the spatial dimension of the data feature map processed by the convolutional layers, even if the size of the data map is halved, it employs twice as many filters.

The bottleneck building block is utilized in the 50-layer ResNet. A bottleneck residual block, sometimes referred to as simply a "bottleneck", employs 11 convolutions to reduce the number of parameters, thus significantly expediting the training of each layer. Instead of using a stack of two levels, it utilizes three layers. An illustration of the ResNet50 architecture is shown in Fig. 6.

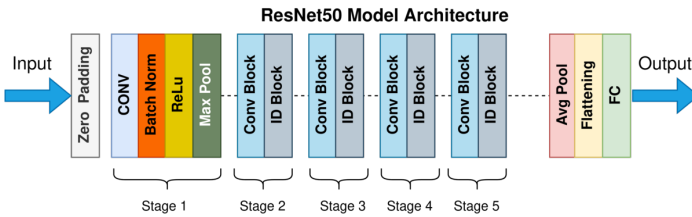


Fig. 6: ResNet50 Architecture [3]

3.4. Evaluation Parameters

The following four metrics are commonly used to assess classifier performance: Accuracy, Precision, Recall, and F1-score. To evaluate these metrics, the following indicators are utilized:

True positives (TP) are outcomes that are both anticipated and actual.

False positives (FP) are predictions of positive results that turn out to be negative.

True negatives (TN) are results that are both predicted and actually negative.

False negatives (FN) are predicted negative results that turn out to be positive.

3.4.1. Accuracy

The accuracy ratio, which measures the number of accurate predictions to all predictions, is the simplest basic metric for evaluating the model's performance. However, accuracy may not perform well with imbalanced datasets. The equation for accuracy is as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

3.4.2. Precision

The precision metric represents the ratio of true positives to the sum of true positives and false positives. It essentially evaluates the accuracy of positive predictions. However, precision does not consider true negatives and false negatives. The equation for precision is as follows:

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

3.4.3. Recall

The recall metric represents the proportion of true positives to the sum of true positives and false negatives. Essentially, it measures the amount of correctly identified positive data. However, recall can lead to a higher rate of false positives. The equation for recall is as follows:

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

3.4.4. F1-Score

The F1 score is the harmonic mean of recall and precision. When precision is improved at the expense of recall, or vice versa, the F1 score aims to balance both precision and recall. The equation for the F1 score is as follows:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (9)$$

4. Proposed method

In this paper, deep learning, particularly CNNs, was employed for the classification of Alzheimer's images. Striving to achieve optimal results, models produced by two algorithms: EfficientNet and ResNet were utilized, through fine-tuning hyperparameters and employing several pre-processing steps. Additionally, extra steps are taken by implementing post-processing ensemble learning algorithm using the results of the models generated from EfficientNet and ResNet. Specifically, this algorithm merges both CNN architectures, thereby elevating predictive performance beyond the capabilities of any singular model. This integration results in more accurate, resilient and reliable predictions. Fig. 7 represents the proposed work methodology, while the ensemble learning post-processing algorithm is detailed in Section 5.4.

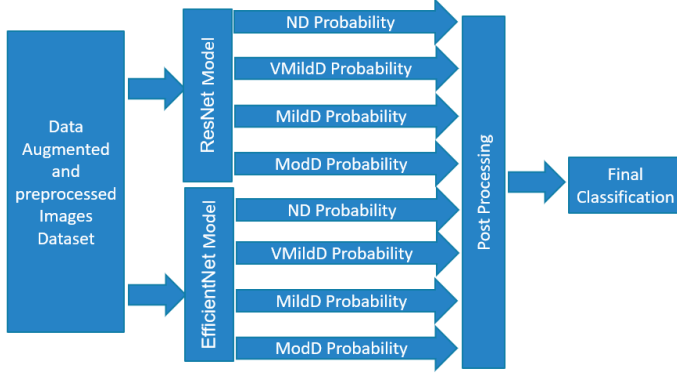


Fig. 7: Methodology of the proposed work, where ND, VMildD, MildD and ModD stand for Non-Demented, Very Mild Demented, Mild Demented and Moderate Demented Probabilities, respectively

4.1. Motivation

This section focused on our innovative ensemble learning strategy aimed at improving the categorization of AD severity from photographs, while Section 3 covered the fundamentals of deep learning.

4.2. Drawbacks of Individual Deep Learning Models

CNNs, in particular, are deep learning models that have shown impressive performance in image categorization tasks. Single models, however, may be biased towards particular data distributions or prone to overfitting. When used to unobserved data, this may result in limitations in generalizability and robustness.

4.3. Ensemble Learning for Improved Outcomes

We propose an ensemble learning strategy that leverages the strengths of two CNN architectures — EfficientNet and ResNet — to overcome these drawbacks and produce the best classification results. When numerous models' predictions are combined through ensemble learning, total performance may be enhanced beyond what would be possible with a single model.

4.4. Suggested Method for Ensemble Learning

Our ensemble technique integrates the predictions from the EfficientNet and ResNet models, trained on images related to AD. We employ a probabilistic rule-based decision-making approach using thresholds. Specifically, a weighted average technique and majority vote approach were employed to enhance decision-making.

1. **Weighted Average Technique:** Predictions from EfficientNet and ResNet are combined using a weighted average. We assign weights based on the performance metrics of each model, giving more weight to the model with higher accuracy and reliability.
2. **Majority Vote Approach:** In cases where the confidence levels of predictions are close, we use a majority vote system. Each model's prediction is considered a vote, and the final decision is based on the majority of votes. This

method helps mitigate the risk of over-reliance on a single model and ensures robust predictions.

Section 5 provides a detailed explanation of the decision-making procedure, including the specific thresholds and conditions used.

4.5. Overfitting

To address the potential issue of overfitting, we included additional experiments with cross-validation to ensure our model generalizes well across different subsets of the data. While the data used was already augmented, we have incorporated another technique, namely the dropout, during training to mitigate overfitting and to validate the model's robustness.

4.6. Novelty and Processing Cost

Even though ensemble learning techniques are not brand-new, our method offers a significant contribution by combining the complementary qualities of the EfficientNet and ResNet architectures to classify the severity of AD from images. This specific combination has not been previously used in this context, nor has the probabilistic rule-based decision-making process we employ (explained in Section 5).

We do, however, recognise that training two deep learning models comes at a significant computational cost (as it is discussed in the Run-Time Performance in Section 5). Training EfficientNet and ResNet separately requires substantial resources, which may not be feasible in all scenarios. We have compared our method with other common approaches, such as using a single model or simpler ensemble techniques like averaging predictions. Our ensemble method has shown superior performance in terms of accuracy and robustness, but at the cost of increased computational demands.

We recognize the need for efficiency and are currently exploring alternative ensemble techniques that could provide comparable results with less computational overhead. For instance, we are investigating methods like model distillation, where the knowledge from multiple models is transferred to a single, smaller model, or using more lightweight architectures in the ensemble. Subsequent work will compare these options, considering both accuracy and computational efficiency, with our current methodology.

The originality of our method lies in the unique combination of EfficientNet and ResNet for AD severity classification and the innovative probabilistic rule-based decision-making process. This approach has not been previously applied in this context, making it a novel contribution to the field.

5. Experiments

The suggested method is evaluated on two publicly available datasets obtained from Kaggle [4], [1], OASIS and ADNI, respectively. The data comprises MRI images collected from various websites, hospitals, and public databases, with each and every label being verified. These datasets underwent pre-processing to ensure consistency before being utilized for training, validation, and testing purposes. The assessment focused

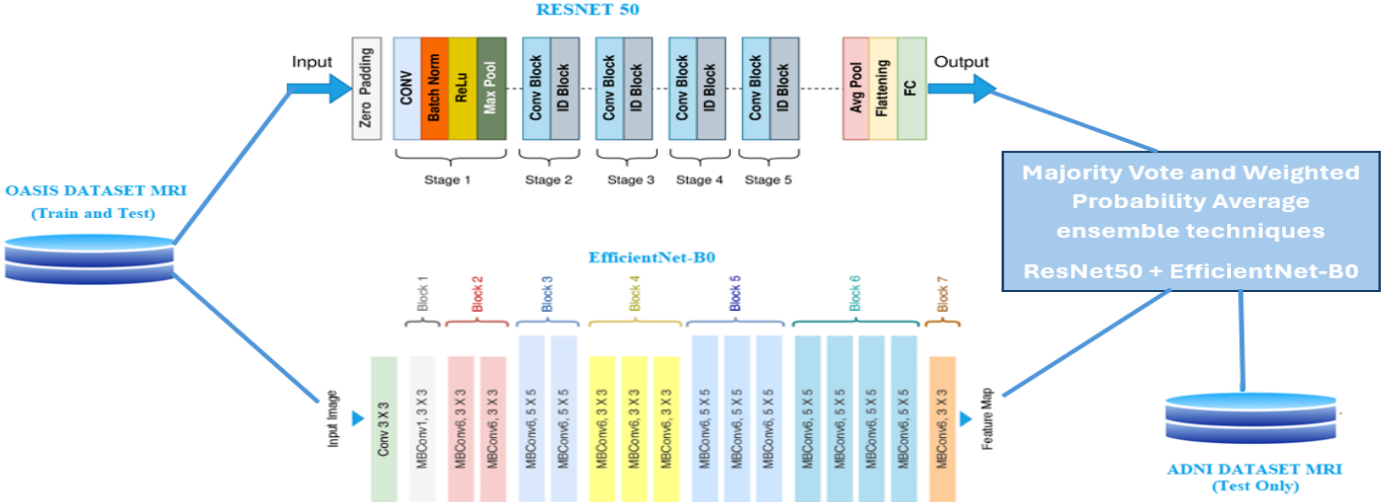


Fig. 8: Ensemble Learning Diagram

on evaluating the performance of two prominent deep learning architectures, EfficientNet and ResNet, across various metrics such as accuracy, sensitivity, and specificity. The utilized approach demonstrates competitive performance across all datasets and the effectiveness of each component is validated.

5.1. Setup

5.1.1. Datasets

The first dataset [4], OASIS, containing 33,984 high-quality augmented Alzheimer's images, was utilized for training, validation, and testing the model. Henceforth, this dataset will be referred to as Dataset1. Furthermore, unique features are found within each of the four classes in the Alzheimer dataset: NonDemented, MildDemented, ModerateDemented, and VeryMildDemented. There are 9600 photos in the "NonDemented" class, most of which are in JPEG format. Of these, 6400 images have the most common size of 200×190 px, with the remaining 3200 images having the size of 180×180 px. This class's photos are all in RGB mode. JPEG is likewise the most common format in the "MildDemented" class, with 8960 photos. With 8064 examples, the size distribution is skewed towards 200×190 px, with 896 photos of size 180×180 px surviving. This class also uses RGB mode for all of its images. Next, "ModerateDemented" class is tackled, where 6464 photos mostly in JPEG format are found. The majority (6400) have dimensions of 200×190 px, while a smaller fraction (64) has dimensions of 180×180 px. Nevertheless, there is a noticeable difference in the size distribution. In a similar vein, every image in this class is RGB mode. Finally, with 8960 photos, JPEG is the most common format in the "VeryMildDemented" class. The distribution of sizes is biased towards 200×190 px, which includes 6720 images. The remaining 2240 images have a size of 180×180 px. Every image in this category is in RGB mode, just like the others. These insights enable more analysis

and modelling efforts by offering a thorough understanding of the dataset's nature. That is why it is decided to make the model accept images with size 200×190 px as it is the dominant size in the dataset.

The second dataset [1], ADNI, consisting of 6,400 pre-processed MRI images, served as a validation dataset to ensure that the pre-trained model provides accurate predictions. Henceforth, this dataset will be referred to as Dataset2. Furthermore, unique features are found in each of the four classes (NonDemented, MildDemented, ModerateDemented, and VeryMildDemented) in the dataset that depicts different levels of dementia. JPEG is the most common format among the 3200 photos in the "NonDemented" class. All of the photos have the same size distribution 128×128 px and are in grayscale mode (L). With 896 photos, JPEG format still has the upper hand in the "MildDemented" class. In a similar vein, every image in this category is in grayscale mode (L) and has a dimension of 128×128 px. Moving on to the "ModerateDemented" class, JPEG remains the only format, albeit there are a significantly fewer number of images—64 total. Each and every image in this class has a measurement of 128×128 px and is in grayscale (L). Finally, JPEG format is dominant in the "VeryMildDemented" class with 2240 photos. The distribution of sizes is identical to the other classes; all images are in grayscale mode (L) and have a measurement of 128×128 px. These insights enable more analysis and modelling efforts by offering a thorough picture of the dataset's nature.

Both datasets contain four classes, arranged in chronological order: Non-Demented, Very Mild Demented, Mild Demented, and Moderate Demented. The distribution of each class for both datasets is presented in Table 1.

Class	Images in Dataset1	Images in Dataset2
Non Demented	9600	3200
Very Mild Demented	8960	2240
Mild Demented	8960	896
Moderate Demented	6464	64
Total	33984	6400

Table 1: Class Distribution for the two datasets

Additionally, a tripartite strategy is employed, dividing the dataset into three essential subsets: the training set, the validation set, and the test set, in a ratio of 7: 1: 2, which is further explained in Section 5.1.2.

Fig. 9 displays a sample of the different Alzheimer images from Dataset1.

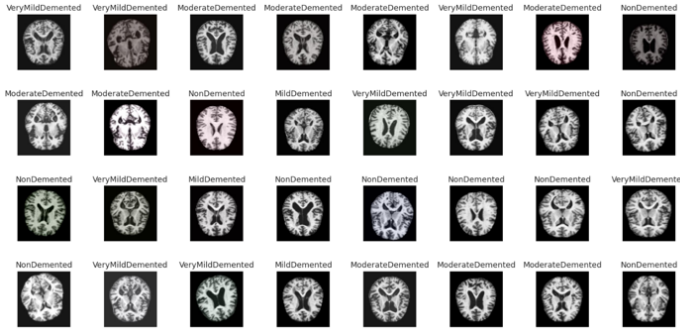


Fig. 9: Example of Alzheimer images from the OASIS dataset showing the four different classes NonDemented, MildDemented, ModerateDemented, and VeryMildDemented

5.1.2. Pre-processing of MRI Datasets

Pre-processing stands as a crucial step in generating optimal classification models. In this paper, the following pre-processing steps were undertaken for Dataset1:

1. The dataset was partitioned into Training (70%), Validation (10%), and Testing (20%) subsets, employing the tripartite strategy, which ensures a balanced distribution.
2. Although most images possessed dimensions of 200x190, a uniform size of 200x190 was enforced by resizing all images.
3. Since the images were already subjected to data augmentation, further augmentation was deemed unnecessary.
4. Given that the dataset was inherently balanced across classes, no additional balancing procedures were performed.

These steps ensure the preparedness of the dataset for the subsequent model training and evaluation.

5.2. Implementation details

In this section the implementation of the two image classification architectures that were employed on the two datasets is described.

5.2.1. EfficientNet

To train EfficientNetB0, MobileNet is utilized as weights with max pooling. Subsequently, batch normalization is incorporated and a dropout of 20% is applied. Following this, two dense layers with 120 units each were added, each employing a ReLU activation function, followed by a dropout of 20%. Finally, the output layer consisted of a dense layer with 4 units (corresponding to the 4 classes) and employed a SoftMax activation function for categorical multiclass classification.

The Adam optimizer and categorical cross-entropy as the loss function are employed. The model was trained using both the training and validation datasets with a batch size of 16 and 15 epochs. Only the best epochs, determined by optimal loss on the validation dataset, were selected.

A key visual depiction of the EfficientNet model's training and validation procedure is provided by Fig. 10, which displays metrics for accuracy and loss. The model's convergence is clearly shown by the way these measures have changed across the training epochs. In particular, the training loss shows a consistent downward trend throughout the course of the epochs, indicating that the model is able to learn and reduce errors on the training set. Simultaneously, the validation loss displays a comparable pattern, confirming that the model performs well outside of the training dataset due to its strong generalisation to new data. Furthermore, the accuracy curves show a steady rise, indicating that the model's capacity to accurately categorise examples increases with time. The training process was effective, as evidenced by the convergence of loss and accuracy gain, confirming the trained EfficientNet model's resilience and dependability in identifying pertinent patterns and features in the data. This convergence indicates that the model is ready to be used in practical applications and gives rise to trust in its forecasting powers.

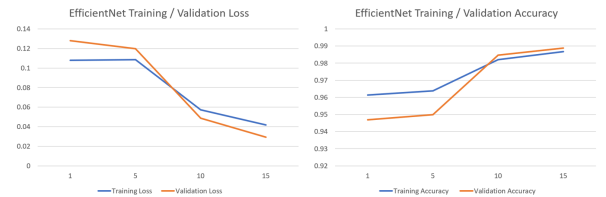


Fig. 10: EfficientNet training and loss accuracy validation

5.2.2. ResNet

To train ResNet50, MobileNet is utilized as weights with max pooling. Batch normalization was then added. Following this, a dense layer with 2048 units employing a ReLU activation function was added, followed by batch normalization. Subsequently, another dense layer with 1024 units employing a ReLU activation function was added, followed by batch normalization. Finally, the output layer consisted of a dense layer with 4 units (corresponding to the 4 classes) and employed a softmax activation function for categorical multiclass classification.

Similar to the EfficientNet model, the Adam optimizer and categorical cross-entropy as the loss function are employed.

The model was trained using both the training and validation datasets with a batch size of 16 and 15 epochs. Only the best epochs, determined by optimal loss on the validation dataset, were selected.

Fig. 11, which shows both accuracy and loss measures, gives a clear picture of the ResNet model’s training and validation dynamics. This visualization provides strong proof of the convergence of the trained ResNet model. Plotted curves indicate how loss values change over time, with a training loss that consistently decreases over the course of epochs. Concurrently, the validation loss curve displays a comparable declining trend, signifying the model’s capacity to expand its generalization much beyond the training set, thus reducing the likelihood of overfitting. Additionally, the accuracy curves exhibit a continuous upward trend, indicating that the model’s ability to correctly categorize examples has improved during the training phase. The training regimen’s effectiveness is demonstrated by the convergence of loss and accuracy improvement, which validates the trained ResNet model’s resilience and ability to identify relevant patterns and features within the dataset. This convergence provides assurance in the model’s prediction ability and supports its suitability for practical application.

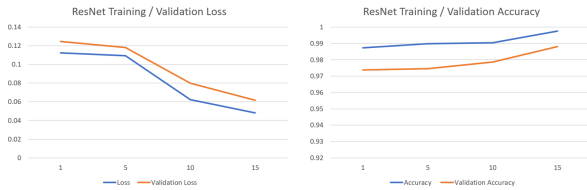


Fig. 11: ResNet training and loss accuracy validation

5.3. Results

The obtained results indicate that both EfficientNet and ResNet architectures exhibit superior performance in AD diagnosis compared to existing methodologies. These high accuracies are explained in the subsequent subsections below.

5.3.1. Results on Dataset1 (OASIS)

As demonstrated in Fig. 12, the applied method surpasses state-of-the-art techniques. Specifically, when utilizing EfficientNet, this method achieved improvements of 7.99%, 2.59%, and 11.67% in prediction accuracy compared to competing methods of [39], [48], and [25], respectively. Notably, the applied method exhibits greater stability across different datasets. Additionally, Fig. 12 illustrates the EfficientNet confusion matrix for the Dataset1 test subset, detailing both correct and incorrect predictions, scoring a 98.59% accuracy.

		Predicted EfficientNet			
		Mild Demented	Moderate Demented	Non Demented	Very Mild Demented
Actual	Mild Demented	1770	0	10	12
	Moderate Demented	0	1294	0	0
	Non Demented	10	0	1877	33
	Very Mild Demented	5	0	26	1761

Fig. 12: EfficientNet confusion matrix for the Dataset1

Furthermore, the performance of EfficientNet for each dementia classification in Dataset 1 (OASIS) is shown in Fig.12, and the computed accuracy, precision, recall, and F1-score percentages are as follows:

Metrics for each individual class:

- **MildDemented:** With a high accuracy of 99.46%, the model proved to be highly capable of identifying cases of mild dementia. Impressive precision (99.16%) showed a low rate of misclassifying other classes as milddemented. Still, there were 37 cases (2.08%) that were incorrectly classified overall, and there were missed cases of mild dementia (98.77% recall).
- **ModerateDemented:** The model achieved perfect scores (100%) in all metrics (accuracy, precision, recall, and F1-score), demonstrating exceptional performance for this category. This shows perfect, error-free identification of cases with moderate dementia.
- **NonDemented:** The model demonstrated low false positive rates (98.12% precision) and good accuracy (98.84%) for nondemented cases. Nevertheless, 79 misclassified cases (4.21%) resulted from some missed NonDemented cases (97.76% recall).
- **VeryMildDemented:** VeryMildDemented performed similarly to MildDemented in terms of accuracy (98.88%), recall (98.27%), and precision (97.51%). This points to a possible mix-up between cases classified as MildDemented and VeryMildDemented, resulting in 31 cases of VeryMildDemented being overlooked and 76 cases overall (4.32%) being incorrectly classified.

Metrics for Comparing Classes:

- **MildDemented vs ModerateDemented and ModerateDemented vs NonDemented and ModerateDemented vs VeryMildDemented:** cored a perfect 100% on all metrics (accuracy, precision, recall, F1-score)
- **MildDemented vs NonDemented:** Attained 99.45% accuracy, 99.44% precision, and 99.44% recall, indicating a low error rate in differentiating between cases of mild dementia and those without it. There were only 20 cases (0.56%) of misclassification.
- **MildDemented vs VeryMildDemented:** displayed a slightly lower accuracy (99.52%) in contrast to the other

comparisons. High precision (99.33%) and nearly perfect recall (99.72%) were demonstrated by mild dementia patients; however, there may have been some misunderstanding between these classes, as evidenced by the 17 misclassified cases (0.48%).

- NonDemented vs VeryMildDemented: 98.40% accuracy was attained. The precision was high (99.75%) but the recall was slightly lower (98.63%) in the NonDemented group. This implies that VeryMildDemented may occasionally be mistakenly assigned to NonDemented individuals.

Similarly, when utilizing ResNet, it is observed that again the proposed method outperforms the state-of-the-art methods. Specifically, this method achieved improvements of 3.99% and 7.67% in prediction accuracy compared to competing methods of [39] and [25], respectively. However, in [48], their approach appears to outperform ours by a 1.41% difference in accuracy score. Additionally, Fig. 13 illustrates the ResNet confusion matrix for the Dataset1 test subset, detailing both correct and incorrect predictions, scoring a 94.59% accuracy.

		Predicted ResNet50			
Actual		Mild Demented	Moderate Demented	Non Demented	Very Mild Demented
	Mild Demented	1774	1	6	11
	Moderate Demented	0	1294	0	0
	Non Demented	71	2	1789	58
	Very Mild Demented	125	8	86	1573

Fig. 13: ResNet confusion matrix for the Dataset1

Example quantitative results are depicted in Fig. 14, showcasing the accuracy score of the utilized approach, which is 98.59% for EfficientNet and 94.59% for ResNet.

The performance of ResNet50 for each dementia classification in Dataset 1 (OASIS) is shown in Fig.13, and the computed accuracy, Precision, Recall, and F1-score percentages are as follows:

Metrics for each individual class:

- MildDemented: The model's accuracy for MildDemented cases dropped to 96.85%, indicating a moderate increase in misclassifications (214 instances, 12.08%). While recall remained high (98.99%), meaning it missed few MildDemented cases, the low precision (90.05%) suggests many other classes were misclassified as MildDemented. The F1-score (94.31%) reflects this trade-off between precision and recall.
- ModerateDemented: The model performed exceptionally well for ModerateDemented, achieving near-perfect scores across all metrics: accuracy (99.84%), precision (99.16%), recall (100%), and F1-score (99.58%). This indicates excellent identification of ModerateDemented

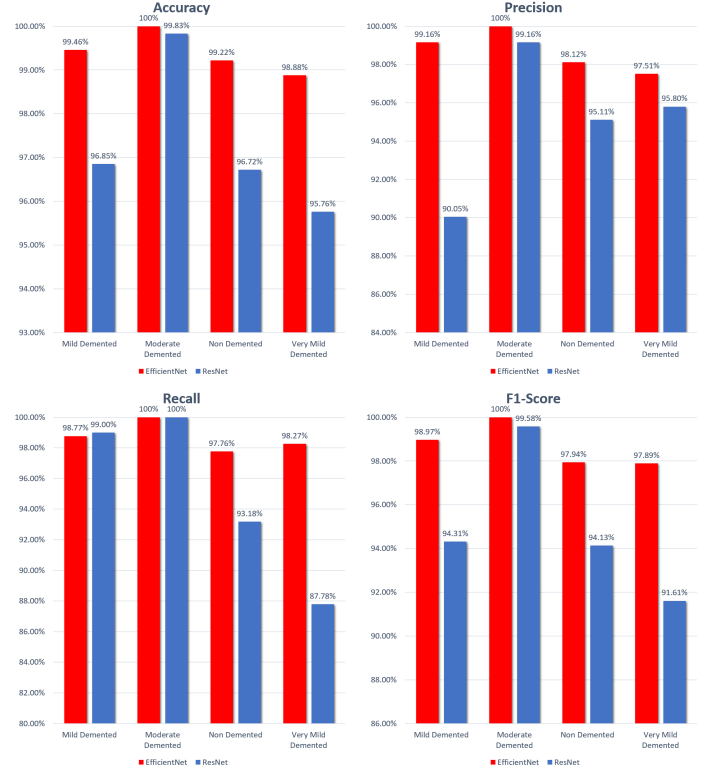


Fig. 14: EfficientNet (with 98.59% accuracy) and ResNet (with 94.59% accuracy) results based on the evaluation parameters

cases with minimal errors (11 misclassified instances, 0.66%).

- NonDemented: NonDemented cases exhibited a decline in accuracy (96.72%), much like MildDemented cases did. With a moderate false positive rate (95.11% precision) and a large number of NonDemented cases missed (93.18% recall), the model misclassified 223 cases (12.47%). This harmony between recall and precision can be seen in the F1-score (94.13%).
- VeryMildDemented: Additionally, there was a moderate decline in accuracy (95.76%) for VeryMildDemented. The model's recall performance was poor (87.78%), meaning that many cases of VeryMildDemented were missed. There were 288 misclassified cases (18.38%) as a result of the moderate false positive rate (95.80% precision). The F1-score of 91.61% is indicative of this overall subpar performance.

Metrics for Comparing Classes:

- ModerateDemented vs. MildDemented: The model achieved nearly perfect accuracy (99.97%), precision (99.94%), recall (100%), and F1-score (99.97%) in distinguishing between MildDemented and ModerateDemented.
- MildDemented vs. NonDemented and MildDemented vs. VeryMildDemented: In contrast to other comparisons, the

accuracy of the comparisons between MildDemented and VeryMildDemented (97.88% and 96.10%, respectively) decreased, indicating a potential for misunderstanding between these classes. There were more misclassifications for MildDemented vs. VeryMildDemented (confusion not quantified) and MildDemented vs. NonDemented (77 instances). This trade-off between accuracy and possible misclassifications is reflected in the F1-score.

- ModerateDemented vs. NonDemented and ModerateDemented vs. VeryMildDemented: These comparisons produced very high accuracy (99.93% and 99.72%, respectively), precision, recall, and F1-scores, as well as excellent differentiation. This suggests that these classes are clearly distinct from one another.
- NonDemented vs. VeryMildDemented: This comparison had a slightly lower recall rate (95.41%) for NonDemented cases, but good accuracy (95.89%) overall. The balance is reflected in the F1-score (96.13%).

In addition, we trained both EfficientNet and ResNet for 15 epochs in order to evaluate their generalizability. For every model, we present the validation accuracy and loss mean and variance. While EfficientNet performed better (mean accuracy: 94.52%, variance: 8.07%) than ResNet (mean accuracy: 92.34%, variance: 4.38%), it was more variable. This implies that more optimisation could improve EfficientNet’s capacity to function consistently across a variety of data distributions. In the Discussion section, we will investigate methods like data augmentation to potentially close the gap and improve EfficientNet’s generalizability as well as go deeper into possible explanations for this observed variance. These findings indicate areas that require more investigation and offer useful standards for choosing models.

We will examine these findings in more detail and go over possible explanations for any observed variations in performance throughout the dataset:

- Mean Validation Accuracy: 94.52% and 92.34%, respectively, were determined to be the mean validation accuracy for EfficientNet and ResNet.
- Variance of Validation Accuracy: ResNet has a variance of 4.38%, while EfficientNet has an 8.07% variance. This suggests that, in comparison to ResNet, EfficientNet’s accuracy varies more between folds.
- Mean Validation Loss: ResNet has a mean validation loss of 0.1578, while EfficientNet’s is 0.2188. Better performance is indicated by a reduced validation loss.
- Variance of Validation Loss: ResNet has a variance of 0.0224 while EfficientNet has a variance of 0.0882. EfficientNet’s validation loss varies considerably over folds, just like accuracy does.

5.3.2. Results on Dataset2 (ADNI)

Dataset 2 was selected as the witness dataset because it is still uncertain whether the results obtained from Dataset 1 are applicable solely to that dataset or can be generalized to multiple Alzheimer’s datasets. It is important to note that Dataset 2 closely resembles Dataset 1.

During this phase, this paper directly utilized the models already trained on Dataset 2, and the ensuing results are discussed below.

Dataset 2 is neither balanced nor augmented with additional data. However, since this dataset was solely employed for testing purposes, there was not any engagement in any pre-processing steps beyond resizing the images to dimensions of 200×190 pixels before classification.

As demonstrated in Fig. 14, when employing EfficientNet improvements of 0.25%, 1.25%, 0.24% and 6.65% are achieved in prediction accuracy compared to competing methods of [20], [8], [16], [39], respectively. The EfficientNet confusion matrix for the Dataset2 test subset, detailing both correct and incorrect predictions, scoring a 97.25% accuracy is illustrated in Fig. 15.

		Predicted EfficientNet			
		Mild Demented	Moderate Demented	Non Demented	Very Mild Demented
Actual	Mild Demented	832	0	8	56
	Moderate Demented	0	64	0	0
	Non Demented	0	0	3196	31
	Very Mild Demented	0	0	81	2159

Fig. 15: EfficientNet confusion matrix

Furthermore, the performance of EfficientNet for each dementia classification in Dataset 2 (ADNI) is shown in Fig.15, and the computed accuracy, precision, recall, and F1-score percentages are as follows:

Metrics for each individual class:

- MildDemented: The model had a high accuracy rate of 99.00% and only 64 misclassifications. Though it had perfect precision (100%, meaning no false positives from other classes were classified as MildDemented), it missed a moderate number of cases (92.86% recall). This harmony between recall and precision can be seen in the F1-score (96.30%).
- ModerateDemented: All metrics (accuracy, precision, recall, and F1-score) were perfectly scored by the model, indicating perfect performance for this category. This shows that every case of moderate dementia was perfectly identified with no errors.
- NonDemented: In comparison to the other classes, the accuracy of the NonDemented cases was slightly lower but still very good at 98.13%. 120 cases were incorrectly classified by the model despite having a low false positive rate (97.29% precision) and missing a small number of NonDemented cases (99.04% recall). This trade-off is reflected in the F1-score (98.16%).

- **VeryMildDemented:** In comparison to other classes, VeryMildDemented showed a somewhat lower accuracy rate of 97.39%. The model's recall performance was poor (96.38%), and it missed a fair amount of cases with VeryMildDemented. 168 cases were incorrectly classified as a result of the moderate false positive rate (96.13% precision). This lower overall performance is reflected in the F1-score of 96.26%.

Metrics for Comparing Classes:

- **MildDemented vs. ModerateDemented and ModerateDemented vs. NonDemented and ModerateDemented vs. VeryMildDemented:** Attained perfect differentiation, scoring 100% on all metrics (F1-score, accuracy, precision, and recall). With zero missed cases and zero false positives for any other class, the model correctly classified all cases of moderate dementia.
- **MildDemented vs. NonDemented:** demonstrated a 99.80% accuracy rate. There were only eight incorrectly classified cases (precision of 99.05%) out of all the MildDemented cases, even though the model correctly recalled every case (no missed cases). The balance is reflected in the F1-score of 99.52%.
- **MildDemented vs. VeryMildDemented:** The model had some difficulty differentiating, despite maintaining a high accuracy of 98.16% (lower precision of 93.69%). Compared to other comparisons, there were more incorrectly classified cases (56), but it accurately recalled every case of mild dementia (not a single one was missed). This trade-off between perfect recall and a higher false positive rate is reflected in the F1-score (96.74%).
- **NonDemented vs. VeryMildDemented:** shown a high degree of accuracy (97.95%). With a high precision of 99.04 percent and a low false positive rate for NonDemented cases, the model's recall (97.53%) was somewhat lowered due to the 81 instances of missed NonDemented cases. This harmony between recall and precision can be seen in the F1-score (98.27%).

Similarly, when utilizing ResNet, improvements of 2.36%, 3.36%, 2.35% and 8.76% are achieved in prediction accuracy compared to competing methods of [20], [8], [16], [39], respectively. Fig. 16, depicts the ResNet confusion matrix for the Dataset2 test subset, detailing both correct and incorrect predictions, scoring a 99.36% accuracy.

		Predicted ResNet50			
		Mild Demented	Moderate Demented	Non Demented	Very Mild Demented
Actual	Mild Demented	896	0	0	0
	Moderate Demented	0	64	0	0
	Non Demented	10	0	3180	10
	Very Mild Demented	8	0	13	2219

Fig. 16: ResNet confusion matrix

The performance of ResNet50 for each dementia classification in Dataset 2 (ADNI) is shown in Fig.16, and the computed accuracy, precision, recall, and F1-score percentages are as follows:

Metrics for each individual class:

- **MildDemented:** Accomplished 99.72% of the time with perfect recall (100.00%), but accuracy was only slightly higher (98.03%) because of a few cases that were incorrectly classified (18).
- **ModerateDemented:** Performed flawlessly (100.00%) across all metrics (recall, accuracy, precision, and F1-score), demonstrating faultless classification.
- **NonDemented:** 99.48% accuracy, 99.59% precision, and 99.38% recall were all demonstrated with good accuracy. Some NonDemented cases (20) were missed, and some instances (33), were incorrectly classified.
- **VeryMildDemented:** demonstrated a high accuracy (99.52%), with good precision (99.55%) and recall (99.06%), was demonstrated by someone . That being said, there were 31 incorrect classifications.

Metrics for Comparing Classes:

- **MildDemented vs. ModerateDemented and ModerateDemented vs. NonDemented and ModerateDemented vs. VeryMildDemented:** Scored a perfect score of 100.00% for all metrics (accuracy, preceision, recall, F1-score), signifying flawless classification between these classes.
- **MildDemented vs. NonDemented:** demonstrated flawless precision (100.00%) and extremely high accuracy (99.76%) with only a few NonDemented cases (10). Although the model identified all predicted MildDemented cases with 100% accuracy, 10 NonDemented cases were incorrectly classified.
- **MildDemented vs. VeryMildDemented:** showed excellent precision (100.00%) and high accuracy (99.74%), but there were also a few VeryMildDemented cases that were missed (8).
- **NonDemented vs. VeryMildDemented:** Good differentiation with some misclassifications is indicated by high accuracy (99.58%) with good precision (99.69%) and recall (99.59%) (23). In this instance, the model yielded balanced precision and recall (about 99.7%) and high accuracy (99.58%). Nevertheless, NonDemented and VeryMildDemented were incorrectly classified in 23 cases.

Example quantitative results are depicted in Fig. 17, showcasing the accuracy score of the applied approach, which is 97.25% for EfficientNet and 99.36% for ResNet.

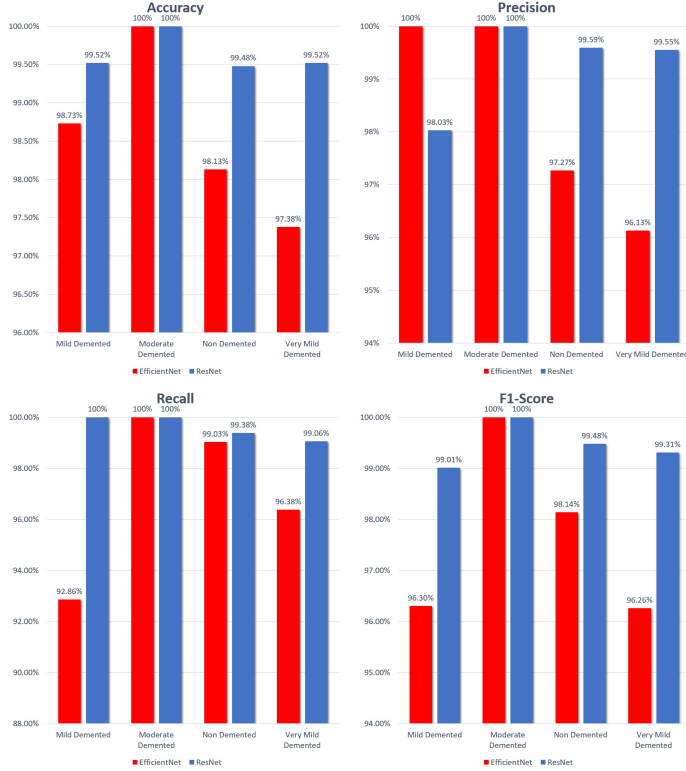


Fig. 17: EfficientNet (with 97.25% accuracy) and ResNet (with 99.36% accuracy) results based on the evaluation parameters

5.4. Post-processing

Leveraging the strengths of multiple models to achieve superior performance is a common practice in deep learning applications. In this work, we propose Algorithm 1 (ADERR), a post-processing method for better classification accuracy in a medical image analysis task, by integrating predictions from two deep learning models: EfficientNet and ResNet50. ADERR takes a data-driven approach influenced by empirical observations from the validation data, in contrast to traditional ensemble methods that frequently rely on intricate mathematical techniques for model fusion.

ADERR's primary strength is its capacity to take advantage of each model's unique advantages while minimising its disadvantages. The algorithm gives priority to the model that demonstrates superior overall results by examining the validation set metrics (accuracy, precision, recall, and F1-score) for every model across various classes. The foundational idea for combining the predictions of the models is this prioritisation.

However, ADERR does more than just use a "majority vote" method. It includes a collection of post-processing rules that, in accordance with predetermined parameters, further hone the selection. These rules take advantage of insights found in the validation data. For example, they give EfficientNet priority for the "ModerateDemented" class because of its clearly better performance in that category. Furthermore, probability thresholds are incorporated into the rules to guarantee that higher confidence level model predictions are given more weight in the ultimate

decision.

The mathematical proof of the model, along with the ADERR that relies on data-driven heuristics and the high results on the validation data on both Dataset 1 (OASIS) and Dataset 2 (ADNI), which will be showed later in this section, are compelling evidence of its efficacy and hold promise for generalizability. This methodology showcases the efficacy of amalgamating model proficiencies via a focused post-processing tactic, culminating in enhanced classification precision within the particular framework of our dataset.

Based on Fig. 14, it is concluded that EfficientNet is a superior classifier compared to ResNet for this dataset, as it achieved better accuracy, precision, recall, and F1-score across almost all classes. Additionally, EfficientNet achieved a perfect score for the ModerateDemented class.

Given these findings, it was decided to develop a post-processing ensemble learning algorithm based on the rules outlined in Algorithm 1 (ADERR). Specifically, a weighted averaging technique has been used by assigning different weights to the predictions of each base model (in this case, EfficientNet and ResNet) based on certain criteria or conditions. In the algorithm stated below, specific conditions were defined for determining when to prioritize the prediction of one model over the other. These conditions include the classification probabilities of each model, the performance of each model in certain classes, and thresholds for probability values. By applying these rules, the algorithm assigns different weights to the predictions of EfficientNet and ResNet, ultimately leading to a final prediction that reflects a combination of both models' outputs, with certain conditions favoring one model's prediction over the other. This approach allows the ensemble model to capitalize on the strengths of each individual model while mitigating overfitting and their weaknesses, leading to improved overall performance.

The rules of such an algorithm are outlined below:

1. By default, the classification by EfficientNet is considered correct.
2. If either EfficientNet or ResNet classifies an image as **"ModerateDemented,"** prioritize the EfficientNet prediction due to its superior performance in this class.
3. If the classification probability by EfficientNet is higher than ResNet's probability, consider the EfficientNet prediction correct.
4. If ResNet's probability is higher than or equal to EfficientNet's probability, **AND**:
 - ResNet's probability is greater than or equal to 83% and EfficientNet's probability is less than or equal to 80%, **OR**
 - ResNet's probability is equal to 100% and EfficientNet's probability is less than or equal to 99%, **OR**
 - ResNet's probability is greater than or equal to 90%, EfficientNet's probability is less than or equal to 95%, and ResNet50 classification is **"VeryMildDemented,"** **OR**

- ResNet’s probability is greater than or equal to 90%, EfficientNet’s probability is less than or equal to 90%, and ResNet50 classification is **”NonDemented,”** THEN adopt the ResNet classification as correct.

5. **ELSE**, retain the EfficientNet classification as correct.

The structured approach outlined in the rules considers the strengths and weaknesses of each model alongside specific conditions related to certain classes and probability thresholds. This approach forms the basis for ADERR, which integrates the probability threshold and empirical results from the training/validation set, subsequently tested on the testing set.

To further support the above argument, we formalized the decision-making process mathematically. This involves showing how the algorithm decides which model’s prediction to use based on their probabilities and predefined conditions. The goal is to demonstrate that the rule adopted by ADERR maximizes accuracy by leveraging the strengths of both models.

Formalization of the Decision Rule

Notations:

- $P_E(x)$: Probability assigned by EfficientNet for the input x
- $P_R(x)$: Probability assigned by ResNet50 for the input x
- $C_E(x)$: Class prediction by EfficientNet for the input x
- $C_R(x)$: Class prediction by ResNet50 for the input x
- $cMOD$ = ”Moderate Demented”
- $cVMID$ = ”Very Mild Demented”
- $cNOD$ = ”Non Demented”

Rules:

1. **Default Rule:** EfficientNet’s classification is considered correct by default:
If none of the specific conditions are met, $C(x) = C_E(x)$
2. **ModerateDemented Class Priority:**

*If either model classifies x as ”cMOD”,
adopt EfficientNet’s prediction :
If $C_E(x) = \text{”cMOD”}$ or $C_R(x) = \text{”cMOD”}$
then $C(x) = C_E(x)$*

3. **Probability Comparison:**

If $P_E(x) > P_R(x)$ then $C(x) = C_E(x)$

4. **ResNet’s Probability Conditions:**

If $P_R(x) \geq P_E(x)$ and

$$\left\{ \begin{array}{l} P_R(x) \geq 0.83 \text{ and } P_E(x) \leq 0.80 \\ \text{or } P_R(x) = 1.00 \text{ and } P_E(x) \leq 0.99 \\ \text{or } P_R(x) \geq 0.90 \text{ and } P_E(x) \leq 0.95 \text{ and } C_R(x) = \text{”cVMID”} \\ \text{or } P_R(x) \geq 0.90 \text{ and } P_E(x) \leq 0.90 \text{ and } C_R(x) = \text{”cNOD”} \end{array} \right\}$$

then $C(x) = C_R(x)$

5. **Fallback Rule:** If none of the above conditions are met:

Adopt EfficientNet’s prediction : $C(x) = C_E(x)$

Proof of the Numerical Basis:

The objective is to show that these rules result in a higher accuracy by combining the strengths of both models. We can frame this problem as a decision-making process where we aim to minimize classification errors.

Step 1: Expected Accuracy Improvement

Given the validation data, assume that EfficientNet has higher accuracy A_E compared to ResNet50 with accuracy A_R for the overall dataset. Let A_M be the accuracy of the model on the ”ModerateDemented” class.

The algorithm prioritizes EfficientNet’s prediction for the ”ModerateDemented” class due to its high performance, thus:

$$A_{combined} = A_E \cdot w_E + A_R \cdot w_R + A_M \cdot w_M$$

where w_E , w_R and w_M are the weights for predictions based on conditions set by the algorithm.

Step 2: Conditional Probabilities

The rules set conditions based on probability thresholds, essentially increasing the weight of more confident predictions:

- EfficientNet is trusted more when its probability is higher than ResNet50.
- For specific classes and high probabilities, ResNet50 is given priority.

Mathematically, the rules can be expressed as a weighted decision function:

$$C(x) = \arg \max_c (w_E \cdot P_E(x) + w_R \cdot P_R(x))$$

where w_E and w_R are determined by the conditions outlined.

Step 3: Majority Vote and Weighted Averaging

The algorithm effectively combines majority voting and weighted averaging:

- Majority Vote: When both models agree on a class, the ensemble adopts this class.
- Weighted Averaging: The probability thresholds ensure that predictions with higher confidence are given more weight, which is critical in cases of disagreement.

Thus, the accuracy improvement can be derived from:

$$A_{ensemble} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(C(x_i) = y_i)$$

where y_i is the true label, and \mathbb{I} is the indicator function. The ensemble accuracy $A_{ensemble}$ is expected to be higher due to the weighted contribution of each model based on their strengths.

The generalized mathematical model above, accurately represents the ensemble learning algorithm’s rules and provides a systematic approach to combining predictions from EfficientNet and ResNet, leading to improved overall performance. The

algorithm for such mathematical model is shown in Algorithm 1, ADERR.

Algorithm 1 ADERR

Constants:

cMOD = "Moderate Demented"
cVMID = "Very Mild Demented"
cNOD = "Non Demented"

Input:

EfficientNet Classification (EC)
EfficientNet Classification Probabbility (ECP)
ResNet Classification (RC)
ReNet Classification Probabbility (RCP)

Output:

Post Processing Classification (PPC)

```

PPC ← EC
if EC = cMOD or RC = cMOD then
    PPC ← EC
else
    if ECP > RCP then
        PPC ← EC
    else
        if (RCP >= ECP)
            and (RCP >= 0.83 and ECP <= 0.8)
            or (RCP = 1 and ECP <= 0.99)
            or (RCP >= 0.9 and ECP <= 0.95 and RC = cVMID)
            or (RCP >= 0.9 and ECP <= 0.9 and RC = cNOD)
        then
            PPC ← RC
        else
            PPC ← EC
        end if
    end if
end if
return PPC

```

ADERR leverages the higher accuracy of EfficientNet while considering specific conditions where ResNet50 performs better. The weighted decision-making approach ensures that the final classification benefits from the strengths of both models, leading to improved overall accuracy. This data-driven approach, while heuristic, is validated by empirical results, showing superior performance in practice.

5.5. Post-processing Results

It is observed that, upon applying the post-processing techniques to the utilized models, the accuracy scores were maximized for both Dataset1 and Dataset2. Particularly, when the post-processing algorithm is applied to Dataset1, a prediction accuracy of 98.97% is achieved, whereas in Dataset2 a remarkable prediction accuracy of 99.41% is achieved.

Hence, Fig. 18 represents the post-processing confusion matrix for the Dataset1 test subset, detailing both correct and incorrect predictions, scoring a 98.97% accuracy.

		Predicted			
		Mild Demented	Moderate Demented	Non Demented	Very Mild Demented
Actual	Mild Demented	1785	0	4	3
	Moderate Demented	0	1294	0	0
	Non Demented	10	0	1888	22
	Very Mild Demented	9	0	22	1761

Fig. 18: Post-processing (with 98.97% accuracy) confusion matrix

The performance of post-processing for each dementia classification in Dataset 1 (OASIS) is shown in Fig.18, and the computed accuracy, precision, recall, and F1-score percentages are as follows:

Metrics for each individual class:

- MildDemented: Due to a few cases that were incorrectly classified, the accuracy was high (99.62%) with a good recall (99.61%), but the precision was slightly lower (98.95%) (26).
- ModerateDemented: 100% performance on all metrics (recall, accuracy, precision, and F1-score) demonstrated perfect classification.
- NonDemented showed good accuracy (99.15%) with a moderate decline in precision compared to MildDemented (98.64%) and recall (98.33%). A small number of NonDemented cases (32) and misclassified instances (58) were overlooked.
- VeryMildDemented: Showed excellent accuracy (99.18%) and comparable trends to NonDemented in terms of recall (98.27%) and precision (98.60%). Still, there were a few incorrect classifications (56).

Metrics for Comparing Classes:

- MildDemented vs. ModerateDemented and ModerateDemented vs. NonDemented and ModerateDemented vs. VeryMildDemented: Achieved perfect score of 100% for all metrics (accuracy, precision, recall, F1-score).
- MildDemented vs. NonDemented: demonstrated near-perfect precision (99.78%) and very good accuracy (99.62%), recall: 99.44%, and F1-score: 99.61%, with a small number of NonDemented cases (10).
- MildDemented vs. VeryMildDemented: Demonstrated high accuracy (99.66%) with excellent precision (99.83%), recall (99.50%), and F1-score (99.66%) but also a small number of missed VeryMildDemented cases (9).
- NonDemented vs. VeryMildDemented: Good scoring is indicated by high accuracy (98.81%) with comparable precision (98.85%), recall (98.85%), and F1-score: 98.85%. Although with some misclassifications (44).

Furthermore, we can say that our model, which scored an accuracy of 98.97% on dataset1 (OASIS) outperformed another ensemble learning used by [15] which scored a test accuracy of 91.96% for The ANN, 85.7% for the gradient boosting and 83.04% for the ensemble learning voting classifier methods.

In conclusion, post-processing appears to have a beneficial effect, potentially improving the model’s performance by correcting some misclassifications across all categories.

Example quantitative results are depicted in Fig. 19, showcasing the accuracy score of the applied approach after the post-processing, which is 98.97% based on the evaluation parameters.

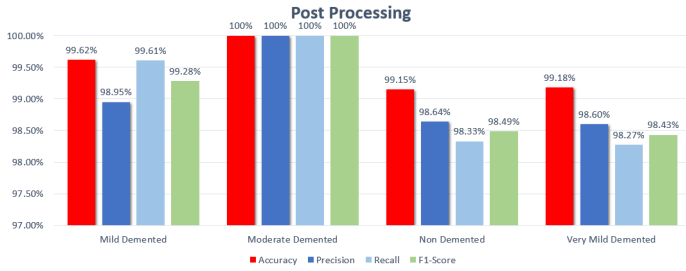


Fig. 19: Post-processing (with 98.97% accuracy) results based on the evaluation parameters

Similarly the confusion matrix for the Dataset2 test subset, detailing both correct and incorrect predictions, scoring a 99.41% accuracy is illustrated in Fig. 20.

		Predicted			
		Mild Demented	Moderate Demented	Non Demented	Very Mild Demented
Actual	Mild Demented	882	0	0	14
	Moderate Demented	0	64	0	0
	Non Demented	1	0	3191	8
	Very Mild Demented	0	0	15	2225

Fig. 20: Post-processing (with 99.41% accuracy) confusion matrix

The performance of post-processing for each dementia classification in Dataset 2 (ADNI) is shown in Fig.20, and the computed accuracy, precision, recall, and F1-score percentages are as follows:

Metrics for each individual class:

- **MildDemented:** Achieved high precision (99.89%), recall (98.44%), and F1-score (99.16%) with an accuracy of 99.77%. Nevertheless, 14 cases (0.23%) had incorrect classifications.
- **ModerateDemented:** Exceptionally well-performed, with no misclassified examples and 100% accuracy across all parameters (precision, recall, and F1-score).
- **NonDemented:** Accuracy was 99.63%, recall, precision, and F1-score were all approximately 99.53%. In 24 cases (0.75%) the classification was incorrect.

- **VeryMildDemented:** Maintained high precision (99.02%), recall (99.33%), and F1-score (99.18%) while demonstrating an accuracy of 99.42%. However, 37 cases (1.67%) were misclassified.

Metrics for Comparing Classes:

- **MildDemented vs ModerateDemented and ModerateDemented vs NonDemented and ModerateDemented vs VeryMildDemented:** scored a perfect 100% on all metrics (accuracy, precision, recall, F1-score).
- **MildDemented vs NonDemented:** Showed accuracy of 99.98% and only misclassified 1 NonDemented instance scoring 100% precision, 99.89% recall, and 99.94% F1-score.
- **MildDemented vs VeryMildDemented:** demonstrated a marginal decline in precision to 98.44% while keeping a high F1-score of 99.21% and a high recall of 100%, indicating that some cases of mild dementia may be mistaken for veryMildDemented.
- **NonDemented vs VeryMildDemented:** Scored an accuracy of 99.58% with high precision of 99.75%, a recall of 99.53%, and F1-score of 99.64%.

[22] previously investigated the accuracy of an Adaptive DBN model based on Teacher-Student interaction on the ADNI dataset for comparable classification tasks. Notably, our approach achieved superior performance across all categories on Dataset 2 (ADNI), which was used as a blind test set for our post-processing method (class labels were hidden). With respect to AD vs. CN, [22]’s model obtained an accuracy of 98.4%, but our post-processing method obtained 100% for ModerateDemented vs. NonDemented. Similarly, their model achieved 98.8% for MCI vs. CN, while ours achieved remarkable accuracy of 99.98% and 99.58%, respectively in differentiating between NonDemented vs. VeryMildDemented and MildDemented vs. NonDemented. Ultimately, our post-processing method achieved a flawless 100% accuracy for both MildDemented vs. ModerateDemented and ModerateDemented vs. VeryMildDemented, while their model scored 97.8% for the difficult MCI vs. AD classification.

Furthermore, [18] previously investigated the accuracy of an ensemble approach on the ADNI dataset for comparable classification tasks. Notably, our approach achieved superior performance across all categories on Dataset 2 (ADNI), which was used as a blind test set for our post-processing method (class labels were hidden). With respect to AD vs. NC, [18]’s model obtained an accuracy of 98.57%, but our post-processing method obtained 100% for ModerateDemented vs. NonDemented. Similarly, their model achieved 96.37% for NC vs. EMCI, while ours achieved remarkable accuracy of 99.98% and 99.58%, respectively in differentiating between NonDemented vs. VeryMildDemented and MildDemented vs. NonDemented. Ultimately, our postprocessing method achieved a flawless 100% accuracy for both MildDemented vs. ModerateDemented and ModerateDemented vs. VeryMildDemented,

while their model scored 94.22% and 99.83% for both EMCI vs LMCI and LMCI vs AD classification.

Finally, we can say that our model, which scored an accuracy of 99.41% on dataset2 (ADNI) outperformed [27] that also used an ensemble learning approach and scored an accuracy of 89.77% before optimization of parameters and an accuracy to 95.75 % after the optimization of parameters.

This direct comparison using the held-out test set (ADNI) highlights how successful our post-processing technique is. Through the integration of data-driven improvements and the utilization of the advantages of both models, our method provides better classification accuracy when discerning between different phases of cognitive decline.

Example quantitative results are depicted in Fig. 21, showcasing the accuracy score of the applied approach after the post-processing, which is 99.41% based on the evaluation parameters.

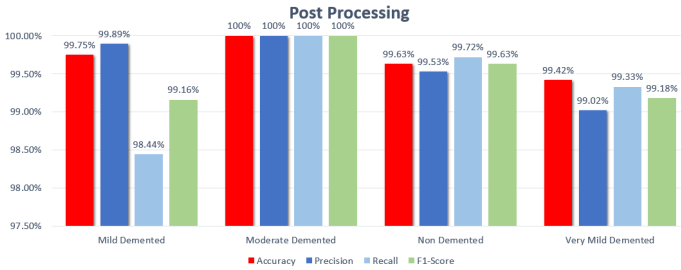


Fig. 21: Post-processing (with 99.41% accuracy) results based on the evaluation parameters

The empirical results, summarized in our paper, demonstrate that ADERR achieves higher accuracy, precision, recall, and F1-score across almost all classes compared to using EfficientNet or ResNet individually. The decision criteria are designed to select the prediction with higher confidence and reliability, reducing the chance of misclassification. These results are obtained through rigorous validation and testing, providing strong evidence for the efficacy of our approach.

5.6. Model performance analysis

In this section the effectiveness and efficiency of the used classification model is discussed.

5.6.1. Comparison of different CNN architectures

The applied models outperform competing models on both datasets. Remarkably, post-processing performed better on Dataset1 (OASIS) and Dataset2 (ADNI) than both EfficientNet and ResNet. EfficientNet and ResNet both obtained 98.59% and 94.59% accuracy in Dataset 1, respectively; however, post-processing outperformed both with a 98.97% accuracy. In Dataset2, EfficientNet and ResNet both obtained 97.25% and 99.36% accuracy, respectively; however, post-processing performed exceptionally well, with 99.41% accuracy. These findings demonstrate the effectiveness of the applied models and

the noteworthy improvement attained by post-processing methods, especially in Dataset1 where the improvement was most pronounced. The increase in performance seen in both datasets indicates the resilience and flexibility of the applied method, suggesting it as a viable option for precise classification tasks in many contexts. Through thorough assessment and comparison, the utilized models exhibit their capacity to attain elevated accuracy rates, providing significant understanding into the potential applications of sophisticated post-processing techniques to improve model performance on a variety of datasets.

5.6.2. Run-time Performance

It is observed that EfficientNet took between 65ms and 150ms per image, while ResNet took between 75ms and 200ms. Consequently, it is inferred that the performance time of the post-processing algorithm, which combines both EfficientNet and ResNet, will be equal to the sum of the individual algorithms' performance times. It is noteworthy that the conditions implemented in the post-processing step take negligible time, and therefore were not included in the performance time calculation of the post-processing algorithm.

6. Discussion

This paper's main contributions are to the discussion of the difficulties associated with Alzheimer's disease and the support of ongoing research efforts to create practical solutions. The proposed work advances the current understanding of ensemble learning techniques in medical image classification by introducing a specialized approach tailored to Alzheimer's disease diagnosis. By integrating EfficientNet and ResNet models and defining a set of decision rules, we provide a systematic framework for combining their predictions effectively.

Importantly, this research extends beyond mere model integration by carefully considering the strengths and weaknesses of each model and incorporating domain-specific knowledge into the decision-making process. For instance, as depicted in Figure 14, EfficientNet achieves a remarkable accuracy of 98.59%, outperforming ResNet, which achieves 94.59% accuracy. Moreover, EfficientNet attains a perfect accuracy score of 100% for the "ModerateDemented" class, correctly identifying all instances without misclassifying other classes. Additionally, specific probability thresholds and class-based conditions are employed, as indicated in Figures 18,19,20 and 21, guiding the ensemble model's decision-making process, resulting in more accurate and reliable predictions.

This work proposes a novel deep learning model that combines an advanced post-processing ensemble learning algorithm using weighted averaging and majority vote techniques with two state-of-the-art CNN algorithms: ResNet and EfficientNet. The ensemble learning approach using existing models, combined with our novel decision-making algorithm, provides a robust method for achieving high numerical accuracy. By diagnosing AD with high accuracy, this work could have a substantial impact on clinical practice and enhance patient outcomes.

The method's robustness and generalizability are guaranteed by a thorough evaluation on two different datasets drawn from

Model	Dataset1 (OASIS)	Dataset2 (ADNI)	Improvement		
			ResNet50	EfficientNetB0	Post Processing
Hypergraph [14]	-	92.11%	+7.25%	+5.14%	+7.3%
PCANet+k-means [16]	-	97.01%	+2.35%	+0.24%	+2.4%
DTI,SVM [17]	-	95.8%	+3.56%	+1.45%	+3.61%
WPBEM [18]	-	88.46%	+10.9%	+8.79%	+10.95%
CAD CNN [8]	-	96%	+3.36%	+1.25%	+3.41%
2D CNN [20]	-	93.61%	+5.75%	+3.64%	+5.8%
3D CNN [20]	-	95.17%	+4.2%	+2.08%	+4.24%
VGG19 [20]	-	97%	+2.36%	+0.25%	+2.41%
DTI,SVM [21]	-	97%	+2.36%	+0.25%	+2.41%
Adaptive DBN model based on Teacher-Student interaction [22]	-	96.7%	+2.66%	+0.55%	+2.71%
Swin Trans.,DCPAN, ADF [45]	-	79.8%	+19.56%	+17.45%	+19.61%
EfficientNetB0 [46]	-	77.2%	+22.16%	+20.05%	+22.21%
AHANet [47]	-	98.53%	+0.83%	-1.28%	+0.88%
SVM, RF, DT, XGBoost [25]	86.92%	-	+7.67%	+11.67%	+12.05%
SVM, RF, DT, XGBoost [39]	-	90.6%	+8.76%	+6.65%	+8.81%
DenseNet-169 [43]	-	82.2%	+17.16%	+15.05%	+17.21%
GNB, RF, DT, XGBoost [48]	96%	-	-1.41%	+2.59%	+2.97%
SVM, RF, DT, XGBoost [50]	93%	-	+1.59%	+5.59%	+5.97%
3D CNN, ResNet [54]	-	94.61%	+4.75%	+2.64%	+4.8%
Ensemble Learning Before Optimization [27]	-	89.77%	+9.59%	+7.48%	+9.64%
Ensemble Learning After Optimization [27]	-	95.75%	+3.61%	+1.5%	+3.66%
Gradient Boosting [15]	85.70%	-	+8.89%	+12.89%	+13.27%
Ensemble Learning Voting Classifier [15]	83.04%	-	+11.55%	+15.55%	+15.93%
Artificial Neural Network model (ANN) [15]	91.96%	-	+2.63%	+6.63%	+7.01%
ResNet50 (ours)	94.59%	99.36%	-	-	-
EfficientNetB0 (ours)	98.59%	97.25%	-	-	-
Post Processing (ours)	98.97%	99.41%	-	-	-

Table 2: Summary of accuracies across Dataset1 (OASIS), Dataset2 (ADNI) and various DL models

OASIS and ADNI. With an accuracy of 98.59% for EfficientNet, 94.59% for ResNet, and 98.97% for the post-processing method on the first dataset, the analysis yields outstanding findings. Similarly, on the second dataset, the post-processing approach achieves 99.41% accuracy, ResNet achieves 99.36% accuracy, and EfficientNet achieves 97.25% accuracy.

Furthermore, as indicated by the results, post-processing exhibited superior performance compared to both EfficientNet and ResNet across most evaluation metrics for all classes. This suggests that utilizing post-processing with both EfficientNet and ResNet yields better results than using either model individually. The mathematical proof and empirical validation support the claim that our method is both effective and reliable, even in the presence of potential adversarial cases.

The significance of the applied method in the fields of medicine and biology stems from its capacity to enhance classification accuracy and diminish the necessity for manual annotation. This capability enables more streamlined and precise analysis of extensive medical image datasets, thereby facilitating improved efficiency and accuracy in research and diagnosis. In conclusion, this work significantly advances the science of Alzheimer’s disease detection by presenting a new deep learning model with higher performance across several datasets, combining state-of-the-art CNN algorithms and post-processing methods.

Table 2 shows a comparison of the accuracy and performance gains made by different deep learning (DL) models on distinct datasets. For the two different datasets designated as Dataset1 and Dataset2, each row represents a particular DL model and displays the accuracy percentage and performance improvement percentage that go along with it. The accuracy attained by each DL model is shown in the "Accuracy" column, represented as a percentage. The next columns, "Improvement (ResNet50)" and "Improvement (EfficientNetB0)," respectively, show the percentage increase in model performance with respect to Dataset1 and Dataset2, in comparison to the baseline or earlier findings. Interestingly, both datasets show notable improvements from the cited DL models, demonstrating their effectiveness in improving classification or prediction tasks. A commentary that provides context for the post-processing results and highlights the overall high accuracy for Datasets 1 (98.97%) and 2 (99.41%) complements the table. This comparison highlights the potential of DL models to improve data-driven decision-making processes by providing insightful information on the developments and efficacy of DL models across a range of application domains.

7. Summary and conclusions

In conclusion, this research presents a significant advancement in Alzheimer's disease detection by introducing a novel deep learning model that combines two state-of-the-art CNN algorithms, ResNet and EfficientNet, with an advanced post-processing ensemble learning method in a weighted manner. This approach demonstrates efficacy in correctly detecting AD, achieving remarkable accuracies on two different datasets from OASIS and ADNI. EfficientNet achieved 98.59% accuracy, ResNet achieved 94.59% accuracy, and the post-processing method on the first dataset achieved an astounding 98.97% accuracy. Similarly, EfficientNet produced accuracies of 97.25% on the second dataset, ResNet achieved accuracies of 99.36%, and the post-processing technique achieved an exceptional 99.41% accuracy.

These outcomes highlight the method's stability and dependability in correctly detecting cases of AD. Additionally, the study demonstrates that EfficientNet performs better than ResNet, and combining both models with the post-processing technique greatly improves performance accuracy, as seen by the remarkable 99.41% score. Combining cutting-edge CNN algorithms with sophisticated post-processing methods represents a significant advancement in the identification of AD, with the potential to enhance patient outcomes and diagnostic precision.

Looking ahead, this discovery paves the way for improvements in the diagnosis of AD. Further validation of the solution across a variety of datasets and optimization of post-processing algorithms are planned to attain even greater levels of accuracy, precision, recall, and F1-Score metrics. These efforts are expected to make meaningful contributions to the field of AD research and clinical treatment, by pushing the frontiers of deep learning and utilizing cutting-edge approaches.

In summary, this research constitutes a noteworthy advancement in the pursuit of more precise and dependable techniques for diagnosing AD. Through the utilization of advanced post-processing techniques and deep learning, this research has the potential to significantly improve the lives of those afflicted with this disabling illness.

Declaration of competing interest

The authors report there are no competing interests to declare.

Acknowledgements

Thanks to everyone who helped making this paper possible.

References

- [1] Alzheimer MRI Preprocessed Dataset. <https://www.kaggle.com/datasets/sachinkumar413/alzheimer-mri-dataset>. Accessed: 2024-03-03.
- [2] The Annotated EfficientNet-B0. <https://iq.opengenus.org/efficientnet/>. Accessed: 2024-04-23.
- [3] The Annotated ResNet-50. <https://towardsdatascience.com/the-annotated-resnet-50-a6c536034758>. Accessed: 2024-04-23.
- [4] Augmented Alzheimer MRI Dataset. <https://www.kaggle.com/datasets/uraninjo/augmented-alzheimer-mri-dataset>. Accessed: 2024-03-03.
- [5] PyTorch ResNet: The basics and a quick tutorial. <https://www.run.ai/guides/deep-learning-for-computer-vision/pytorch-resnet>. Accessed: 2024-03-03.
- [6] What is deep learning? — How It Works, Techniques Applications. <https://www.mathworks.com/discovery/deep-learning.html>. Accessed: 2024-03-03.
- [7] What is supervised learning? — IBM. <https://www.ibm.com/topics/supervised-learning>. Accessed: 2024-03-03.
- [8] Ahila A, Mounir Hamdi, Sami Bourouis, Kulhanek Rastislav, and Faizaan Mohmed. Evaluation of neuro images for the diagnosis of alzheimer's disease using deep learning neural network. *Frontiers in Public Health*, 10:834032, 2022.
- [9] Michele Alessandrini, Giorgio Biagetti, Paolo Crippa, Laura Falaschetti, Simona Luzzi, and Claudio Turchetti. EEG-based Alzheimer's disease recognition using robust-PCA and LSTM recurrent neural network. *Sensors*, 22(10):3696, 2022.
- [10] Emre Altinkaya, Kemal Polat, and Burhan Barakli. Detection of Alzheimer's Disease and Dementia States Based on Deep Learning from MRI images: A Comprehensive Review. *Journal of the Institute of Electronics and Computer*, 1(1):39–53, 2020.
- [11] Ning An, Huitong Ding, Jiaoyun Yang, Rhoda Au, and Ting F.A. Ang. Deep ensemble learning for Alzheimer's disease classification. *Journal of Biomedical Informatics*, 105:103411, 2020.
- [12] Anza Aqeel, Ali Hassan, Muhammad Attique Khan, Saad Rehman, Usman Tariq, Seifedine Kadry, Arnab Majumdar, and Orawit Thinnukool. A long short-term memory biomarker-based prediction framework for Alzheimer's disease. *Sensors*, 22(4):1475, 2022.
- [13] Alzheimer Association. Alzheimer's disease facts and figures, 2023.
- [14] Angelica I Aviles-Rivero, Christina Runkel, Nicolas Papadakis, Zoe Kourtzi, and Carola-Bibiane Schönlieb. Multi-modal hypergraph diffusion network with dual prior for Alzheimer classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 717–727. Springer, 2022.
- [15] Ahana Bandyopadhyay, Sourodir Ghosh, Moinak Bose, Arun Singh, Alice Othmani, and KC Santosh. Alzheimer's Disease Detection Using Ensemble Learning and Artificial Neural Networks. pages 12–21, 2023.
- [16] Xiuli Bi, Shutong Li, Bin Xiao, Yu Li, Guoyin Wang, and Xu Ma. Computer aided Alzheimer's disease diagnosis by an unsupervised deep learning technology. *Neurocomputing*, 392:296–304, 2020.
- [17] Bahare Bigham, Seyed Amir Zamanpour, and Hoda Zare. Features of the superficial white matter as biomarkers for the detection of Alzheimer's disease and mild cognitive impairment: A diffusion tensor imaging study. *Heliyon*, 8(1), 2022.
- [18] Sina Fathi, Ali Ahmadi, Afsaneh Dehnad, Mostafa Almasi-Dooghaee, and Melika Sadegh. A Deep Learning-Based Ensemble Method for Early Diagnosis of Alzheimer's Disease using MRI Images. *Neuroinformatics*, 22:89–105, 2023.
- [19] Taher M Ghazal, Sagheer Abbas, Sundus Munir, M Adnan Khan, Munir Ahmad, Ghassan F Issa, Syeda Binish Zahra, Muhammad Adnan Khan, and Mohammad Kamrul Hasan. Alzheimer Disease Detection Empowered with Transfer Learning. *Computers, Materials & Continua*, 70(3), 2022.
- [20] Hadeer A Helaly, Mahmoud Badawy, and Amira Y Haikal. Deep Learning Approach for Early Detection of Alzheimer's disease. *Cognitive computation*, 14:1711–1727, 2022.
- [21] Latifa Houria, Nouredine Belkhamza, Assia Cherfa, and Yazid Cherfa. Multi-modality MRI for Alzheimer's disease detection using deep learning. *Physical and Engineering Sciences in Medicine*, 45(4):1043–1053, 2022.
- [22] Takumi Ichimura, Shin Kamada, Toshihide Harada, and Ken Inoue. A teacher-student-based adaptive structural deep learning model and its estimating uncertainty of image data. *Artificial Intelligence*, 49:129, 2023.
- [23] Muhammad Tausif Irshad, Muhammad Adeel Nisar, Xinyu Huang, Jana Hartz, Olaf Flak, Frédéric Li, Philip Gouverneur, Artur Piet, Kerstin M. Oltmanns, and Marcin Grzegorzec. SenseHunger: Machine Learning Approach to Hunger Detection Using Wearable Sensors. *Sensors*, 22:7711,

- 2022.
- [24] Wenjie Kang, Lan Lin, Shen Sun, and Shuicai Wu. Three-round learning strategy based on 3D deep convolutional GANs for Alzheimer's disease staging. *Sci Rep*, 13:5750, 2023.
 - [25] C Kavitha, Vinodhini Mani, SR Srividhya, Osamah Ibrahim Khalaf, and Carlos Andrés Tavera Romero. Early-Stage Alzheimer's Disease Prediction using Machine Learning Models. *Frontiers in public health*, 10:853294, 2022.
 - [26] Md Saikat Islam Khan, Anichur Rahman, Tanoy Debnath, Md Razaul Karim, Mostofa Kamal Nasir, Shahab S Band, Amir Mosavi, and Iman Dehzangi. Accurate brain tumor detection using deep convolutional neural network. *Computational and Structural Biotechnology Journal*, 20:4733–4745, 2022.
 - [27] Yusera Farooq Khan, Baijnath Kaushik, Chiranjee Lal Chowdhary, and Gautam Srivastava. Ensemble Model for Diagnostic Classification of Alzheimer's Disease Based on Brain Anatomical Magnetic Resonance Imaging. *Diagnostics (Basel)*, 12:3193, 2022.
 - [28] Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *ICLR: arXiv:1412.6980v9*, 2014.
 - [29] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.
 - [30] Rongjian Li, Wenlu Zhang, Heung-II Suk, Li Wang, Jiang Li, Dinggang Shen, and Shuiwang Ji. Deep learning based imaging data completion for improved brain disease diagnosis. *Medical Image Computing and Computer-Assisted Intervention—MICCAI*, 17(Pt 3):305–312, 2014.
 - [31] Zhihua Liu, Lei Tong, Long Chen, Zheheng Jiang, Feixiang Zhou, Qianni Zhang, Xiangrong Zhang, Yaochu Jin, and Huiyu Zhou. Deep learning based brain tumor segmentation: A survey. *Complex & intelligent systems*, 9(1):1001–1026, 2023.
 - [32] Md Ishtyaq Mahmud, Muntasir Mamun, and Ahmed Abdelgawad. A Deep Analysis of Brain Tumor Detection from MR Images using Deep Learning Networks. *Algorithms*, 16(4):176, 2023.
 - [33] Ahmad Naeem, Tayyaba Anees, Rizwan Ali Naqvi, and Woong-Kee Loh. A Comprehensive Analysis of Recent Deep and Federated-Learning-Based Methodologies for Brain Tumor Diagnosis. *Journal of Personalized Medicine*, 12(2):275, 2022.
 - [34] Justin S Paul, Andrew J Plassard, Bennett A Landman, and Daniel Fabbrì. Deep learning for brain tumor classification. *Medical Imaging: Biomedical Applications in Molecular, Structural, and Functional Imaging*, 10137:253–268, 2017.
 - [35] Angela Rizk-Jackson, Diederick Stoffers, Sarah Sheldon, Josh Kuperman, Anders Dale, Jody Goldstein, Jody Corey-Bloom, Russell A Pol-drack, and Adam R Aron. Evaluating imaging biomarkers for neurodegeneration in pre-symptomatic huntington's disease using machine learning techniques. *Neuroimage*, 56(2):788–796, 2011.
 - [36] Snehashis Roy, Andrew Knutsen, Alexandru Korotcov, Asamoah Bosomtvi, Bernard Dardzinski, John A Butman, and Dzung L Pham. A deep learning framework for brain extraction in humans and animals with traumatic brain injury. *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 687–691, 2018.
 - [37] Rabia Sajjad, Muhammad Faheem Khan, Asif Nawaz, Malik Taimur Ali, and Muhammad Adil. Systematic Analysis of Ovarian Cancer Empowered with Machine and Deep Learning: A Taxonomy and Future Challenges. *Journal of Computing & Biomedical Informatics*, 3(2):64–87, 2022.
 - [38] Muhammad Shafiq and Zhaoquan Gu. Deep Residual Learning for Image Recognition: A Survey. *Applied Sciences*, 12(18):8972, 2022.
 - [39] Rajesh Kumar Shrivastava, Simar Preet Singh, and Gagandeep Kaur. Machine Learning Models for Alzheimer's Disease Detection Using OASIS Data. *Data Analysis for Neurodegenerative Disorders*, pages 111–126, 2023.
 - [40] Amar Shukla, Rajeev Tiwari, and Shamik Tiwari. Review on Alzheimer Disease Detection Methods: Automatic Pipelines and Machine Learning Techniques. *Sci*, 5(1):13, 2023.
 - [41] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *ICLR: arXiv:1409.1556v6*, 2015.
 - [42] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *ICML: arXiv:1905.11946v5*, 2020.
 - [43] Illakiya Thayumanasamy and Karthik Ramamurthy. Performance Analysis of Machine Learning and Deep Learning Models for Classification of Alzheimer's Disease from Brain MRI. *Traitement du Signal*, 39:1961–1970, 2022.
 - [44] Illakiya Thayumanasamy and Karthik Ramamurthy. Automatic Detection of Alzheimer's Disease using Deep Learning Models and Neuro-Imaging: Current Trends and Future Perspectives. *Neuroinformatics*, 21(2):339–364, 2023.
 - [45] Illakiya Thayumanasamy and Karthik Ramamurthy. A Dimension Centric Proximate Attention Network and Swin Transformer for Age-Based Classification of Mild Cognitive Impairment From Brain MRI. *IEEE Access*, 11:128018–128031, 2023.
 - [46] Illakiya Thayumanasamy and Karthik Ramamurthy. A deep feature fusion network with global context and cross-dimensional dependencies for classification of mild cognitive impairment from brain MRI. *Image and Vision Computing*, 144:104967, 2024.
 - [47] Illakiya Thayumanasamy, Karthik Ramamurthy, MV Siddharth, Rashmi Mishra, and Ashish Udainiya. AHANet: Adaptive Hybrid Attention Network for Alzheimer's Disease Classification Using Brain Magnetic Resonance Imaging. *Bioengineering*, 10(6):714, 2023.
 - [48] Khandaker Mohammad Mohi Uddin, Mir Jafikul Alam, Jannat-E-Anwar, Md Ashraf Uddin, and Sunil Aryal. A Novel Approach Utilizing Machine Learning for the Early Diagnosis of Alzheimer's Disease. *Biomedical Materials & Devices*, 10:1–17, 2023.
 - [49] Hannelore K van der Burgh, Ruben Schmidt, Henk-Jan Westeneng, Marcel A de Reus, Leonard H van den Berg, and Martijn P van den Heuvel. Deep learning predictions of survival based on MRI in amyotrophic lateral sclerosis. *Neuroimage Clin*, 13:361–369, 2017.
 - [50] Anuradha Vashishtha, Anuja Kumar Acharya, and Sujata Swain. A Comparative Study on Various Machine Learning Approaches for the Detection of Alzheimer Disease. *International Journal of Intelligent Systems and Applications in Engineering*, 10(3):294–304, 2022.
 - [51] S Venkatasubramanian, Jaiprakash Narain Dwivedi, S Raja, N Rajeswari, J Logeshwaran, Avvaru Praveen Kumar, and Ardashir Mohammadzadeh. Prediction of Alzheimer's Disease Using DHO-Based Pretrained CNN Model. *Mathematical Problems in Engineering*, 2023:1–11, 2023.
 - [52] Qin Yang, Xin Li, Xinyun Ding, Feiyang Xu, and Zhenhua Ling. Deep learning-based speech analysis for Alzheimer's disease detection: A literature review. *Alzheimer's Research & Therapy*, 14(1):186, 2022.
 - [53] Shisheng Zhang, Simon K Poon, Kenny Vuong, Alexandra Sneddon, and Clement T Loy. A Deep Learning-Based Approach for Gait Analysis in Huntington Disease. *Stud Health Technol Inform*, 264:477–481, 2019.
 - [54] Yanteng Zhang, Xiaohai He, Yixin Liu, Charlene Zhi Lin Ong, Yan Liu, and Qizhi Teng. An end-to-end multimodal 3D CNN framework with multi-level features for the prediction of mild cognitive impairment. *Knowledge-Based Systems*, 281:111064, 2023.
 - [55] Zhen Zhang, Zongren Zou, Ellen Kuhl, and George Em Karniadakis. Discovering a reaction-diffusion model for Alzheimer's disease by combining PINNs with symbolic regression. *Computer Methods in Applied Mechanics and Engineering*, 419:116647, 2024.