Deep Learning-Based Multi-Classification of Alzheimer's Disease Stages Using Fine-Tuned VGG19 and MLP on MRI Scans

Siham Amrouch[*1,2], Ryma Guefrouchi[3,4], Islam Chabbi[1,2].
[1]Computer Science department, Mohamed Cherif Messaadia University, Souk Ahras, Algeria
[2]LIM Laboratory, Faculty of Sciences and Technology, Mohamed Cherif Messaadia University, Souk Ahras, Algeria
[3]Abdelhamid Mehri Constantine2 University, Constantine, Algeria
[4]MISC Laboratory, Abdelhamid Mehri Constantine2 University, Constantine, Algeria
E-mail: s.amrouche@univ-soukahras.dz, ryma.guefrouchi@univ-constantine2.dz, i.chabbi@univ-soukahras.dz.
[*]Corresponding author

Keywords: alzheimer's disease, image analysis, deep learning, fine-tuning, MLP, VGG19, MRI.

Received: Februar 3, 2025

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder and a leading cause of dementia worldwide, making early and accurate diagnosis essential for effective clinical intervention. Although recent Deep Learning (DL) approaches have shown promising results for automated AD diagnosis from brain Magnetic Resonance Imaging (MRI), many existing methods remain limited by insufficient domain adaptation of pretrained models, severe class imbalance in clinical datasets, and suboptimal classification performance, which restrict their generalisability and clinical reliability. To address these challenges, this paper proposes a robust DL framework for multi-class classification of AD stages from structural MRI scans. The proposed framework is built upon a fine-tuned VGG19 model, enabling end-to-end learning and effective domain adaptation from natural images to medical imaging data, followed by a Multi-Layer Perceptron (MLP) for accurate stage-level classification. To mitigate the inherent class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) is applied at the image level to balance class distributions during training. The framework is evaluated on OASIS dataset, publicly available via Kaggle, comprising 6,400 MRI images categorised into four classes. All MRI scans undergo a standardised preprocessing pipeline including image resizing, intensity normalisation, grayscale-to-RGB conversion, and training data augmentation. The dataset is split into training (80%) and testing (20%) sets using a stratified strategy. In addition, the proposed model is independently validated on ADNI dataset to assess robustness and cross-dataset generalisability. Experimental results demonstrate the effectiveness of the proposed framework, achieving an accuracy of 98.83% on OASIS and 99.13% on ADNI, along with high precision, recall, F1-score, and AUC, outperforming state-of-the-art methods.

Povzetek: V tem članku združujemo VGG19 kot izvleček značilk z MPL kot klasifikatorjem za zgodnje odkrivanje Alzheimerjeve bolezni.

1 Introduction

Dementia is a progressive neurological disorder characterised by a gradual decline in cognitive abilities [1,2]. AD is the most prevalent form of dementia, accounting for approximately 60–80% of all dementia cases worldwide. Alarmingly, one person develops dementia every three seconds, with AD constituting nearly 60% of these cases [3].

AD is a chronic neurodegenerative disease that progressively damages brain cells, leading to impairments in memory, reasoning, and other cognitive functions. As the disease advances, patients gradually lose the ability to perform even basic daily activities independently [4]. In its early stages, AD primarily affects brain regions responsible for memory and language, resulting in difficulties in writing, reading, and speech. Memory loss is a hallmark symptom of AD [5], and affected individuals may struggle to recognise family members or recall significant life events [7]. Mental and physical health play a crucial role throughout the human life cycle, and older adults are particularly vulnerable to diseases such as cancer, cardiovascular disorders, arthritis, diabetes, pneumonia, depression, kidney disease, dementia, and AD [6].

The global population is projected to reach 11.2 billion by 2100, representing an increase of nearly 50%. By the middle of the twenty-first century, it is estimated that approximately two billion elderly individuals will be living worldwide. According to the World Alzheimer's Association, nearly seven million people aged 65 years and older in the United States are currently affected by AD, and this number is expected to rise significantly. By 2050, approximately 640 million individuals worldwide are projected to be affected by AD [8]. To identify Mild Cognitive Impairment (MCI) and diagnose AD, clinicians commonly rely on brain MRI, which provides detailed structural information about the brain [9].

Image classification is a fundamental task in image analysis, computer vision, and pattern recognition. It involves assigning one or more labels to an image by extracting low-, mid-, and high-level features and training a classifier using labelled datasets. In the context of medical imaging, accurate image classification plays a critical role in disease detection and diagnosis.

Early diagnosis of AD offers several clinical benefits, including slowing disease progression, reducing healthcare costs, and improving treatment outcomes. Early intervention has the potential to decrease mortality rates and reduce the frequency of medical emergencies associated with advanced stages of the disease [11,12]. Recent clinical research failures in AD treatment further emphasise the importance of early and accurate diagnosis.

In recent years, numerous Machine Learning (ML) and Deep Learning (DL) techniques have been proposed and successfully applied across various domains, including healthcare, computer vision, and pattern recognition. These techniques have been extensively used to analyse and classify MRI scans for early detection and diagnosis of AD [13,14]. Among them, convolutional neural networks (CNNs) have demonstrated remarkable performance due to their ability to automatically extract high-level and discriminative features from imaging data, outperforming traditional feature extraction methods. A key advantage of CNNs lies in their end-to-end learning capability, where feature extraction and classification are jointly optimised.

Several pre-trained CNN architectures have shown strong performance in predicting AD stages from MRI scans, including AlexNet, ResNet, VGG variants, SqueezeNet, DenseNet, and Inception models. These architectures effectively capture structural brain patterns that are crucial for distinguishing between different stages of AD.

In this work, we propose a robust DL framework based on a fine-tuned VGG19 model, combined with an MLP, for multi-class classification of AD stages from brain MRI scans. Unlike approaches that rely on fixed feature extraction, the proposed framework employs end-to-end fine-tuning of VGG19, enabling effective domain adaptation and learning of highly discriminative MRI-specific representations.

To address the severe class imbalance commonly observed in clinical datasets, the SMOTE technique is applied at the image level to balance class distributions and improve generalisation [15]. The primary evaluation is conducted on OASIS dataset, publicly available via Kaggle platform which includes four AD stages: Non-Demented (ND), Very Mild Demented (VMD), Mild Demented (MD), and Moderate Demented (MOD). Furthermore, the proposed framework is independently validated on ADNI dataset to assess robustness and cross-dataset generalisability. Experimental results demonstrate highly competitive performance, highlighting the strong potential of the proposed approach to support accurate diagnosis and early intervention in AD. Comparative analyses with state-of-the-art methods consistently confirm the superiority of the proposed framework.

The main contributions of this work are summarised as follows:

- Proposing a fine-tuned VGG19-based deep learning framework for multi-class classification of AD stages from MRI scans.
- Addressing class imbalance through the application of SMOTE to improve model generalisation and reduce bias toward majority classes.
- Integrating an MLP for effective stage-level prediction based on deep features learned via end-to-end optimisation.
- Conducting extensive experimental evaluations on two public datasets (OASIS and ADNI), demonstrating high accuracy and strong generalisability.
- Providing a comprehensive comparison with state-of-the-art approaches using multiple evaluation metrics.

The remainder of this paper is organised as follows. Section 2 reviews related work, Section 3 describes the proposed methodology, Section 4 presents experimental results and comparative analyses, and Section 5 concludes the paper and outlines future research directions.

2 Related work

In recent years, DL combined with ML has shown strong potential for AD diagnosis. Numerous studies have proposed DL-based models as effective clinical decision-support tools. This section reviews representative and closely related works, emphasising the growing importance of advanced learning techniques for accurate AD detection and staging.

Ghazal et al. [17] applied DL and transfer learning (TL) techniques for the early detection of AD. Their model, named ADDTLA, was evaluated on a multi-class MRI dataset representing different stages of AD and achieved an overall classification accuracy of 91.70%.

Revathi et al. [18] investigated the impact of blood pressure and diabetes on cognitive impairment using support vector machines (SVM) and random forest classifiers. They further employed multinomial logistic regression for early AD identification. The proposed approach was tested on the Cognitive Ability Test (CAT) dataset and the reported accuracies for SVM, random forest, and multinomial logistic regression were 71%, 86%, and 89%, respectively.

Khan et al. [19] proposed a three-tier hybrid classification framework for AD and MCI prediction using baseline characteristics and psychometric test data. The model integrates nineteen classifiers, including logistic regression, naïve Bayes, SVM, decision trees, random forest, and extreme gradient boosting, using a two-layer stacking strategy. Evaluated on ADNI dataset, the approach achieved accuracies of 90.24% and 95.12% for layer1 and layer2, respectively, demonstrating effective identification of early AD and MCI cases.

Pirrone et al. [20] introduced a new method based on finite impulse response (FIR) filters to extract features in the double time domain for distinguishing AD, MCI, and

HC. Power intensity differences in high- and low-frequency bands were computed and classified using SVM, k-nearest neighbour (KNN), and decision tree (DT) classifiers. EEG recordings from 105 subjects were used, achieving accuracies of 97%, 95%, and 83% for binary classifications, and 75% accuracy for multi-class classification.

Liu et al. [21] proposed a Multi-Modality Cascaded CNN (MC-CNN) to learn multi-level and multimodal features from MRI and PET images. The method employs deep 3D CNNs on local patches to extract compact representations, followed by a 2D CNN to model cross-modal correlations. Evaluated on ADNI dataset, the model achieved accuracies of 93.26% for AD vs. NC and 82.95% for MCI vs. NC classification, albeit at a high computational cost.

Lu et al. [22] introduced a Multiscale Deep Neural Network (MDNN) for early AD detection using FDG-PET images. The model was evaluated on ADNI dataset and achieved an overall classification accuracy of 82.51%. Although the approach demonstrated stable performance, it lacked the ability to integrate multimodal data.

Islam et al. [23] proposed an ensemble CNN (e-CNN) model for AD stage prediction and classification. The method was evaluated on OASIS dataset and achieved an accuracy of 92.48%, with a precision of 0.940, recall of 0.930, and F-measure of 0.920. While effective, the model showed limited adaptability to other AD datasets.

Marzban et al. [24] developed a cost-effective shallow CNN-based framework for classifying AD and MCI against healthy controls using ADNI dataset. The approach incorporated diffusion tensor imaging (DTI) and MRI-derived features and achieved an AUC of 94%, accuracy of 93.5%, sensitivity of 92.5%, and specificity of 93.9% using ten-fold cross-validation.

Deepa et al. [25] proposed an optimised VGG16 model using the Arithmetic Optimisation Algorithm (AOA) for AD classification. The framework integrates MRI preprocessing and segmentation via the CAT12 toolkit. Evaluated on the ADNI dataset, it achieved 92.34% accuracy, with 0.969 precision, 0.958 recall, 0.962 sensitivity, 0.967 specificity, and 0.957 F1-score.

Sharma et al. [26] combined a VGG16 feature extractor with a neural network classifier for early AD diagnosis. Experiments on two MRI datasets (6,400 and 6,330 images) achieved 90.4% accuracy, 0.969 AUC, 0.904 and F1-score on the first dataset, while accuracy dropped to 71.1% on the second, indicating dataset sensitivity.

Murugan et al. [27] introduced Dementia Network (DEMNET), a framework designed to extract AD-specific characteristics from MRI images while considering dementia stages. The model employed an MLP to generate high-resolution probability maps, providing interpretable visualisations of AD risk. DEMNET achieved an accuracy of 95.23%, AUC of 97%, and a Cohen's kappa value of 0.93 on OASIS dataset, and its effectiveness was further validated on ADNI dataset.

Table 1 provides a systematic comparison of the aforementioned studies on AD classification, summarising the datasets used, input modalities, classification types, model architectures, reported accuracies, and key limitations.

As shown in Table 1, most existing approaches employ pretrained networks without fine-tuning, neglect class imbalance issues, or exhibit high computational complexity. Moreover, several studies focus on binary classification tasks or rely on limited datasets, which restricts their applicability to real clinical scenarios involving multiple AD stages. These limitations collectively hinder generalisability and robustness. In contrast, the proposed framework addresses these gaps by performing fine-tuned multi-class classification on a large-scale public MRI dataset while explicitly handling class imbalance, thereby achieving superior performance with improved clinical relevance.

3 Materials and methods

The proposed framework is designed as an end-to-end DL system for multi-class classification of AD stages from structural MRI scans. It consists of three main stages: data collection and preprocessing, deep feature extraction and learning using a fine-tuned VGG19-based model, and classification using an MLP-based classifier. This section describes in detail the materials used and the methodological steps followed to construct and evaluate the proposed AD classification system.

3.1 Feature learning using fine-tuned VGG19

Feature learning is a fundamental step in medical image analysis and pattern recognition, as it enables the extraction of discriminative representations that enhance classification performance while reducing the complexity of raw image data [28, 29]. By transforming high-dimensional image inputs into meaningful feature representations, DL-based approaches facilitate efficient and accurate analysis of brain MRI scans. Several feature learning techniques have been proposed in the literature, including Principal Component Analysis, kernel-based methods, autoencoders, local feature descriptors, and deep CNNs, depending on the nature of the data and the target application [29, 30]. In this study, the VGG19 architecture is employed as the backbone of the proposed framework. VGG19 is a deep CNN originally developed for large-scale image classification, it consists of 19 layers, including 16 convolutional layers and 3 fully connected layers [31]. Its uniform architecture, based on small 3×3 convolutional filters and a consistent design pattern, makes it highly reproducible and effective for hierarchical feature learning. These characteristics justify its selection for MRI-based AD classification.

Table 1: Comparative analysis of existing approaches for Alzheimer's disease classification.

Ref	Architecture	Dataset	Input Modality	Classification Type	Accuracy (%)	Key Limitations
[17]	ADDTLA	Private MRI dataset	MRI	Multi-class	91.70	Limited reproducibility
[18]	SVM	CAT	MRI	Multi-class	71.00	Suboptimal performance
	Random Forest	CAT	MRI	Multi-class	86.00	Limited discriminative capacity
	Multinomial Logistic Regression	CAT	MRI	Multi-class	89.00	Classical machine learning limitations
[19]	Three-tier cognitive hybrid ML	ADNI	MRI	Multi-class	90.24 (HCM-1); 95.12 (HCM-2)	Highly complex architecture (19 ML models)
[20]	SVM, DT, KNN	EEG dataset	EEG	Multi-class	75.00	Suboptimal multi-class performance
	SVM, DT, KNN	EEG dataset	EEG	Binary	97.00 (HC vs. AD); 95.00 (HC vs. MCI); 83.00 (MCI vs. AD)	Binary-focused evaluation
[21]	Multi-Modality Cascaded CNN (MC-CNN)	ADNI	MRI + PET	Binary	93.26 (AD vs. NC); 82.95 (MCI vs. NC)	Binary tasks only; high computational cost
[22]	Multiscale MDNN	ADNI	MRI	Multi-class	82.51	Suboptimal classification accuracy
[23]	e-CNN	OASIS	MRI	Multi-class	92.48	No class imbalance handling + poor generalization
[24]	CNN	ADNI	MRI + DTI	Binary	93.50	Binary classification only
[25]	Optimised VGG16 with AOA	ADNI	MRI	Multi-class	92.34	No fine-tuning
[26]	Neural Network + VGG16	OASIS	MRI	Multi-class	90.40	No fine-tuning
[27]	CNN + MLP	OASIS	MRI	Multi-class	95.23	High computational complexity
[37]	Ensemble-based classifier	ADNI	MRI	Binary	85.55	Binary classification only; suboptimal performance
		NRCD	MRI	Binary	90.05	Binary classification only
[38]	MLP	ADNI	MRI + PET	Multi-class	89.00	No fine-tuning; suboptimal performance
	CBLSTM + GAIN	ADNI	MRI + PET	Multi-class	82.00	Limited performance
	CBLSTM + SMOTE	ADNI	MRI + PET	Multi-class	82.00	Limited performance

Unlike approaches that use VGG19 solely as a fixed feature extractor, the proposed framework adopts a fine-tuning strategy, allowing the network to be optimised end-to-end on MRI data. This strategy enables effective domain adaptation from natural images (ImageNet) to medical imaging data and allows the model to learn disease-specific features associated with AD progression.

The architecture of VGG19 is illustrated in Figure 1 and can be summarised as follows:

Input Layer: It receives preprocessed MRI scans resized to 224×224 pixels, using a kernel size of 3×3 [32,33].

Convolutional Layers: The network is composed of multiple convolutional blocks, each containing two or more convolutional layers with 3×3 receptive fields and an increasing number of filters. As shown in Figure 1, the first block applies two convolutional layers with 64 filters, followed by deeper blocks with higher filter counts. Padding is set to one pixel and the stride to one, ensuring that spatial resolution is preserved and that the resulting activation maps retain the structural characteristics of the input images.

Figure 1: Architecture of VGG19 model.

Max-pooling layers: They are inserted between convolutional blocks to reduce spatial dimensionality and computational complexity. Each max-pooling operation selects the maximum value within a local window, as defined by Equation (1):

$$MP = Floor\left(\frac{Ix - P}{s} + 1\right) \qquad (1)$$

where I_x denotes the input size, P the pooling window size, and S the stride.

In this work, max-pooling layers use a 2×2 window with a stride of 2, reducing feature map dimensions by half.

Flatten layer: Following the convolutional blocks, a flattening layer converts the resulting high-dimensional feature maps into a one-dimensional feature vector suitable for input to the fully connected layers.

Fully connected layers: The original architecture of VGG19 includes three fully connected layers, with 4096 neurons in the first two and an output layer matching class numbers [34]. In the proposed framework, these layers are replaced by a task-specific MLP classifier for AD stage classification.

Activation functions: Rectified Linear Unit (ReLU) activation functions are employed in all convolutional and fully connected layers to introduce non-linearity, while the Softmax activation function is used in the output layer for multi-class classification.

3.2 Neural networks

Neural networks, inspired by the human brain, are widely applied for classification, regression, and pattern recognition tasks [20, 24, 35]. They consist of an input layer, hidden layers, and an output layer, where interconnected neurons iteratively optimise weights and biases to learn complex patterns. In the proposed framework, a convolutional neural network (VGG19) extracts hierarchical features from MRI images, while a feedforward MLP classifies these features into AD stages. This combination captures both spatial patterns and high-level discriminative representations, enhancing the accuracy of AD diagnosis.

3.3 Methodology

The central objective of this study is to design and evaluate a robust deep learning framework for multi-class classification of AD disease stages from structural MRI scans. Accordingly, the main research question addressed in this work is: Can a fine-tuned VGG19 model combined with an MLP classifier outperform existing state-of-the-art methods for multi-class AD classification using MRI data?

Based on this question, the research hypothesis is that end-to-end fine-tuning of VGG19, coupled with a task-specific MLP classifier and appropriate class balancing, can significantly improve classification accuracy and generalisation across datasets. The proposed system processes MRI scans via data preprocessing, dataset balancing, deep feature extraction, and classification, with the complete framework architecture shown in Figure 2 and detailed methodology described below.

3.3.1 Data collection

This study employs the OASIS (Open Access Series of Imaging Studies) dataset [10, 16], publicly available via Kaggle. It contains 6,400 T1-weighted structural MRI images across four AD classes: Non-Demented, Very Mild Demented, Mild Demented, and Moderate Demented. The original images have a resolution of 176×208 pixels and are resized to 224×224 pixels to

match the input requirements of VGG19. Sample images from each class are presented in Figure 3, and the class distribution is summarised in Table 2.

3.3.2 Data preprocessing

Data preprocessing is a critical step in medical image analysis, as it improves model robustness and learning efficiency. The preprocessing pipeline includes image processing, dataset balancing, and data splitting :

Image processing: Image processing is performed using the ImageDataGenerator class provided by Keras. The MRI images are loaded from disk, decoded, and converted into RGB format. Pixel values originally ranging from 0 to 255 are rescaled to the range [0,1] to accelerate training and stabilise optimisation.

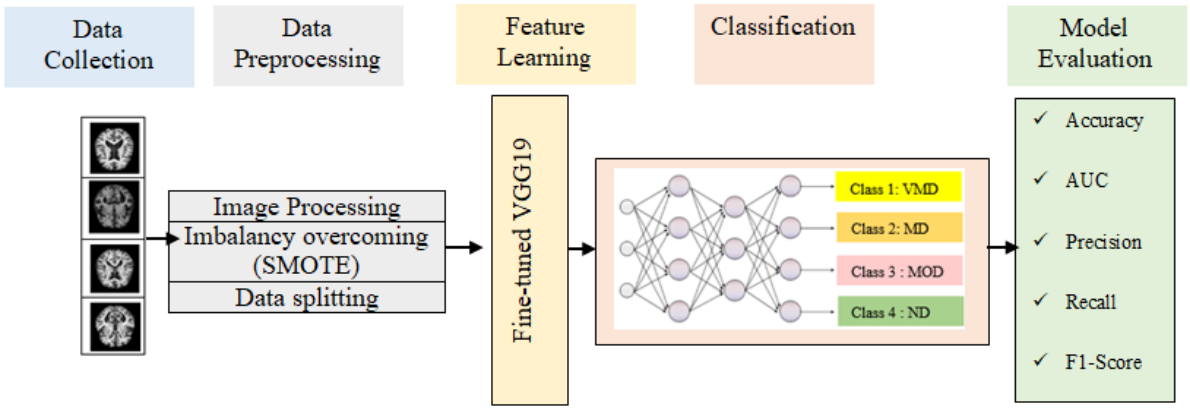

Figure 2: Overall architecture of the proposed model.

Table 2 : Data distribution in obtained dataset.

Class	# images	Train 80%	Test 20%
Mild Demented	896	717	179
Moderate Demented	2240	1792	448
Non Demented	3200	2560	640
Very Mild Demented	64	51	13

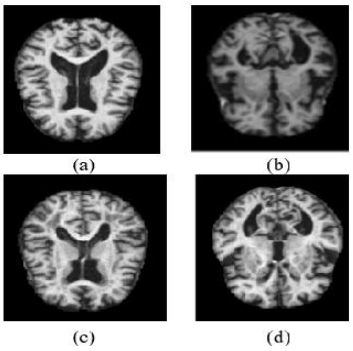

Figure 3: Sample images based on OASIS dataset, (a) MildDemented, (b) ModerateDemented, (c) NonDemented, (d) VeryMildDemented.

Dataset Balancing and Augmentation: The dataset exhibits a significant class imbalance, which can negatively affect classification performance and lead to biased predictions toward majority classes. To address this issue, SMOTE is applied as a data balancing strategy [36]. Although SMOTE was originally designed for structured data, in this work it is applied in a controlled manner at the image level using vectorised representations. Specifically, each grayscale MRI image is first resized and intensity-normalised, then flattened into a one-dimensional pixel-intensity vector. Standard SMOTE is subsequently applied to these flattened image vectors exclusively on the training set, generating synthetic samples for minority classes by interpolating between existing vectors in pixel space. The resampled vectors are then reshaped back into image form and used as input to the model. After applying SMOTE, the training dataset is expanded to 10,240 images, with 2,560 images per class, as reported in Table 3.

To further reduce the risk of overfitting or artefact amplification, SMOTE-based balancing is combined with conventional image augmentation techniques, including rotation (±15°), width and height shifts (±10% of the image size), and slight zooming (±10%).

No horizontal or vertical flipping is applied, as preserving anatomical orientation is essential in brain MRI analysis. Finally, the balanced and augmented images are organised into batches to facilitate efficient and stable model training.

Table 3 : Training data distribution after SMOTE

Class	# images
Mild Demented	2560
Moderate Demented	2560
Non Demented	2560
Very Mild Demented	2560

Data splitting: The dataset is divided into two subsets: 80% for training and 20% for testing, using the original 6,400 images. A stratified splitting strategy is adopted to preserve class proportions in both subsets and ensure unbiased evaluation.

3.3.3 Feature learning

In the proposed framework, feature learning is performed using the fine-tuned VGG19 network. The model is initialised with ImageNet pretrained weights to leverage its prior knowledge of visual patterns. During training, the layers of the fourth block (the deeper layers) of VGG19 are fine-tuned to adapt the learned representations to the MRI domain. The early layers capture low-level features such as edges and textures, while deeper layers learn higher-level structural patterns associated with AD, including ventricular enlargement, cortical thinning, and hippocampal atrophy.

3.3.4 Classification

The choice of employing an MLP classifier after the VGG19 feature extractor is motivated by its ability to model non-linear relationships between high-level deep features and output classes while maintaining relatively low computational complexity. Using a dedicated MLP also provides flexibility to tailor the classifier to the specific characteristics of AD stage classification.

The deep features learned by the fine-tuned VGG19 network are fed into a compact MLP classifier that serves as the final decision head of the framework. The MLP consists of a flattening input layer, followed by two fully connected hidden layers with 512 and 256 neurons, respectively, both using ReLU activation functions. Dropout regularisation is applied after each hidden layer, with rates of 0.5 and 0.3, to reduce overfitting and improve generalisation. The output layer contains four neurons corresponding to the four AD stages and employs a Softmax activation function.

This design choice was intentional: since the fine-tuned VGG19 backbone already produces highly discriminative, task-specific feature representations, a shallow classification head is sufficient to model class boundaries effectively. Deeper or more complex classifiers were found to increase model complexity without consistent performance gains and posed a higher risk of overfitting. The adopted MLP thus achieves a favourable balance between classification performance, robustness, and computational efficiency, making it suitable for clinical decision-support applications.

3.3.5 Experimental setup

The model is trained using the Adam optimiser for its adaptive learning rate and fast convergence, with an initial learning rate of 0.0001, dynamically adjusted via a ReduceLROnPlateau scheduler. Categorical cross-entropy serves as the loss function, and dropout layers mitigate overfitting and improve generalisation. Training is performed for 30 epochs with a batch size of 32. Experiments are implemented in TensorFlow/Keras on a GPU-accelerated Google Colab environment with 16 GB GPU memory and 32 GB RAM, averaging 90–100 seconds per epoch (~45 minutes total). Full source code will be publicly available upon acceptance for reproducibility.

3.3.6 Model evaluation

The proposed model is evaluated using standard performance metrics, including accuracy, Area Under the Curve (AUC), precision, recall, F1-score, and the Receiver Operating Characteristic (ROC) curve. All metrics are derived from the confusion matrix, which compares predicted and true class labels to assess prediction correctness and reliability. These metrics are described below.

Accuracy: It quantifies the proportion of correctly classified samples among all predictions. Accuracy is calculated using Equation (2):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

where:

- TP denotes the number of positive samples correctly predicted as positive,
- FP denotes the number of negative samples incorrectly predicted as positive,
- TN denotes the number of negative samples correctly predicted as negative, and
- FN denotes the number of positive samples incorrectly predicted as negative.

Precision: It measures the proportion of correctly predicted positive samples among all samples predicted as positive. It reflects the model's ability to avoid false positive predictions and is computed using Equation (3):

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

Recall: Also referred to as sensitivity, represents the proportion of correctly predicted positive samples relative to the total number of actual positive samples. It indicates the model's ability to identify positive cases and is calculated using Equation (4):

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

F1-score: It represents the harmonic mean of precision and recall. This metric is particularly valuable when dealing with imbalanced datasets, as it accounts for both false positives and false negatives. The F1-score is computed using Equation (5):

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (5)$$

AUC: It measures a model's ability to distinguish positive and negative classes across varying thresholds. Higher values indicate better performance, with 1 representing perfect discrimination and 0.5 equivalent to random guessing, providing a robust summary of predictive capability.

ROC Curve: It plots the true positive rate (sensitivity) against the false positive rate (1 − specificity) as the classification threshold varies. Each point on the ROC curve corresponds to a specific threshold, illustrating the trade-off between sensitivity and specificity. The ROC curve offers a comprehensive

visual assessment of the model's performance across a range of operating conditions.

Although ROC and AUC are traditionally used for binary classification, they are extended to the multi-class setting in this study using a one-vs-rest (OvR) strategy. For each AD stage, the corresponding class is treated as the positive class while the remaining classes are considered negative. The final AUC score is computed as a macro-average by averaging the AUC values obtained for each class, ensuring equal contribution of all disease stages regardless of class size.

4 Results and discussion

This section presents a comprehensive evaluation of the proposed DL framework for AD stage classification. The performance of the model is analysed using quantitative metrics, confusion matrices, learning curves, and ROC analyses. In addition, the robustness and generalisability of the proposed approach is validated through cross-dataset evaluation on the ADNI dataset. All reported results are obtained using the fine-tuned VGG19–MLP architecture, trained end-to-end on MRI data.

4.1 Performance evaluation on OASIS dataset

By fine-tuning VGG19 backbone on structural MRI scans, the proposed framework is able to adapt pretrained convolutional filters to disease-specific anatomical patterns associated with AD progression. This end-to-end optimisation enables the network to learn more discriminative and task-relevant features compared to fixed feature extraction approaches. On the OASIS dataset, the framework achieves: 98.83% accuracy, 0.993 precision, 0.988 recall and 0.991 F1-score. These results demonstrate the effectiveness of the fine-tuned VGG19–MLP framework in discriminating between the four AD stages. The confusion matrix illustrated in Figure 4, shows that the majority of MRI scans are correctly classified. Misclassifications are limited and mainly occur between neighbouring disease stages, such as Very Mild Demented and Mild Demented. This behaviour is clinically expected due to subtle structural differences between adjacent stages and highlights the capability of our model to capture progressive neurodegenerative patterns.

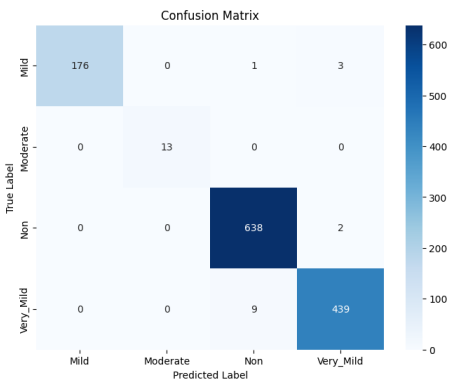

Figure 4 : Confusion matrix of proposed model on OASIS dataset.

The training and validation accuracy and loss curves, presented in Figure 5, demonstrate stable convergence behaviour. The absence of significant divergence between training and validation curves confirms that the fine-tuning strategy, combined with SMOTE-based dataset balancing and regularisation, effectively mitigates overfitting while preserving generalisation.

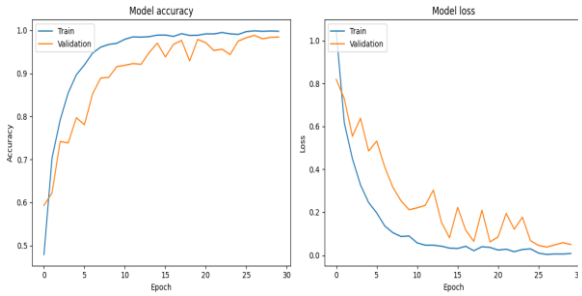

Figure 5: Accuracy and Loss curves of the proposed model during train and test on OASIS dataset.

The ROC curves for each class, shown in Figure 6, further confirm the discriminative power of the proposed model, with AUC values consistently close to unity. These results indicate that the fine-tuned VGG19 backbone provides robust feature representations under varying classification thresholds.

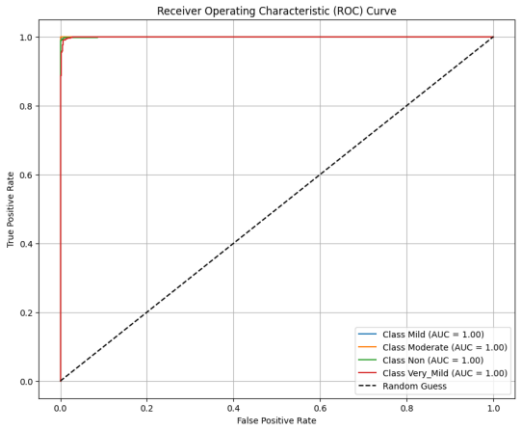

Figure 6: Obtained ROC curve on OASIS dataset.

4.2 Cross-dataset validation on ADNI dataset

To further evaluate the robustness and generalisability of the proposed framework, an independent validation was conducted on ADNI dataset [39], which comprises 5,154 T1-weighted structural MRI scans belonging to three classes: Cognitively Normal (CN), Mild Cognitive Impairment (MCI), and Alzheimer's Disease (AD). This dataset differs from OASIS in acquisition protocols, scanner characteristics, and subject demographics.

Thanks to the end-to-end fine-tuning of VGG19, the proposed model demonstrates excellent cross-dataset generalisation. On ADNI dataset, the framework achieves: 99.13% accuracy, 0.993 precision, 0.987 recall and 0.990 F1-score.

These results confirm that the fine-tuned convolutional layers successfully learn domain-invariant yet disease-relevant features, allowing the model to maintain high performance on unseen data. The confusion matrix for ADNI, illustrated in Figure 7, shows near-perfect classification, while the accuracy and loss curves in Figure 8 indicate stable inference behaviour. The ROC curve, presented in Figure 9, exhibits near-ideal separability, further validating the effectiveness of the fine-tuned VGG19 architecture.

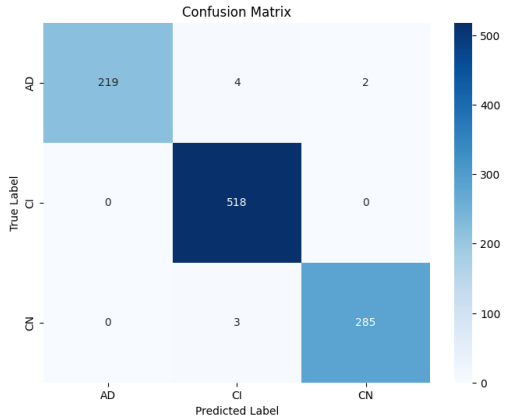

Figure 7 : Confusion matrix of proposed model on ADNI dataset

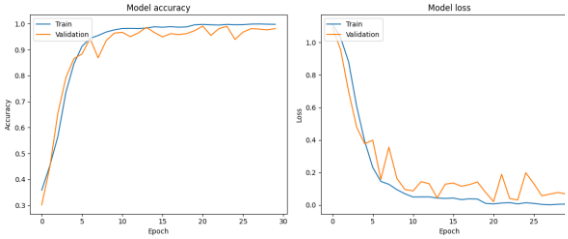

Figure 8: Accuracy and Loss curves of the proposed model during train and test on ADNI dataset.

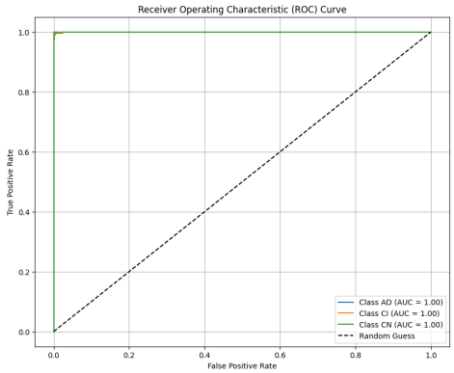

Figure 9: Obtained ROC curve on ADNI dataset.

4.3 Comparison with state-of-the-art methods

A quantitative comparison between the proposed approach and existing state-of-the-art methods is presented in Table 4.

4.4 Discussion

As summarised in Table 4, the proposed fine-tuned VGG19–MLP framework consistently outperforms existing methods for AD stage classification across multiple datasets.

On the OASIS dataset, the proposed model achieves an accuracy of 98.83%, exceeding previously reported multi-class approaches evaluated on the same dataset, such as CNN+MLP [27] (95.23%), e-CNN [23] (92.48%), and VGG16-based models without fine-tuning [26] (90.40%). This corresponds to an absolute improvement of +3.6% to +8.4%, which is significant for clinical-grade medical image classification.

On the ADNI dataset, the proposed framework achieves an accuracy of 99.13%, outperforming most existing multi-class AD classification methods reported in the literature, including hybrid ML models [19], CNN-based approaches [21,23], optimised VGG variants [24], and ensemble- or MLP-based classifiers [37,38]. These ADNI-based methods typically report accuracies ranging from 82.51% to 95.12%, even when employing complex hybrid architectures or multimodal inputs. Accordingly, the proposed approach yields an absolute performance improvement of approximately +4.0% to +16.6%, depending on the reference model. Notably, this gain is achieved using a comparatively lightweight architecture and MRI data only, without relying on multimodal fusion strategies or highly complex ensemble systems.

To provide a broader perspective on existing approaches to AD diagnosis, contextualise the proposed method within the wider body of related research, and highlight the diversity of modelling strategies explored in the literature, our model is also juxtaposed with approaches evaluated on different data modalities, such as EEG-based methods [20], or addressing alternative problem formulations, including binary classification or distinct clinical targets [17, 18, 21]. Owing to differences in imaging modality, class definitions, and experimental protocols, these studies are not directly comparable with the present multi-class MRI-based framework. Their inclusion is therefore intended to offer contextual insight into the broader research landscape rather than to establish direct performance comparisons.

Factors contributing to performance improvements: Several methodological factors explain the observed performance gains. First, unlike earlier studies that rely on pretrained CNNs as fixed feature extractors, the proposed framework adopts a fine-tuned VGG19 architecture, enabling effective domain adaptation from natural images to MRI scans. This end-to-end optimisation allows the model to capture AD-specific anatomical patterns, such as cortical thinning and ventricular enlargement, more effectively. Second, the combination of fine-tuned deep features with an MLP classifier provides strong discriminative power while avoiding excessive architectural complexity, in contrast to prior multi-stage or heavily cascaded models with limited generalisability. Third, the use of SMOTE for handling class imbalance significantly improves

minority-class recognition, directly contributing to consistently high precision, recall, and F1-scores across all AD stages.

Overfitting analysis and impact of SMOTE

Although the original datasets are relatively small and SMOTE is applied, multiple observations indicate that the proposed model does not suffer from overfitting. The training and validation accuracy curves exhibit stable convergence with minimal gap, while the corresponding loss curves decrease smoothly, indicating effective learning rather than memorisation. Cross-dataset validation on ADNI further confirms generalisation. If SMOTE had induced overfitting, performance on the independent ADNI dataset would likely degrade. Instead, the model achieves even higher accuracy (99.13%), alongside high precision (0.993), recall (0.987), and F1-score (0.990), confirming robustness. SMOTE is applied exclusively to the training set, while evaluation is conducted on untouched test data. Combined with data augmentation, regularisation, and stratified splitting, this strategy effectively mitigates overfitting risks.

In summary, the fine-tuned VGG19–MLP framework demonstrates state-of-the-art performance on both OASIS and ADNI datasets, with clear margins of improvement over existing methods. Strong cross-dataset generalisation, stable learning behaviour, and imbalance-aware training highlight the framework's suitability for reliable clinical decision-support applications.

It is worth noting that, in medical imaging tasks with limited data, architectural simplicity at the classification stage often improves robustness and reproducibility. Here, the combination of a fine-tuned deep backbone with a lightweight MLP classifier proved effective, achieving strong performance while limiting unnecessary model complexity.

5 Conclusion and future scope

This paper proposed an effective DL framework for multi-class classification of AD stages from structural MRI scans. The framework is built on a fine-tuned VGG19 network for end-to-end feature learning, followed by an MLP for stage-level classification into Non-Demented, Very Mild Demented, Mild Demented, and Moderate Demented categories.

Table 4: A comparative analysis of the proposed model with the state-of-the-art models.

Ref	Architecture	Dataset	Accuracy (%)	Precision	Recall	F1-score	AUC
[17]	ADDTLA	Private MRI dataset	91.70	NM	NM	NM	NM
[18]	SVM	CAT	71.00	NM	NM	NM	0,900
	Random Forest	CAT	86.00	NM	NM	NM	0,740
	Multinomial Logistic Regression	CAT	89.00	NM	NM	NM	NM
[19]	Three-tier cognitive hybrid ML, (HCM-1)	ADNI	90.24	0.905	0.947	0.903	0.968
	Three-tier cognitive hybrid ML, (HCM-2)	ADNI	95.12	0.952	0.975	0.952	0.990
[20]	SVM, DT, KNN	EEG dataset	75.00	NM	NM	NM	NM
	SVM, DT, KNN	EEG dataset	97.00 (HC vs. AD), 95.00 (HC vs. MCI), 83.00 (MCI vs. AD)	NM	NM	NM	NM
[21]	Multi-Modality Cascaded CNN (MC-CNN)	ADNI	93.26 (AD vs. NC), 82.95 (MCI vs. NC)	NM	NM	NM	0.956(AD vs. NC), 0.884(MCI vs. NC)
[22]	Multiscale MDNN	ADNI	82.51	NM	NM	NM	NM
[23]	e-CNN	OASIS	92.48	0.940	0.930	0.920	NM
[24]	CNN	ADNI	93.50	NM	NM	NM	0.94
[25]	Optimised VGG16 with AOA	ADNI	92.34	0,969	0,958	0,957	NM
[26]	Neural Network + VGG16	OASIS	90.40	0,905	0,904	0,904	0,969
[27]	CNN + MLP	OASIS	95.23	0,960	0,950	0,952	0,970
[37]	Ensemble-based classifier	ADNI	85.55	0,854	0,855	0,855	NM
		NRCD	90.05	0,888	0,898	0,893	NM
[38]	MLP	ADNI	89.00	0,850	0,870	0,890	NM
	CBLSTM + GAIN	ADNI	82.00	0,790	0,820	0,820	NM
	CBLSTM + SMOTE	ADNI	82.00	0,780	0,880	0,820	NM
Proposed	FineTunnedVGG19 +MLP	OASIS	98.83	0.993	0.988	0.991	1,000
Proposed	FineTunnedVGG19 +MLP	ADNI	99.13	0.993	0.987	0.990	1,000

The model was evaluated on the OASIS dataset, where class imbalance was addressed using the SMOTE technique. Experimental results on the held-out test set demonstrate strong performance, achieving an accuracy of 98.83% with consistently high precision, recall, F1-score, and AUC. To assess robustness and generalisability, the framework was further validated on the ADNI dataset, achieving an accuracy of 99.13%, confirming its strong cross-dataset performance.

Comparative analysis with state-of-the-art methods shows that the proposed framework achieves superior accuracy while maintaining a relatively simple and computationally efficient architecture. These results indicate that the proposed model can serve as a reliable decision-support tool for assisting clinicians in the early diagnosis and severity assessment of AD.

As future work, the proposed framework will be evaluated on additional large-scale and real-world clinical datasets to further validate its applicability in diverse diagnostic settings. Moreover, the integration of alternative deep learning architectures (e.g., ResNet, EfficientNet, Capsule Networks) and explainability techniques will be explored to enhance interpretability and clinical trust. Finally, optimising the framework for deployment on resource-constrained environments and investigating multimodal data integration represent promising directions for future research.

References

[1] N. Bansal and M. Parle, Dementia: an overview, Journal of Pharmaceutical Technology, Research and Management, vol. 2, no. 1, pp. 29–45, 2014. doi: 10.15415/jptrm.2014.21003.

[2] Z. Arvanitakis and D. A. Bennett, What is dementia? JAMA, vol. 322, no. 17, 2019. doi: 10.1001/jama.2019.11653

[3] C. Patterson, World Alzheimer Report 2018: The state of the art of dementia research—New frontiers, Technical Report, Alzheimer's Disease International, Jun. 2018.

[4] S. Przedborski, M. Vila, and V. Jackson-Lewis, Series introduction: Neurodegeneration—What is it and where are we? Journal of Clinical Investigation, vol. 111, no. 1, pp. 3–10, 2003. doi: https://doi.org/10.1172/JCI200317522.

[5] G. M. McKhann et al., The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging–Alzheimer's Association workgroups, Alzheimer's & Dementia, vol. 7, no. 3, pp. 263–269, 2011. doi: 10.1016/j.jalz.2011.03.005

[6] S. Sharma and K. Guleria, A deep learning model for early prediction of pneumonia using VGG19 and neural networks, in Proc. 3rd Int. Conf. Mobile Radio Communications & 5G Networks (MRCN), Springer, 2022, pp. 1–12. doi: 10.1007/978-981-19-7982-8_50.

[7] J. Giorgio et al., Modelling prognostic trajectories of cognitive decline due to Alzheimer's disease, NeuroImage: Clinical, vol. 26, Art. no. 102199, 2020. doi: https://doi.org/10.1016/j.nicl.2020.102199.

[8] C. Lian, M. Liu, J. Zhang, and D. Shen, Hierarchical fully convolutional network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 4, pp. 880–893, 2020. doi: 10.1109/TPAMI.2018.2889096.

[9] E. Moser et al., Magnetic resonance imaging methodology, European Journal of Nuclear Medicine and Molecular Imaging, vol. 36, suppl. 1, pp. 30–41, 2009. https://doi.org/10.1007/s00259-008-0938-3.

[10] D. S. Marcus et al., Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI data in young, middle-aged, nondemented, and demented older adults, Journal of Cognitive Neuroscience, vol. 19, no. 9, pp. 1498–1507, 2007. doi: https://doi.org/10.1162/jocn.2007.19.9.1498.

[11] S. Kumar, K. Guleria, and S. Tiwari, An analytical study of the COVID-19 pandemic: Fatality rate and influential factors, Design Engineering, pp. 12213–12225, 2021. doi: https://doi.org/10.1049/PBHE038E_ch4.

[12] A. Algani, Y.M. Ritonga, M. Bala, A. Ansari, M.S. Badr, M. Taloba, Machine learning in health condition check-up using Breiman's random forest algorithm, Measurement: Sensors, vol. 23, Art. no. 100406, 2022. doi: https://doi.org/10.1016/j.measen.2022.100406.

[13] S. Jindal, A. Sharma, A. Joshi, M. Gupta, Artificial intelligence fuelling healthcare, in Mobile Radio Communications and 5G Networks, Springer, Singapore, pp. 501–507, 2021. Doi: https://doi.org/10.1007/978-981-15-7130-5_40.

[14] N. K. Trivedi, A. Anand, U. K. Lilhore, and K. Guleria, Deep learning applications on edge computing, in Machine Learning for Edge Computing, CRC Press.

[15] H. Mansourifar and W. Shi, Deep synthetic minority over-sampling technique, arXiv:2003.09788, 2020. Available: http://arxiv.org/abs/2003.09788.

[16] S. Dubey, Alzheimer's dataset, 2019. Available: https://www.kaggle.com/tourist55/alzheimers-dataset-4-class-of-images.

[17] T. M. Ghazal and G. Issa, Alzheimer disease detection empowered with transfer learning, Computers, Materials & Continua, vol. 70, pp. 5005–5019, 2022. doi: https://doi.org/10.32604/cmc.2022.020866

[18] A. Revathi et al., Early detection of cognitive decline using machine learning and cognitive ability test, Security and Communication Networks, 2022(2):1-13, pp. 1–13, 2022. doi: https://doi.org/10.1155/2022/4190023.

[19] A. Khan and S. Zubair, Development of a three-tiered cognitive hybrid machine learning algorithm for effective diagnosis of Alzheimer's disease, Journal of King Saud University – Computer and

Information Sciences, 2022. doi: https://doi.org/10.1016/j.jksuci.2022.07.016.

[20] D. Pirrone, E. Weitschek, P. Di Paolo, S. De Salvo, M.C. De Cola, EEG signal processing and supervised machine learning to early diagnose Alzheimer's disease, Applied Sciences, vol. 12, no. 11, Art. no. 5413, 2022. doi: https://doi.org/10.3390/app12115413.

[21] M. Liu, D. Cheng, K. Wang, Y. Wang, Multi-modality cascaded convolutional neural networks for Alzheimer's disease diagnosis, Neuroinformatics, vol. 16, nos. 3–4, pp. 295–308, 2018. doi: 10.1007/s12021-018-9370-4.

[22] D. Lu, K. Popuri, G.W. Ding, R. Balachandar, M.F. Beg, Multiscale deep neural network-based analysis of FDG-PET images for early diagnosis of Alzheimer's disease, Medical Image Analysis, vol. 46, pp. 26–34, 2018. doi: 10.1016/j.media.2018.02.002.

[23] J. Islam and Y. Zhang, Brain MRI analysis for Alzheimer's disease diagnosis using an ensemble system of deep convolutional neural networks, Brain Informatics, vol. 5, no. 2, pp. 1–14, 2018. doi: 10.1186/s40708-018-0080-3.

[24] E.N. Marzban, A.M. Eldeib, I.A. Yassine, Y.M. Kadah, S.D. Ginsberg, For the Alzheimer's Disease Neurodegenerative Initiative (2020) Alzheimer's disease diagnosis from diffusion tensor images using convolutional neural networks, PLoS One, vol. 15, no. 3, Art. no. e0230409, 2020. doi: https://doi.org/10.1371/journal.pone.0230409.

[25] N. Deepa and S. P. Chokkalingam, Optimization of VGG16 using arithmetic optimization algorithm for early detection of Alzheimer's disease, Biomedical Signal Processing and Control, vol. 74, Art. no. 103455, 2022. https://doi.org/10.1016/j.bspc.2021.103455.

[26] S. Sharma, K. Guleria, S. Tiwari, and S. Kumar, A deep learning-based CNN model with VGG16 feature extractor for Alzheimer's disease detection using MRI scans, Measurement: Sensors, vol. 24, Art. no. 100506, 2022. https://doi.org/10.1016/j.measen.2022.100506.

[27] S. Murugan et al., DEMNET: A deep learning model for early diagnosis of Alzheimer's disease and dementia from MR images, IEEE Access, vol. 9, pp. 90319–90329, 2021. doi: 10.1109/ACCESS.2021.3090474.

[28] M. To˘gaçar, B. Ergen, Z. C¨omert, F. ¨Ozyurt, A deep feature learning model for pneumonia detection, IRBM, vol. 41, no. 4, pp. 212–222, 2020. doi: 10.1109/ACCESS.2021.3090474.

[29] P. Bansal, A. Vanjani, A. Mehta, J.C. Kavitha, S. Kumar, Improving melanoma classification accuracy using binary Harris hawks optimization, Soft Computing, vol. 26, no. 17, pp. 8163–8181, 2022. doi: https://doi.org/10.1007/s00500-022-07234-1.

[30] S. Gupta, P. Bansal, S. Kumar, ULBP-RF: A hybrid approach for malware image classification, in Proc. 5th Int. Conf. Parallel, Distributed and Grid Computing (PDGC), 2018. doi: 10.1109/PDGC.2018.8745989.

[31] K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv:1409.1556, 2014.

[32] J. Pardede, B. Sitohang, S. Akbar, and M. L. Khodra, Transfer learning using VGG16 for fruit ripeness detection, International Journal of Intelligent Systems and Applications, vol. 13, no. 2, pp. 52–61, 2021. doi: 10.5815/ijisa.2021.02.04.

[33] P. Xu, J. Zhao, J. Zhang, Identification of intrinsically disordered protein regions using deep neural networks, Algorithms, vol. 14, no. 4, Art. no. 107, 2021. doi: 10.3390/a14040107.

[34] A. De and A. S. Chowdhury, DTI-based Alzheimer's disease classification with rank-modulated fusion of CNNs and random forest, Expert Systems with Applications, vol. 169, Art. no. 114338, 2021. doi: 10.1016/j.eswa.2020.114338.

[35] E.K. Fanar, O. Bayat, A.D. Duru, Diagnosis of Alzheimer's disease using 2D MRI slices by CNN, Applied Bionics and Biomechanics, Art. no. 6690539, 2021. doi: 10.1155/2021/6690539.

[36] T. Pan, J. Zhao, W. Wu, and J. Yang, Learning imbalanced datasets based on SMOTE and Gaussian distribution, Information Sciences, vol. 512, pp. 1214–1233, 2020. doi: 10.1016/j.ins.2019.10.048.

[37] S. Ahmed et al., Ensembles of patch-based classifiers for Alzheimer's disease diagnosis, IEEE Access, vol. 7, pp. 73373–73383, 2019. doi: 10.1109/ACCESS.2019.2920011.

[38] D. Stamate et al., Applying deep learning to predicting dementia and mild cognitive impairment, in Artificial Intelligence Applications and Innovations, IFIP AICT, vol. 584, Springer, 2020, pp. 308–319. doi: 10.1007/978-3-030-49186-4_26.

[39] R. C. Petersen et al., Alzheimer's Disease Neuroimaging Initiative (ADNI): Clinical characterization, Neurology, vol. 74, no. 3, pp. 201–209, 2010. doi: 10.1212/WNL.0b013e3181cb3e25