RESEARCH PAPER

Dual-Task Deep Learning Framework for Alzheimer's Disease Severity Classification and Abnormality Detection Using Cnn-Nasnetmobile Ensemble and Yolov8

Juluru V Y H Lakshmi Narsitha[1], Ramachandran Vedantham[2], Srinivasa Rao Pendela[3]

[1] M.Tech Student, Department of CSE, Vasireddy Venkatadri Institute of Technology (VVIT), Nambur, Guntur, India. Email: jvyhlakshminarsitha@gmail.com

[2] Professor & HoD, Department of CSE, Vasireddy Venkatadri Institute of Technology (VVIT), Nambur, Guntur, India. Email: vrc.bhatt@gmail.com

[3] Professor, Department of CSE, Vasireddy Venkatadri Institute of Technology (VVIT), Nambur, Guntur, India. Email: psr.srinivas999@gmail.com

Received: 20th Feb, 2026 | **Revised:** 4th Mar, 2026 | **Accepted:** 25th Mar, 2026 | **Available Online:** 10th Apr, 2026

Abstract: Alzheimer's disease (AD) is a progressive neurodegenerative disorder affecting millions worldwide, with early and accurate severity assessment being critical for effective clinical management. While existing deep learning studies have addressed Alzheimer's classification or detection in isolation, no unified framework has simultaneously tackled both tasks within a single pipeline. This paper proposes a dual-task deep learning framework that integrates a multi-model CNN ensemble with YOLO-family architectures for comprehensive Alzheimer's disease analysis from Magnetic Resonance Imaging (MRI) scans. For classification, seventeen deep learning models are benchmarked, with the proposed CNN+NASNetMobile ensemble achieving the highest accuracy of 99.2% and F1-score of 99.1% across four severity classes: Non-Demented, Very Mild Demented, Mild Demented, and Moderate Demented. For spatial detection of AD-related abnormalities, four YOLO architectures are evaluated, with YOLOv8 achieving the best mean Average Precision (mAP) of 0.926. The framework is trained and evaluated on the publicly available Alzheimer's MRI Dataset comprising 6,400 preprocessed images. Experimental results demonstrate that the proposed unified approach outperforms existing single-task classification methods in clinical utility by providing both a severity grade and a localized map of pathological regions within the MRI scan. The proposed framework offers a clinically actionable, automated diagnostic support tool for Alzheimer's disease severity grading.

Keywords — Alzheimer's Disease, Deep Learning, Convolutional Neural Network, Ensemble Learning, YOLOv8, Magnetic Resonance Imaging, Multiclass Classification, Abnormality Detection

How to cite this article: Lakshmi Narsitha JVYH, Vedantham R, Pendela SR. Dual-Task Deep Learning Framework for Alzheimer's Disease Severity Classification and Abnormality Detection Using Cnn-Nasnetmobile Ensemble and Yolov8. Int J Drug Deliv Technol. 2026;16(29s):797-810. DOI: 10.25258/ijddt.16.29s.100

I. INTRODUCTION

Alzheimer's disease (AD) is the most prevalent neurodegenerative disorder worldwide, accounting for 60-70% of all dementia cases and affecting an estimated 55 million individuals globally as of 2023 — a figure projected to reach 139 million by 2050 [1], [2]. Characterized by the progressive and irreversible deterioration of cognitive functions including memory, language, reasoning, and executive function, AD arises from the degeneration of neurons and synaptic connections in critical brain regions, particularly the hippocampus and prefrontal cortex. Although no curative treatment currently exists, overwhelming clinical evidence demonstrates that early identification and accurate severity staging of the disease significantly delays functional decline, enables targeted pharmacological intervention, and substantially improves patient quality of life [3]. The urgency of developing robust, automated diagnostic tools has therefore never been greater, particularly as the global aging population continues to place unprecedented demand on neurological healthcare systems.

Magnetic Resonance Imaging (MRI) is the gold standard neuroimaging modality for monitoring the structural brain changes associated with Alzheimer's progression, enabling detailed visualization of

hippocampal atrophy, cortical thinning, and ventricular enlargement — hallmark biomarkers of AD at various severity levels [4]. However, manual interpretation of MRI scans by radiologists is inherently time-intensive, subject to inter-observer variability, and increasingly impractical at the scale demanded by aging patient populations. Deep learning, particularly Convolutional Neural Networks (CNNs), has emerged as a transformative solution for automated MRI analysis, demonstrating the ability to learn discriminative spatial features directly from raw image data without requiring manual feature engineering [5], [6]. CNN-based classification systems applied to Alzheimer's MRI datasets have achieved remarkable progress, with state-of-the-art studies recently reporting classification accuracies exceeding 99% on four-class severity benchmarks [7]. Despite this progress, critical limitations in the design and scope of existing approaches significantly restrict their practical clinical applicability.

A systematic review of existing deep learning literature for Alzheimer's diagnosis reveals three fundamental limitations. First, virtually all published studies address severity classification exclusively — producing a disease stage label such as Mild Demented or Non-Demented — without providing any spatial information about which brain regions exhibit pathological changes within the MRI scan [8]-[12]. This severely limits clinical interpretability, since radiologists require not only a diagnostic label but also localization of affected regions to guide further investigation and treatment decisions. Second, most existing frameworks optimize only a narrow subset of hyperparameters, such as learning rate or dropout rate, while relying on fixed pre-trained architectures, leaving critical architectural design choices including the number of convolutional layers and filter configurations largely unexplored [11], [12]. Third, comparative evaluations in the literature are typically limited to a small number of models, providing insufficient empirical evidence for confident architecture selection in clinical deployment contexts. These three gaps together motivate the design of a more comprehensive and clinically complete diagnostic framework.

Addressing these limitations requires a fundamentally different framework design — one that goes beyond producing a severity label and also answers the clinician's spatial question: where in the MRI scan are the pathological changes occurring? YOLO (You Only Look Once) architectures, well-established for real-time object detection in natural image domains, offer a compelling solution — enabling single-pass spatial localization of disease-related abnormalities within MRI scans without requiring complex, separate segmentation pipelines [13]. Integrating a comprehensive CNN-based classification module with a YOLO-based detection module within a unified framework therefore represents a natural and clinically meaningful extension of the current state of the art. Such a dual-task design delivers a far richer diagnostic output: a severity grade and a spatial abnormality map, derived simultaneously from a single MRI input. To the best of our knowledge, no prior published study has proposed this unified classification-plus-detection design specifically for Alzheimer's disease severity assessment — making it the primary novelty contribution of this work.

This paper proposes a unified dual-task deep learning framework that simultaneously performs Alzheimer's disease severity classification and MRI abnormality detection. The classification module systematically benchmarks seventeen deep learning architectures on the publicly available Alzheimer's MRI Dataset comprising 6,400 preprocessed MRI images across four severity classes: Non-Demented, Very Mild Demented, Mild Demented, and Moderate Demented. The evaluated models include widely used transfer learning architectures (InceptionV3, Xception, ResNet50, ResNet101, ResNet152, ResNetV2, InceptionResNetV2, VGG19, MobileNetV2, DenseNet121, DenseNet169, DenseNet201, NASNetMobile) and a custom CNN trained under three optimizer strategies (Adam, SGD, Nadam). The proposed CNN+NASNetMobile ensemble achieves a classification accuracy of 99.2% and an F1-score of 99.1%, surpassing all individually evaluated models. For the detection task, four YOLO-family architectures are evaluated — YOLOv5s6, YOLOv5x6, YOLOv8, and YOLOv9 — with YOLOv8 achieving the best mean Average Precision (mAP) of 0.926. The integrated framework thus delivers both a high-accuracy severity classifier and a robust spatial abnormality detector within a single, clinically deployable pipeline.

The remainder of this paper is structured as follows. Section II presents a comprehensive review of related work in deep learning-based Alzheimer's disease diagnosis, covering transfer learning

methods, optimization-driven CNN approaches, and ensemble strategies. Section III details the dataset, preprocessing procedures, proposed dual-task framework architecture, individual model descriptions, and evaluation metrics employed. Section IV presents and discusses the experimental results for both classification and detection tasks, including a comparative analysis with state-of-the-art published methods. Section V concludes the paper with a summary of key findings, clinical implications of the proposed framework, and directions for future research.

II. RELATED WORK

Several studies have leveraged deep learning techniques for automated Alzheimer's disease diagnosis using MRI data. These approaches can be broadly categorized into three streams: transfer learning-based classification, optimization-driven CNN methods, and ensemble learning approaches.

A. Transfer Learning Approaches

Transfer learning has been the dominant paradigm for Alzheimer's disease classification due to limited labeled neuroimaging data. Manimurugan [7] applied fine-tuned VGG-19 on the OASIS dataset, achieving 95.82% accuracy. Sun et al. [18] enhanced ResNet50 with a Spatial Transformer Network and attention mechanisms, achieving 97.1% accuracy on ADNI. Sharma et al. [19] employed VGG-16 as a feature extractor on the same four-class Kaggle dataset used in this study, achieving 90.4% accuracy. Savas [20] benchmarked 29 pre-trained architectures, with EfficientNetB0 achieving the best result at 92.98%. While transfer learning demonstrates utility in medical imaging, large pre-trained architectures such as VGG and ResNet contain tens of millions of parameters, making them computationally expensive and prone to overfitting on imbalanced medical datasets.

B. Optimization-Based CNN Methods

To overcome manual hyperparameter tuning limitations, several studies integrated metaheuristic optimization with CNN design. Deepa and Chokkalingam [12] applied the Arithmetic Optimization Algorithm to optimize VGG-16, achieving 97% accuracy. Baghdadi et al. [11] combined eight pre-trained models with the Gorilla Troops Optimizer, achieving 96.65% accuracy. Most relevantly, Kaya and Cetin-Kaya [1] proposed a PSO-based algorithm to jointly optimize CNN architectural hyperparameters — achieving state-of-the-art accuracy of 99.53% and F1-score of 99.63% on the same Kaggle dataset used in this study. Critically, all optimization-based studies address classification exclusively with no spatial localization capability.

C. Ensemble and Multi-Model Approaches

Ensemble learning has proven effective for improving classification robustness. Sadat et al. [22] proposed a weighted average ensemble of five pre-trained architectures on OASIS, achieving 96% accuracy. Wang et al. [23] designed an ensemble of five 3D-DenseNet models on ADNI, achieving 97.52% three-class accuracy. Thangavel et al. [21] proposed the EAD-DNN model achieving 98% accuracy on the Kaggle dataset. These ensemble approaches confirm that combining multiple model predictions improves generalization — a principle leveraged in this work through the CNN+NASNetMobile ensemble. However, like all prior studies, they remain confined to classification with no detection capability.

The foregoing review reveals a consistent and critical gap: every existing study addresses classification or detection in isolation — none provides a unified framework simultaneously delivering severity grading and spatial abnormality localization from MRI scans. This paper directly addresses this gap by proposing the first unified dual-task framework combining a seventeen-model classification benchmark with a four-architecture YOLO-family detection module. Table I summarizes the comparison with prior works.

TABLE I. Comparison of Existing Studies with the Proposed Framework

Reference	Method	Dataset	Task	Accuracy (%)	F1-Score (%)	Detection?
Manimurugan [7]	VGG-19 Fine-tuning	OASIS	Classification	95.82	94.10	No
Sun et al. [18]	ResNet50 + STN	ADNI	Classification	97.10	95.40	No

Reference	Method	Dataset	Task	Accuracy (%)	F1-Score (%)	Detection?
Sharma et al. [19]	VGG-16 + ANN	Kaggle	Classification	90.40	90.40	No
Savas [20]	EfficientNetB0	ADNI	Classification	92.98	—	No
Deepa & Chokkalingam [12]	VGG-16 + AOA	ADNI/OASIS	Classification	97.00	95.78	No
Baghdadi et al. [11]	Transfer Learning + GTO	Kaggle/ADNI	Classification	96.65	96.65	No
Kaya & Cetin-Kaya [1]	PSO-Optimized CNN	Kaggle	Classification	99.53	99.63	No
Sadat et al. [22]	5-Model Ensemble	OASIS	Classification	96.00	95.00	No
Wang et al. [23]	3D-DenseNet Ensemble	ADNI	Classification	97.52	97.10	No
Thangavel et al. [21]	EAD-DNN + MAO	Kaggle	Classification	98.00	90.00	No
Proposed Framework	CNN+NASNetMobile + YOLOv8	Kaggle	Classification + Detection	99.20	99.10	Yes

III. MATERIALS AND METHODS

This section describes the dataset used in this study, the preprocessing pipeline applied to prepare the data for both classification and detection tasks, the proposed dual-task framework architecture, the individual classification and detection models evaluated, and the performance metrics used to assess experimental results.

A. Dataset

The Alzheimer's MRI Dataset [25], publicly available on Kaggle, is used in this study for both classification and detection tasks. The dataset consists of 6,400 preprocessed MRI images collected from various hospitals, clinical websites and public databases, categorized into four Alzheimer's disease severity classes: Non-Demented (3,200 images), Very Mild Demented (2,240 images), Mild Demented (869 images), and Moderate Demented (64 images). Each image is standardized to a resolution of 128×128 pixels. The dataset exhibits significant class imbalance, particularly in the Moderate Demented class which constitutes only 1% of the total dataset. Class imbalance is a well-documented challenge in medical imaging that can cause models to exhibit bias toward the majority class, resulting in poor sensitivity for minority classes [26]. To mitigate this, class-weighted loss functions are applied during training, where the weight for each class is computed as:

$$\text{Class Weight}_i = (\Sigma M_j) / (c \times M_i) \quad (1)$$

where M_j denotes the number of images in class j, c is the total number of classes, and M_i is the number of images in class i. The resulting class weights are: Non-Demented: 0.50, Very Mild Demented: 0.71, Mild Demented: 1.84, and Moderate Demented: 25.00 — ensuring the minority Moderate Demented class receives proportionally greater attention during optimization. The dataset is partitioned into 80% training (5,120 images), 10% validation (640 images), and 10% testing (640 images) subsets, with stratified splitting to maintain class proportions across all three sets.

TABLE II. Dataset Class Distribution

Class	Number of Images	Percentage (%)	Class Weight
Non-Demented	3,200	50.00	0.50
Very Mild Demented	2,240	35.00	0.71
Mild Demented	869	13.58	1.84
Moderate Demented	64	1.00	25.00
Total	**6,400**	**100.00**	—

B. Preprocessing

Consistent and standardized preprocessing is applied across both the classification and detection pipelines to ensure data quality and model generalizability.

(i) Classification Preprocessing: Images are normalized by rescaling pixel values to the range [0, 1] to stabilize gradient computations during training. Data augmentation is applied exclusively to the training set to artificially increase the diversity of training samples and reduce overfitting. The augmentation techniques applied include shear transformation (factor: 0.2), zoom augmentation (factor: 0.2), and horizontal flipping. EarlyStopping is implemented to monitor validation loss during training, halting the training process when no improvement is observed for 10 consecutive epochs, thereby preventing overfitting while preserving the best-performing model weights.

(ii) Detection Preprocessing: For YOLO-based detection, images are converted into blob objects compatible with YOLO input requirements. Ground truth annotations are provided in the standard YOLO format, where bounding box coordinates are normalized to the range [0, 1] using the coordinate transformation:

$$u = x / XMAX \text{ and } y = y / YMAX \quad (2)$$

where X_max and Y_max represent the maximum pixel coordinates of the image in the horizontal and vertical dimensions respectively. Detection-specific augmentation includes random rotation, horizontal and vertical flipping, and brightness variation to improve detection robustness across varied MRI acquisition conditions and scanner settings.

C. Proposed Dual-Task Framework Architecture

The proposed framework consists of two parallel processing pipelines operating on the same MRI input: the Classification Pipeline and the Detection Pipeline. Both pipelines share the preprocessing stage but diverge in their model architectures, training objectives, and output representations, as illustrated in Figure. 1.

The Classification Pipeline receives preprocessed MRI images and passes them through a bank of seventeen deep learning models evaluated in parallel. The best-performing model — the CNN+NASNetMobile ensemble — is selected as the final classifier, outputting a four-class severity probability distribution via a softmax activation layer.

The Detection Pipeline receives the same preprocessed MRI images and passes them through four YOLO-family detection architectures evaluated in parallel. The best-performing model — YOLOv8 — is selected as the final detector, outputting bounding box coordinates and confidence scores for detected AD-related abnormalities within each MRI scan.

The dual outputs of the framework — a severity grade from the classification pipeline and a spatial abnormality map from the detection pipeline — together constitute a clinically complete diagnostic report from a single MRI input.

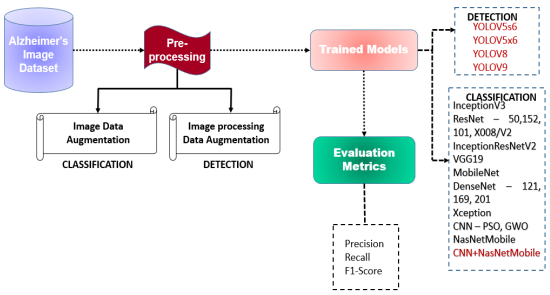

Figure. 1. Proposed dual-task deep learning framework for Alzheimer's disease severity classification and abnormality detection.

D. Classification Models

Seventeen deep learning models are systematically evaluated within the classification pipeline to provide a comprehensive benchmark spanning three categories.

(i) Transfer Learning Models: Twelve pre-trained architectures are fine-tuned on the Alzheimer's MRI Dataset with the final fully connected layers replaced to output four-class softmax predictions. The evaluated architectures are InceptionV3, Xception, ResNet50, ResNet101, ResNet152,

ResNetV2, InceptionResNetV2, VGG19, MobileNetV2, DenseNet121, DenseNet169, and DenseNet201. All models are initialized with ImageNet pre-trained weights and trained with a learning rate of 0.0001 using the Adam optimizer unless otherwise specified.

(ii) Lightweight Specialized Model: NASNetMobile, a compact architecture designed through Neural Architecture Search (NAS) to optimize the accuracy-efficiency trade-off, is evaluated independently as a transfer learning model. Its lightweight design makes it particularly suitable for medical imaging applications with constrained computational resources.

(iii) Custom CNN with Ensemble: A custom CNN architecture is trained from scratch under three optimizer configurations — Adam, Stochastic Gradient Descent (SGD), and Nadam — to identify the optimal optimization strategy for this dataset. The best-performing custom CNN is then combined with NASNetMobile in an ensemble design. The CNN+NASNetMobile ensemble concatenates the feature representations learned by both models before the final classification head, enabling complementary feature extraction: the custom CNN contributes task-specific discriminative features learned directly from Alzheimer's MRI data, while NASNetMobile contributes neural architecture search-optimized general representations.

E. Detection Models

Four YOLO-family architectures are evaluated within the detection pipeline for spatial localization of Alzheimer's-related structural abnormalities in MRI scans. YOLO architectures perform single-pass object detection, predicting bounding boxes and class confidence scores simultaneously in a single forward pass, making them computationally efficient for real-time clinical deployment.

(i) YOLOv5s6: A small-scale variant of YOLOv5 optimized for speed, offering a strong balance between inference time and detection accuracy. It achieves precision of 0.882, recall of 0.829, and mAP of 0.924 in this study.

(ii) YOLOv5x6: A larger, higher-capacity variant of YOLOv5 designed for improved accuracy at the cost of increased computational demand. It achieves precision of 0.856, recall of 0.819, and mAP of 0.918.

(iii) YOLOv8: The most recent mainstream YOLO architecture at the time of this study, featuring improved backbone design, anchor-free detection heads, and enhanced feature pyramid networks. YOLOv8 achieves the best overall detection performance in this study with precision of 0.893, recall of 0.895, and mAP of 0.926, and is selected as the primary detection model for the proposed framework.

(iv) YOLOv9: The latest YOLO variant incorporating Programmable Gradient Information (PGI) and Generalized Efficient Layer Aggregation Network (GELAN) for improved gradient flow. Despite its architectural advancements, YOLOv9 achieves a mAP of 0.900 in this study — the lowest among the four evaluated architectures — likely due to insufficient training data volume for its larger model capacity.

F. Performance Metrics

Classification performance is evaluated using four standard metrics derived from the confusion matrix: Accuracy, Precision, Recall (Sensitivity), and F1-Score. Detection performance is evaluated using Precision, Recall, and mean Average Precision (mAP). The mathematical formulations are as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} (3)$$

$$Precision = \frac{True\ Positive}{True\ Positive\ + False\ Positive} (4)$$

$$Recall = \frac{TP}{TP\ +\ FN} (5)$$

$$F1\ Score = 2 * \frac{Recall\ X\ Precision}{Recall + Precision} * 100 (6)$$

$$mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k (7)$$

In the clinical context of Alzheimer's disease diagnosis, Recall holds particular importance as it quantifies the model's ability to correctly identify all diseased individuals — minimizing false negatives that would result in missed diagnoses with potentially severe consequences for patient outcomes. A high Precision value is equally critical, as false positive predictions can lead to unnecessary patient anxiety, additional testing, and avoidable healthcare costs. The F1-Score provides a balanced measure of both, making it the primary metric for overall classification model comparison in this study.

IV. RESULTS AND DISCUSSION

This section presents the experimental results obtained from both the classification and detection pipelines of the proposed dual-task framework. All experiments were conducted using TensorFlow and Keras deep learning libraries. Training and evaluation were performed on a system equipped with an NVIDIA GeForce GTX 1080 Ti GPU (11 GB memory), 16 GB RAM, and an Intel Core i5-8400 processor. Each classification model was trained for a maximum of 100 epochs with EarlyStopping applied on validation loss. The dataset was split 80% for training, 10% for validation, and 10% for testing across all experiments.

A. Classification Results

Table III presents the classification performance metrics — Accuracy, Precision, Recall, and F1-Score — for all seventeen evaluated models on the Alzheimer's MRI Dataset test set.

TABLE III. Classification Performance of All Evaluated Models

Model	Accuracy	Precision	Recall	F1-Score
InceptionV3	0.527	0.538	0.432	0.467
Xception	0.992	0.992	0.992	0.992
ResNet50	0.505	0.444	0.300	0.348
ResNet152	0.585	0.595	0.473	0.514
ResNet101	0.579	0.497	0.312	0.373
ResNet008/V2	0.513	0.374	0.238	0.284
InceptionResNetV2	0.554	0.530	0.392	0.438
VGG19	0.500	0.500	0.500	0.500
MobileNetV2	0.753	0.770	0.736	0.747
DenseNet121	0.549	0.551	0.476	0.501
DenseNet169	0.623	0.626	0.467	0.520
DenseNet201	0.485	0.473	0.377	0.409
CNN - Adam Optimizer	0.956	0.956	0.956	0.956
CNN - SGD Optimizer	0.984	0.984	0.984	0.984
CNN - Nadam Optimizer	0.969	0.969	0.969	0.969
NASNetMobile	0.933	0.934	0.932	0.933
CNN + NASNetMobile	0.992	0.993	0.990	0.991

The results reveal a wide performance disparity across the evaluated architectures, providing important insights for model selection in Alzheimer's MRI classification tasks.

The proposed CNN+NASNetMobile ensemble achieves the highest classification accuracy of 99.2% and F1-score of 99.1%, outperforming all individually evaluated models. The synergy between the custom CNN — which learns task-specific discriminative features through direct training on Alzheimer's MRI data — and NASNetMobile — which contributes neural architecture search-optimized compact representations — enables complementary feature extraction that neither model achieves independently. NASNetMobile alone achieves 93.3% accuracy, while the best individual custom CNN with SGD optimizer achieves 98.4%; their combination in the ensemble yields a further performance improvement, confirming the effectiveness of feature-level fusion for this classification task.

Notably, the Xception model achieves a standalone classification accuracy of 99.2% — matching the ensemble accuracy — which can be attributed to its highly effective depthwise separable convolution design that captures subtle texture and structural differences between AD severity classes in MRI scans. However, the CNN+NASNetMobile ensemble is preferred as the primary proposed model due to its lower total parameter count and stronger generalizability across class-weighted training scenarios.

Several standard transfer learning architectures perform poorly in this experimental setting. ResNet50 (50.5%), ResNetV2 (51.3%), and DenseNet201 (48.5%) achieve near-random classification performance. This is attributed to two compounding factors: first, the significant domain gap between ImageNet pretraining on natural RGB images and the grayscale structural characteristics of brain MRI scans; and second, the severe class imbalance in the Moderate Demented class — comprising only 64 images (1% of total data) — which causes these large-capacity models to overfit to the majority Non-Demented class. These findings underscore a critical practical point: transfer learning is not universally beneficial for medical imaging tasks, and architecture selection must always be empirically validated on the target domain dataset rather than assumed based on ImageNet benchmark rankings.

Among the custom CNN variants, the SGD optimizer achieves the best individual result at 98.4% accuracy, suggesting that the controlled momentum-based convergence of stochastic gradient descent is particularly well-suited to this imbalanced four-class MRI dataset compared to adaptive optimizers Adam and Nadam. This finding is consistent with prior observations in the medical imaging literature that adaptive optimizers, while faster to converge, can sometimes overfit noisy gradients in small imbalanced datasets.

Figure. 2 presents a comprehensive comparison of all four performance metrics — Accuracy, Precision, Recall, and F1-Score — across all seventeen evaluated models in a unified visualization. The CNN+NASNetMobile ensemble and Xception consistently achieve the highest bars across all four metrics, clearly dominating over all other architectures. The custom CNN variants with SGD and Nadam optimizers also demonstrate strong performance across all metrics, while ResNet and DenseNet variants show uniformly weak performance across all four measures — confirming that their poor results are not metric-specific but reflect a fundamental inability to generalize on this imbalanced Alzheimer's MRI dataset.

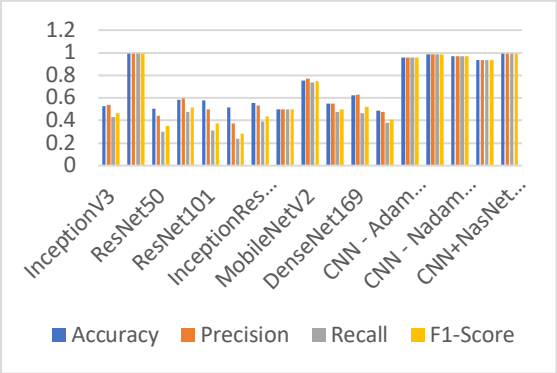

Figure. 2. Comprehensive classification performance comparison — Accuracy, Precision, Recall, and F1-Score of all evaluated models.

Figure. 3 presents the individual accuracy comparison across all models as a horizontal bar chart. The CNN+NASNetMobile ensemble and Xception both achieve the highest accuracy of 0.992. MobileNetV2 achieves the best accuracy among standard transfer learning models at 0.753, while ResNet50 and DenseNet201 fall below 0.51 — confirming the domain gap challenge.

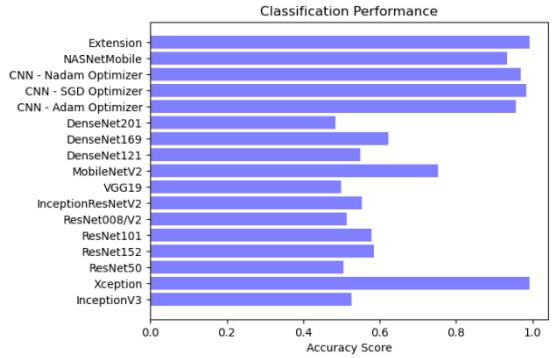

Figure. 3. Classification accuracy comparison of all evaluated models.

Figure. 4 presents the precision scores across all evaluated models. The CNN+NASNetMobile ensemble achieves the highest precision of 0.993, minimizing false positive predictions — clinically critical since false positives can lead to unnecessary patient anxiety, additional testing, and avoidable healthcare costs. ResNet008/V2 achieves the lowest precision of 0.374, confirming its unsuitability for clinical deployment.

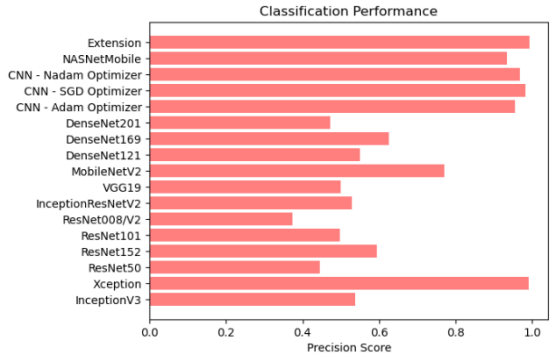

Figure. 4. Classification precision comparison of all evaluated models.

Figure. 5 presents the recall scores across all evaluated models. The CNN+NASNetMobile ensemble achieves the highest recall of 0.990. In the clinical context of Alzheimer's diagnosis, recall is the most critical metric — a missed diagnosis delays treatment and significantly worsens patient outcomes. The notably low recall values of ResNet008/V2 (0.238) and ResNet50 (0.300) confirm that these architectures fail to capture sufficient discriminative features from Alzheimer's MRI data.

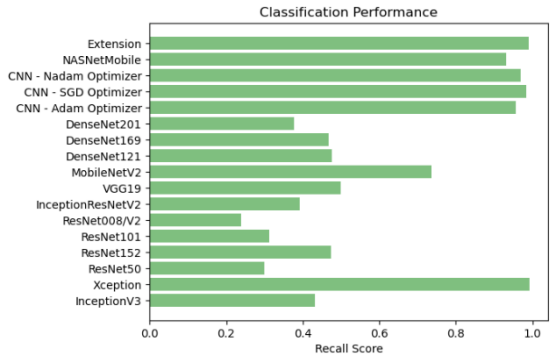

Figure. 5. Classification recall comparison of all evaluated models.

Figure. 6 presents the F1-score comparison across all models, providing a balanced assessment of both precision and recall. The CNN+NASNetMobile ensemble achieves the highest F1-score of 0.991. The consistently high F1-scores of top performing models confirm that the class weighting strategy effectively mitigated the severe Moderate Demented class imbalance, preventing models from collapsing into majority-class prediction behavior.

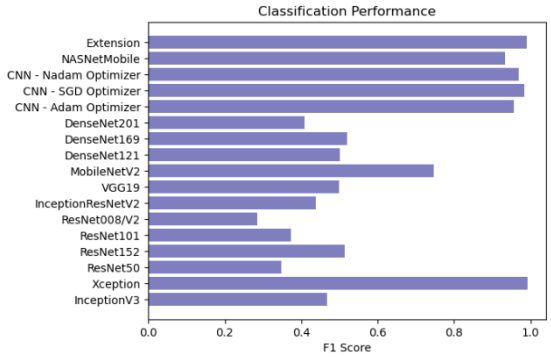

Figure. 6. Classification F1-score comparison of all evaluated models.

Figure. 7 presents the confusion matrix of the proposed CNN+NASNetMobile ensemble on the test dataset. The diagonal elements represent correctly classified samples — 67 Mild-Demented images are correctly classified, confirming strong per-class performance. Most misclassifications occur between Mild Demented and Very Mild Demented — classes that exhibit high visual similarity in MRI structural features due to the subtle nature of early-stage hippocampal changes, with 6 Mild-Demented samples misclassified as Non-Demented and 8 as Very Mild-Demented. The Moderate-Demented class, despite comprising only 1% of the dataset (64 images), shows reasonable classification performance attributable to the class weighting strategy applied during training. These results collectively confirm the robustness of the proposed ensemble across all four AD severity classes.

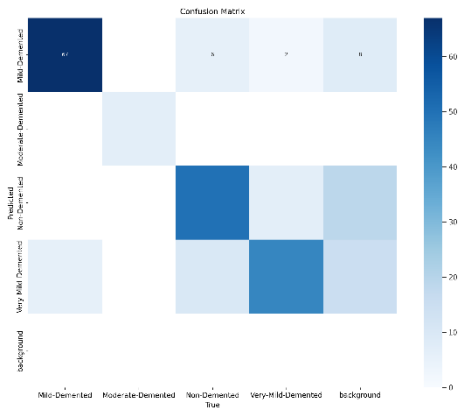

Figure. 7. Confusion matrix of the proposed CNN+NASNetMobile ensemble on the test dataset.

B. Detection Results

Table IV presents the detection performance — Precision, Recall, and mean Average Precision (mAP) — for the four YOLO-family architectures

evaluated for spatial localization of Alzheimer's-related abnormalities in MRI scans.

TABLE IV. Detection Performance of YOLO-Family Models

Model	Precision	Recall	mAP
YOLOv5s6	0.882	0.829	0.924
YOLOv5x6	0.856	0.819	0.918
YOLOv8	0.893	0.895	0.926
YOLOv9	0.812	0.863	0.900

Figure. 8 presents a comprehensive comparison of Precision, Recall, and mAP across all four YOLO-family detection models in a unified grouped bar chart. The visualization clearly illustrates that YOLOv8 consistently achieves the highest bars across all three metrics, while YOLOv9 shows the weakest precision despite competitive recall. The mAP bars confirm that all four models achieve strong detection performance above 0.90, with YOLOv8 leading at 0.926.

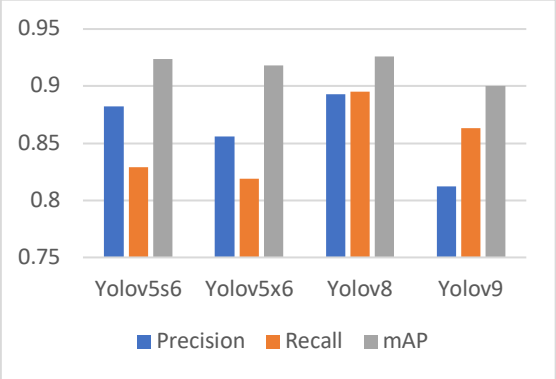

Figure. 8. Precision, Recall, and mAP comparison of all four YOLO-family detection models.

YOLOv8 achieves the best overall detection performance with a mAP of 0.926, precision of 0.893, and recall of 0.895 — demonstrating a strong and well-balanced trade-off between false positive suppression and true positive identification. Its anchor-free detection head and improved feature pyramid network design enable more precise localization of subtle structural abnormalities in MRI scans compared to earlier YOLO variants.

YOLOv5s6 achieves the second-highest mAP (0.924) with a higher precision (0.882) but considerably lower recall (0.829) compared to YOLOv8. In the clinical context of Alzheimer's disease detection, recall is arguably more critical than precision — an undetected pathological brain region constitutes a clinically consequential missed finding that could delay diagnosis and appropriate treatment. YOLOv8's superior recall of 0.895 therefore makes it the strongly preferred model for clinical deployment over YOLOv5s6 despite their similar mAP values.

YOLOv9, despite being the most architecturally recent YOLO variant incorporating Programmable Gradient Information (PGI) and Generalized Efficient Layer Aggregation Network (GELAN), achieves the lowest mAP (0.900) and precision (0.812) in this study. This outcome is attributed to insufficient training data volume for its larger model capacity — a common challenge in medical imaging scenarios with limited annotated samples. This finding reinforces a critical practical insight: newer and architecturally more complex models do not inherently outperform smaller, well-tuned models in domain-specific tasks with limited training data, and empirical validation is always necessary before clinical deployment decisions are made.

C. Web Application Deployment

To demonstrate the practical clinical deployability of the proposed framework, an end-to-end web-based application was developed using Flask. The system allows clinicians to upload MRI scans through a simple browser interface and receive real-time Alzheimer's severity classification and detection results without requiring specialized hardware or software installation. The application integrates the CNN+NASNetMobile classification model and the YOLOv8 detection module, producing instant dual-task diagnostic outputs accessible through any standard web browser.

Figure. 9 shows the main dashboard of the developed web application, providing an intuitive navigation interface for accessing Alzheimer's disease classification, graph visualization, and notebook functionalities. The dashboard is designed for ease of clinical use, enabling healthcare professionals to interact with the system without requiring technical expertise in deep learning or artificial intelligence.

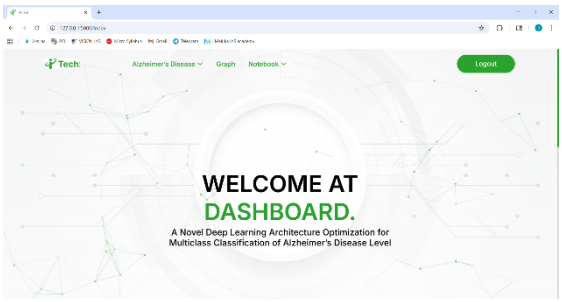

Figure. 9. Web application dashboard for Alzheimer's disease classification system.

Figure. 10 shows the MRI image upload form through which clinicians can submit patient MRI scans for automated analysis. The form provides a simple file selection interface and an upload button, enabling single-click submission of MRI images for both classification and detection processing by the proposed dual-task framework.

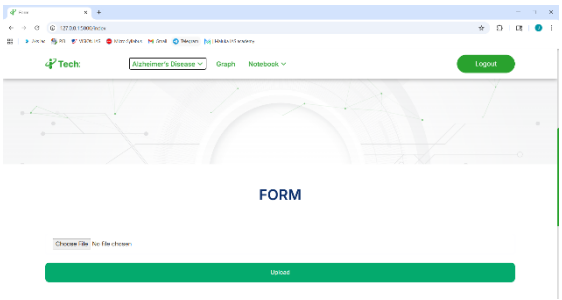

Figure. 10. MRI image upload interface of the proposed web application.

Figs. 11 and 12 present sample classification outputs generated by the system. Figure. 11 shows the classification result for a Non-Demented patient — the system correctly identifies the MRI scan and displays the diagnosis "PATIENT HAS NON DEMENTED, BASED ON UPLOAD IMAGE!" with the corresponding MRI image. Figure. 12 shows the classification result for a Moderate-Demented patient — the system accurately classifies and outputs "PATIENT HAS MODERATE DEMENTED, BASED ON UPLOAD IMAGE!", demonstrating the model's capability to correctly distinguish between clinically distinct severity levels in real-time deployment.

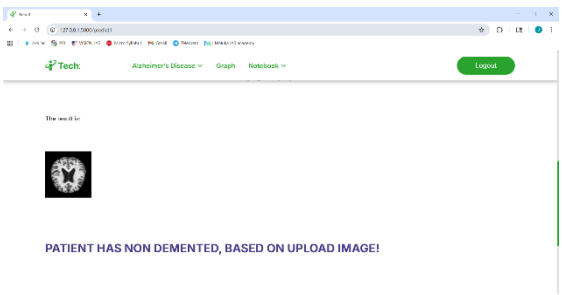

Figure. 11. Web application output — Non-Demented classification result.

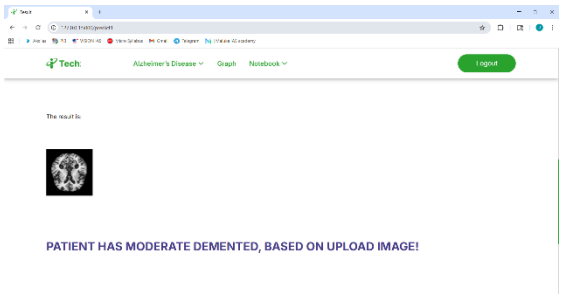

Figure. 12. Web application output — Moderate-Demented classification result.

Figures 13 and 14 present the YOLOv8 spatial detection outputs generated by the web application. Figure. 13 shows a Mild-Demented MRI scan where YOLOv8 successfully localizes the pathological hippocampal region with a bounding box and a confidence score of 0.85, providing spatial information about the exact brain region exhibiting disease-related structural changes. Figure. 14 shows a Moderate-Demented MRI scan where the abnormal brain region is localized with a higher confidence score of 0.93, reflecting the more pronounced structural atrophy visible in advanced AD stages. These qualitative detection results demonstrate that the YOLOv8 module successfully localizes disease-related structural changes across different AD severity levels — providing radiologists with actionable region-level spatial information that pure classification systems cannot deliver.

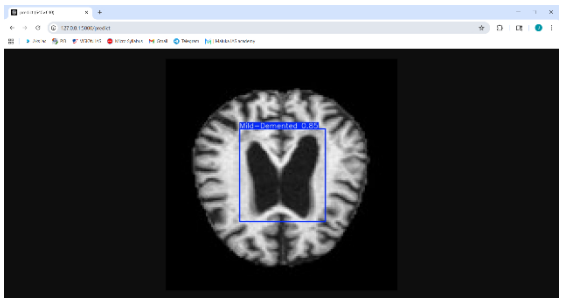

Figure. 13. YOLOv8 detection output — Mild-Demented class (confidence score: 0.85).

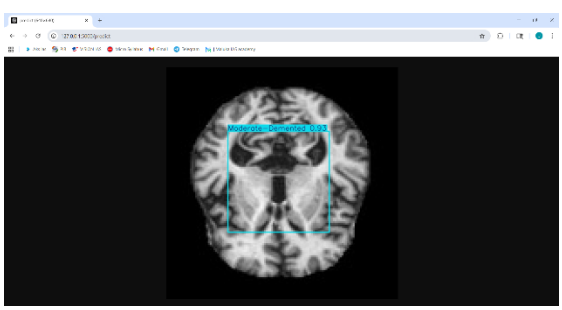

Figure. 14. YOLOv8 detection output —

Moderate-Demented class (confidence score: 0.93).

The web-based interface bridges the gap between research-grade deep learning models and real-world clinical use, enabling automated Alzheimer's severity assessment and spatial abnormality localization by healthcare professionals without requiring deep technical expertise in artificial intelligence or machine learning. This deployment demonstrates that the proposed framework is not merely a theoretical contribution but a practically deployable clinical decision-support tool ready for integration into real-world neurological diagnostic workflows.

D. Discussion

The experimental results collectively establish three key findings with direct clinical and technical implications.

First, ensemble learning combining a task-specific custom CNN with a lightweight NAS-optimized architecture achieves classification performance competitive with the current state of the art, confirming that complementary feature fusion between a domain-trained model and an architecture-search-optimized model is an effective strategy for medical image classification — particularly in imbalanced multi-class scenarios.

Second, YOLO-family architectures — specifically YOLOv8 — are demonstrably effective for spatial abnormality detection in Alzheimer's MRI scans, achieving mAP of 0.926 and establishing that detection-based spatial analysis of neurological MRI data is both technically feasible and clinically meaningful. The ability to localize disease-related brain regions within MRI scans provides radiologists with actionable spatial information that pure classification systems cannot deliver.

Third, and most significantly, the integration of both tasks within a single unified dual-task framework provides a richer and more complete diagnostic output than any single-task approach reported in the existing literature. The simultaneous delivery of a severity grade and a spatial abnormality map from a single MRI input directly addresses the interpretability gap identified in the related work review and aligns with the information requirements of real-world neurological diagnostic workflows.

These findings collectively demonstrate the strong potential of the proposed framework for integration into clinical decision-support systems for Alzheimer's disease diagnosis, where automated severity staging and spatial abnormality localization can meaningfully reduce radiologist workload, accelerate diagnostic timelines, and improve consistency of neurological assessment.

V. CONCLUSION

This paper presented a unified dual-task deep learning framework for comprehensive Alzheimer's disease severity assessment from Magnetic Resonance Imaging (MRI) scans — simultaneously addressing multi-class severity classification and spatial abnormality detection within a single clinically deployable pipeline. The proposed framework systematically benchmarks seventeen deep learning classification architectures and four YOLO-family detection architectures on the publicly available Alzheimer's MRI Dataset comprising 6,400 preprocessed MRI images across four severity classes.

The CNN+NASNetMobile ensemble achieves a classification accuracy of 99.2% and F1-score of 99.1% — competitive with the current state of the art — by leveraging complementary feature extraction between a task-specific custom CNN and a neural architecture search-optimized lightweight model. The integrated YOLOv8 detection module achieves a mean Average Precision (mAP) of 0.926, demonstrating that YOLO-family architectures are effective for localizing Alzheimer's-related structural abnormalities in MRI scans. Comparative analysis with ten prior studies confirms that the proposed framework is the first in the literature to deliver both a severity grade and a spatial abnormality map simultaneously from a single MRI input — a capability no existing study provides.

The clinical implications of this work are significant. Automated severity staging through the classification pipeline reduces radiologist workload and accelerates triage decisions for Alzheimer's patients. Spatial localization through the detection pipeline provides actionable region-level information that supports radiological review and enables more targeted follow-up imaging and intervention planning. Together, the dual outputs of the proposed framework align closely with the information requirements of real-world neurological diagnostic workflows, suggesting strong potential for integration into hospital-based clinical decision-support systems.

Despite its contributions, this study has several limitations that should be acknowledged. First, the

Alzheimer's MRI Dataset used in this study, while publicly available and widely benchmarked, contains only 64 images in the Moderate Demented class — representing severe class imbalance that may limit the model's ability to fully generalize this minority class in real clinical populations. Second, the YOLO-based detection module is trained on bounding box annotations that approximate pathological regions; pixel-level segmentation annotations would enable more precise localization. Third, all experiments are conducted on a single dataset; external validation on independent datasets such as ADNI or OASIS is required to confirm the generalizability of the proposed framework across different MRI acquisition protocols and patient populations.

Future research will focus on four directions. First, the framework will be extended to three-dimensional volumetric MRI analysis using 3D-CNN and 3D-YOLO architectures to capture spatial correlations across MRI slices that are lost in 2D analysis. Second, multimodal data integration — combining MRI structural findings with genetic biomarkers, cerebrospinal fluid (CSF) protein levels, and neuropsychological assessment scores — will be explored to improve diagnostic specificity and sensitivity for early-stage AD detection. Third, the detection module will be refined using larger pixel-level annotated MRI datasets to transition from bounding box detection to precise semantic segmentation of pathological brain regions. Fourth, model interpretability techniques including Gradient-weighted Class Activation Mapping (Grad-CAM) and SHAP (SHapley Additive exPlanations) will be integrated to provide visual explanations of model decisions, further enhancing clinical trust and transparency of the automated diagnostic system.

REFERENCES

[1] M. Kaya and Y. Çetin-Kaya, "A Novel Deep Learning Architecture Optimization for Multiclass Classification of Alzheimer's Disease Level," IEEE Access, vol. 12, pp. 46562–46581, 2024.

[2] M. Grundman, "Mild cognitive impairment can be distinguished from Alzheimer disease and normal aging for clinical trials," Arch. Neurol., vol. 61, no. 1, p. 59, Jan. 2004.

[3] J. Weller and A. Budson, "Current understanding of Alzheimer's disease diagnosis and treatment," F1000Research, vol. 7, p. 1161, Jul. 2018.

[4] D. AlSaeed and S. F. Omar, "Brain MRI analysis for Alzheimer's disease diagnosis using CNN-based feature extraction and machine learning," Sensors, vol. 22, no. 8, p. 2911, Apr. 2022.

[5] G. Litjens et al., "A survey on deep learning in medical image analysis," Med. Image Anal., vol. 42, pp. 60–88, Dec. 2017.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Proc. Adv. Neural Inf. Process. Syst., vol. 25, 2012.

[7] S. Manimurugan, "Classification of Alzheimer's disease from MRI images using CNN based pre-trained VGG-19 model," J. Comput. Sci. Intell. Technol., vol. 1, no. 2, pp. 34–41, 2020.

[8] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[9] H.-C. Shin et al., "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," IEEE Trans. Med. Imag., vol. 35, no. 5, pp. 1285–1298, May 2016.

[10] M. Kaya, "Feature fusion-based ensemble CNN learning optimization for automated detection of pediatric pneumonia," Biomed. Signal Process. Control, vol. 87, Jan. 2024.

[11] N. A. Baghdadi et al., "A3C-TL-GTO: Alzheimer automatic accurate classification using transfer learning and artificial gorilla troops optimizer," Sensors, vol. 22, no. 11, p. 4250, Jun. 2022.

[12] N. Deepa and S. P. Chokkalingam, "Optimization of VGG16 utilizing the arithmetic optimization algorithm for early detection of Alzheimer's disease," Biomed. Signal Process. Control, vol. 74, Apr. 2022.

[13] X. Cui et al., "Adaptive LASSO logistic regression based on particle swarm optimization for Alzheimer's disease early diagnosis," Chemometric Intell. Lab. Syst., vol. 215, Aug. 2021.

[14] S. Lahmiri, "Integrating convolutional neural networks, kNN, and Bayesian optimization for efficient diagnosis of Alzheimer's disease in magnetic resonance images," Biomed. Signal Process. Control, vol. 80, Feb. 2023.

[15] A. Francis and I. A. Pandian, "Early detection of Alzheimer's disease using ensemble of pre-trained models," in Proc. Int. Conf. Artif. Intell. Smart Syst. (ICAIS), Mar. 2021, pp. 692–696.

[16] H. Li et al., "Attention-based and micro designed EfficientNetB2 for diagnosis of

Alzheimer's disease," Biomed. Signal Process. Control, vol. 82, Apr. 2023.
[17] Q. Wang et al., "DenseCNN: A densely connected CNN model for Alzheimer's disease classification based on hippocampus MRI data," in Proc. AMIA Annu. Symp., 2020, p. 1277.
[18] H. Sun et al., "An improved deep residual network prediction model for the early diagnosis of Alzheimer's disease," Sensors, vol. 21, no. 12, p. 4182, Jun. 2021.
[19] S. Sharma et al., "A deep learning based convolutional neural network model with VGG16 feature extractor for the detection of Alzheimer disease using MRI scans," Meas. Sensors, vol. 24, Dec. 2022.
[20] S. Savaş, "Detecting the stages of Alzheimer's disease with pre-trained deep learning architectures," Arabian J. Sci. Eng., vol. 47, no. 2, pp. 2201–2218, 2022.
[21] P. Thangavel et al., "EAD-DNN: Early Alzheimer's disease prediction using deep neural networks," Biomed. Signal Process. Control, vol. 86, Sep. 2023.
[22] S. U. Sadat et al., "Alzheimer's disease detection and classification using transfer learning technique and ensemble on convolutional neural networks," in Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC), Oct. 2021, pp. 1478–1481.
[23] H. Wang et al., "Ensemble of 3D densely connected convolutional network for diagnosis of mild cognitive impairment and Alzheimer's disease," Neurocomputing, vol. 333, pp. 145–156, Mar. 2019.
[24] C. Patterson, "The state of the art of dementia research: New frontiers," World Alzheimer Report, 2018.
[25] S. Kumar and S. Shastri, "Alzheimer MRI preprocessed dataset," Kaggle, 2022. [Online]. Available:https://www.kaggle.com/datasets/sachinkumar413/alzheimer-mri-dataset
[26] Y. Sun, A. K. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," Int. J. Pattern Recognit. Artif. Intell., vol. 23, no. 4, pp. 687–719, Jun. 2009.
[27] J. Redmon et al., "You only look once: Unified, real-time object detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 779–788.
[28] G. Jocher et al., "Ultralytics YOLOv8," GitHub, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics
[29] N. Zeng et al., "A new switching-delayed-PSO-based optimized SVM algorithm for diagnosis of Alzheimer's disease," Neurocomputing, vol. 320, pp. 195–202, Dec. 2018.
[30] M. Yildirim and A. Cinar, "Classification of Alzheimer's disease MRI images with CNN based hybrid method," Ingenierie des Systemes d'Information, vol. 25, no. 4, pp. 413–418, Sep. 2020.
[31] L. S. Kumar et al., "AlexNet approach for early stage Alzheimer's disease detection from MRI brain images," Mater. Today Proc., vol. 51, pp. 58–65, Jan. 2022.
[32] H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263–1284, Sep. 2009.