Transfer Learning and Stacked Ensembles for Neurological Disorder Classification

[1]Andhavarapu Tejtha, [2]Ravi Kumar T, [3]Panduranga Vital Terlapu, [4]Ramkishor Pondreti and [5]Suneel Gowtham Karudumpa

[1 2 3 4]Department of Computer Science and Engineering, Aditya Institute of Technology and Management,
Tekkali, Srikakulam, Andhra Pradesh, India.

[5]Department of Electrical and Electronics Engineering, Aditya Institute of Technology and Management,
Tekkali, Srikakulam, Andhra Pradesh, India.

[1]andhavaraputejitha@gmail.com, [2]ravi.4u.kumar@gmail.com, [3]vital2927@gmail.com, [4]ramkishor.pondreti@gmail.com, [5]goutham.suneel@gmail.com

Correspondence should be addressed to Ravi Kumar T : ravi.4u.kumar@gmail.com

Article Info
Journal of Machine and Computing (https://anapub.co.ke/journals/jmc/jmc.html)
Doi: https://doi.org/10.53759/7669/jmc202505107.
Received 29 October 2024; Revised from 30 January 2025; Accepted 22 March 2025.
Available online 05 July 2025.

Abstract – To enhance the identification and categorization of Alzheimer's Disease (AD) across four stages—Very Mild Dementia, Moderate Dementia, Mild Dementia, and Non-Dementia (Healthy Subjects). Leveraging a Kaggle dataset comprising 3,382 MRI brain images, the proposed methodology integrates transfer learning with the Inception V3 convolutional neural network to extract high-dimensional features, followed by ensemble stacking of machine learning (ML) models, including Neural Networks (NN 100x100, NN 70x70), XGBoost, CatBoost, AdaBoost, and a meta-learner. The dataset is enlarged to 299x299 pixels. It undergoes 10-fold cross-validation to check its performance. The features are saved in *.csv format for use in machine learning. Performance is assessed using AUC, Correctness Accuracy (CA), F1-score, Precision, and Recall, revealing the Stacking model's standout performance with an AUC of 0.959, CA of 0.870, and balanced metrics of 0.871, alongside NN 100x100's leading AUC of 0.967 and CA of 0.863. While XGBoost (AUC 0.928, CA 0.775) and CatBoost (AUC 0.881, CA 0.704) show moderate success, AdaBoost lags with an AUC of 0.681 and CA of 0.568, highlighting challenges with imbalanced data, particularly for the underrepresented Moderate Dementia class (64 images). The hybrid approach is good at identifying complex patterns in AD. It can help with early diagnosis and treatment. Future efforts will aim to augment the dataset volume, enhance configurations for the model, try different structures, and combine Various types of data.

Keywords – XGBoost, CatBoost, Self-Attention, Incetion V3, Steel Strength Estimation, Moderate Dementia Class, Meta-Learner, Data-Driven Analysis.

I. INTRODUCTION

The **Table 1** shows the ADStages and Diagnosis has five stages. Each stage shows different symptoms. The stages range from early changes to severe cognitive decline. Doctors use tests like PET scans and MRIs to diagnose the disease. They also use cognitive assessments. Treatment changes as the disease progresses. In the early stages, lifestyle management and cognitive therapies are used.

Table 1. Alzheimer's Disease Stages and Diagnosis

Stage	Effects & Symptoms	Diagnosis	Treatment
Preclinical Stage	No noticeable symptoms, but brain changes begin.	Biomarker tests (CSF analysis, PET scans).	No treatment required; lifestyle modifications may help delay onset.
Mild Cognitive Impairment (MCI)	Memory lapses, trouble finding words, and mild confusion.	Cognitive tests (MoCA, MMSE), MRI, PET scans.	Healthy diet, exercise, and monitoring; possible clinical trials.
Early-Stage Alzheimer's	Increased forgetfulness, difficulty with problem-solving, and mild personality changes.	Neurological exams, blood tests, and cognitive assessments.	Cholinesterase inhibitors (Donepezil, Rivastigmine).

Table 1. Continued

Stage	Effects & Symptoms	Diagnosis	Treatment
Moderate-Stage Alzheimer's	Significant memory loss, difficulty recognizing people, mood swings, confusion, and trouble with daily tasks.	Brain imaging (MRI, CT), cognitive tests.	Cholinesterase inhibitors, NMDA receptor antagonists (Memantine), behavioural therapy.
Severe/Late-Stage Alzheimer's	Loss of communication, inability to recognize family, severe cognitive decline, bedridden state.	Clinical evaluation based on symptoms and history.	Palliative care, support for caregivers, medications to manage symptoms (antipsychotics, antidepressants).

II. LITERATURE REVIEW

Mahamud et al. (2025) [1] addressed the critical need for early and explainable Alzheimer's detection using machine learning (ML) models. The authors developed a pipeline integrating feature extraction from clinical data and imaging modalities, followed by classification using interpretable ML algorithms such as decision trees and SHAP values. The model not only achieved competitive accuracy but also offered visual explanations for its predictions, thereby improving clinician trust and decision support in real-world settings. Topsukal et al. (2024) [2] proposed an ensemble of deep learning architectures with an enhanced Xception model as the core for detecting ADusing brain MRI images. The framework included advanced data augmentation techniques, image preprocessing, and fine-tuning of model layers. This approach improved convergence speed and achieved high classification performance across AD, MCI, and normal control classes, demonstrating the strength of using optimized pre-trained models in neuroimaging. Jenber Belay et al. (2024) [3] explored a combination of ensemble deep learning and quantum ML for Alzheimer's classification. The authors employed traditional CNN backbones and integrated them with quantum classifiers based on variational quantum circuits. The fusion approach outperformed classical methods on standard benchmarks, indicating quantum-enhanced computing could provide a new dimension in neurodegenerative disease detection.

Nasir et al. (2024) [4] focused on MRI-based classification, this research utilized multiple deep learning architectures including ResNet, DenseNet, and CNN-LSTM combinations. The models were trained on public datasets (e.g., ADNI), and performance metrics were evaluated using accuracy, precision, and AUC-ROC. The study highlighted the importance of spatial feature extraction in MRIs and presented a comparative performance analysis of architectures. Rana et al. (2024) [5] introduced a hybrid deep-learning model using InceptionV3 for feature extraction and a custom CNN classifier for prediction. It employed clinically relevant preprocessing like skull stripping and intensity normalization. The model was validated on real-world patient scans and demonstrated robustness in early-stage detection, positioning it as suitable for clinical deployment. Mujahid et al. (2023) [6] implemented an ensemble of EfficientNet-B2 and VGG-16 architectures to detect Alzheimer's Disease. Each model contributed features that were concatenated before final classification. Data preprocessing involved histogram equalization and slice selection from 3D MRIs. The ensemble model achieved superior accuracy and generalization compared to individual baselines. Bhushanm (2023) [7] presented a custom-designed Inception V3 model optimized for detecting Alzheimer's signs from neuroimaging data. Modifications included tuned activation functions and reduced parameter redundancy for faster training. The model was benchmarked against other CNN variants and achieved notable improvement in early-stage detection with minimal computational overhead.

Alatrany et al. (2023) [8], transfer learning was applied to CNN models pre-trained on ImageNet, which were then fine-tuned for Alzheimer's classification. Models like ResNet50 and DenseNet121 were ensembled using voting and averaging techniques. The approach demonstrated that transfer learning could effectively compensate for the small size of Alzheimer's datasets, providing significant accuracy gains. Sharma et al. (2022) [9] developed a modified Inception model integrating transfer learning and preprocessing steps such as normalization and contrast enhancement. The authors tested the model on 2D MRI slices and incorporated dropout layers to prevent overfitting. Results showed improved diagnostic precision and reduced training time, emphasizing the utility of architectural customization. Agarwal et al. (2021) [10] examined over 50 papers using transfer learning on neuroimaging data for Alzheimer's detection. The review categorized studies by model type, imaging modality (MRI, PET, CT), and training strategy. It concluded that transfer learning significantly enhances performance, particularly when domain adaptation is used between imaging datasets.

Helaly et al. (2022) [11] provided a structured review of deep learning applications in AD detection across five domains: image acquisition, preprocessing, model selection, evaluation metrics, and deployment. It highlighted advances in multimodal learning and fusion techniques, advocating for integrated imaging and clinical data analysis for future developments. In combined transfer learning with ensemble approaches to improve Alzheimer's classification using data from multiple MRI datasets. Augmentation techniques like rotation, scaling, and elastic deformation were applied to expand training data. The combined models showed improved sensitivity, particularly in detecting mild cognitive impairment (MCI), often missed in standard models. Sadat et al. (2021) [12] presented a comparative study of ensemble methods including majority voting, stacking, and bagging, applied to deep CNN models. Using ADNI dataset, the study found that stacking yielded the best performance, suggesting that model diversity plays a key role in improving classification outcomes. Jansi et al. (2024) [13] focused on the InceptionV3 model, the paper explored its performance on

Alzheimer's classification using both 2D and 3D MRIs. It reported an accuracy of 87.69% and emphasized the architecture's efficient handling of varying spatial resolutions, which is critical in detecting fine-grained neuro degeneration patterns.

The [14] introduced a CNN model using transfer learning for multiclass classification of Alzheimer's stages (CN, MCI, AD). The model was evaluated on balanced and imbalanced datasets using SMOTE and achieved a maximum accuracy of 93%, suggesting strong potential for clinical support tools that provide stage-wise diagnosis.

III. METHODOLOGY

The proposed model detects Alzheimer's Disease. It classifies brain images into four stages: Mild Dementia, Moderate Dementia, Non-Dementia (Healthy Subjects), and Very Mild Dementia. The process has two phases. The first phase is Data Collection and Feature Extraction. The second phase is the Prediction Process. ML techniques and transfer learning are used to diagnose AD accurately. The workflow involves collecting ADimages from a Kaggle dataset, pre-processing them to standardize dimensions and enhance diversity, and storing them in *.PNG or *.JPEG formats for deep learning framework compatibility. The Inception V3 transfer learning model extracts features, which are transformed and stored in *.csv format for ML model integration. The dataset is split using 10-fold cross-validation for robust performance evaluation. ML models help identify patterns and classify stages of Alzheimer's. A stacking approach combines predictions from individual models to create a meta-learner. The process involves building the best model by evaluating the stacked and individual models. An unknown Alzheimer's image is input into the model. The model predicts the probability of the image being associated with one of four classes. Performance analysis is done to assess the model's effectiveness across all classes.

ML models help identify patterns and classify stages of Alzheimer's. A stacking approach combines predictions from individual models to create a meta-learner. The process involves building the best model by evaluating the stacked and individual models. An unknown Alzheimer's image is input into the model. The model predicts the probability of the image being associated with one of four classes. Performance analysis is done to assess the model's effectiveness across all classes.

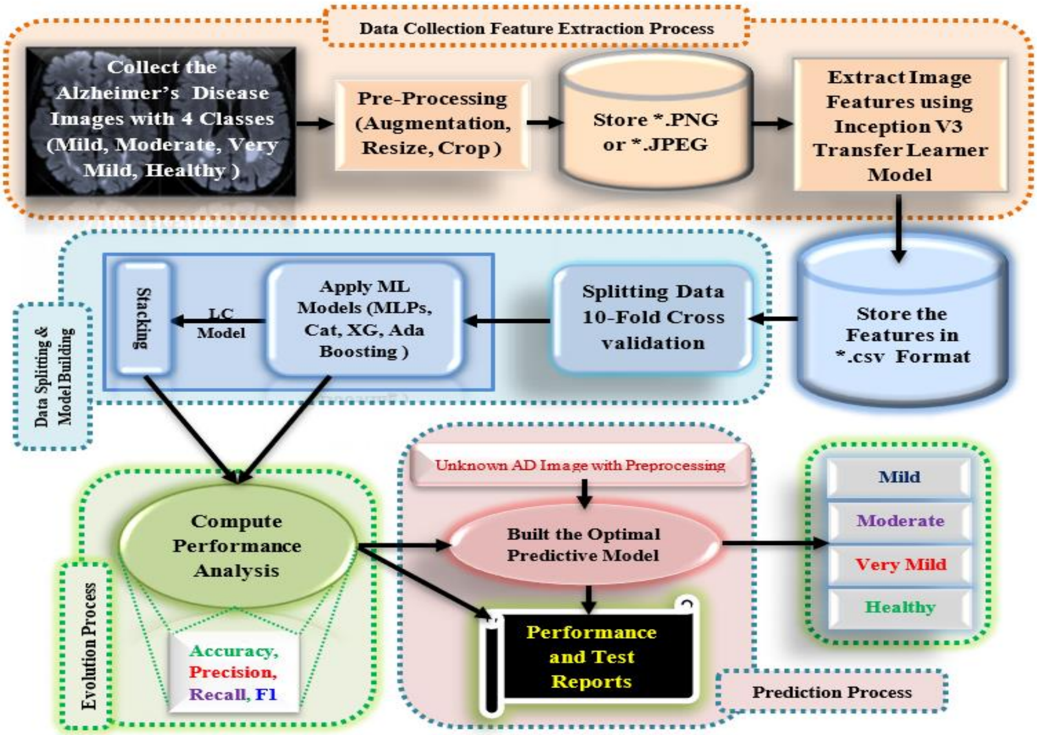

Data Collection Feature Extraction Process
Collect the Alzheimer's Disease Images with 4 Classes (Mild, Moderate, Very Mild, Healthy)
Pre-Processing (Augmentation, Resize, Crop)
Store *.PNG or *.JPEG
Extract Image Features using Inception V3 Transfer Learner Model
Store the Features in *.csv Format
Splitting Data 10-Fold Cross validation
Apply ML Models (MLPs, Cat, XG, Ada Boosting)
LC Model
Stacking
Data Splitting & Model Building
Compute Performance Analysis
Accuracy, Precision, Recall, F1
Evolution Process
Unknown AD Image with Preprocessing
Built the Optimal Predictive Model
Performance and Test Reports
Mild
Moderate
Very Mild
Healthy
Prediction Process

Fig 1. Proposal Model for The Identification of Alzheimer's Disease.

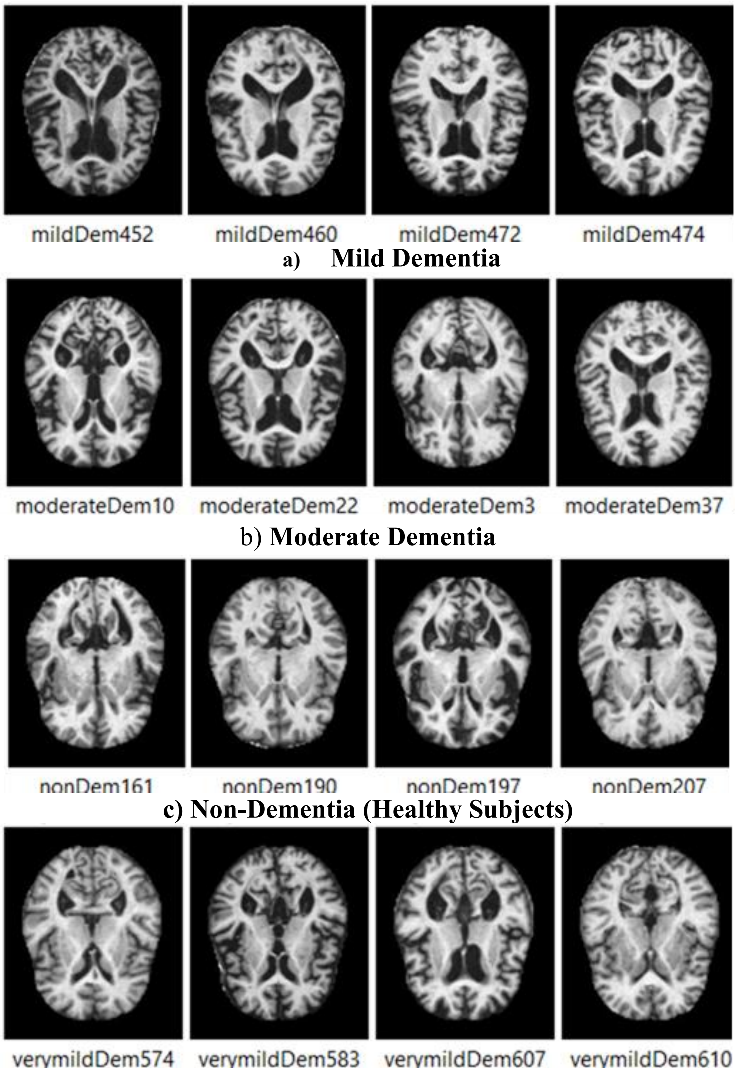
Dataset Description

The **Table 2** presents the AD dataset, which is classified into four classes: Mild Dementia (896 images), Moderate Dementia (64 images), non-dementia (1,200 images), and Very Mild Dementia (1,222 images). The Alzheimer's image dataset, sourced from Kaggle's datastore, comprises four distinct classes: Mild Dementia, Moderate Dementia, Non-Dementia, and Very Mild Dementia. Each class represents different stages of dementia, with 896 images labeled as Mild Dementia **Fig 1**, 64 images as Moderate Dementia, 1,200 images as non-dementia, and 1,222 images as Very Mild Dementia. This diverse representation is suitable for model training and supports experimental analysis in ML applications for Alzheimer's detection. **Fig 3** shows MRI brain images from four groups in the AD dataset. The groups are: (a) Mild

Dementia, (b) Moderate Dementia, (c) Non-Dementia (Healthy Subjects), and (d) Very Mild Dementia. The visual differences across classes aid in understanding structural brain changes at various dementia stages.

Table 2. Alzheimer's Disease Dataset Description

Class Code	Class Name	No. of Images
1	Mild Dementia	896
2	Moderate Dementia	64
3	Non-Dementia	1,200
4	Very Mild Dementia	1,222



a) **Mild Dementia**

b) **Moderate Dementia**

c) Non-Dementia (Healthy Subjects)

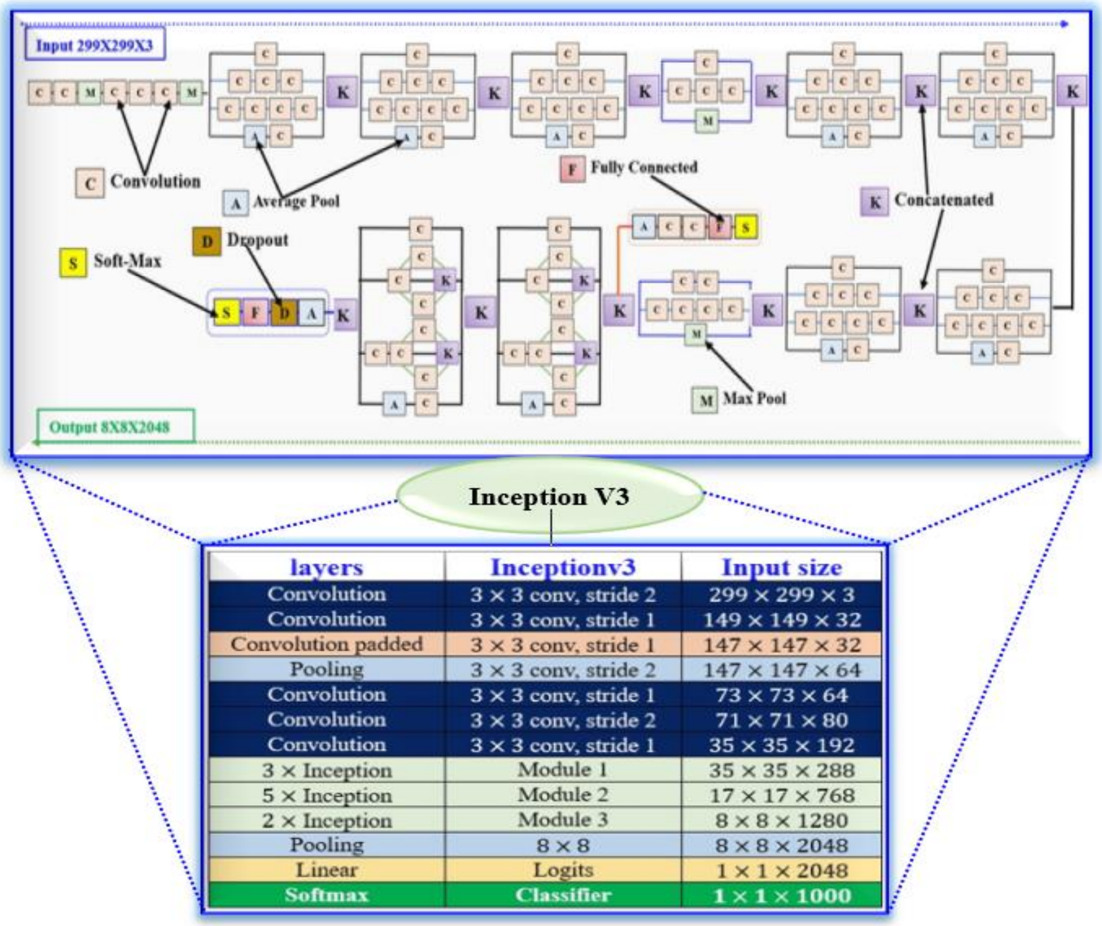
d) Very Mild Dementia

Fig 2. Sample Images of Alzheimer's Disease Experimental Image Dataset.

Inception V3

Inception V3 is a deep architecture that optimizes computational complexity, improves feature extraction, and reduces overfitting by using "Inception modules." It allows the network to process input data at multiple scales simultaneously, capturing fine-grained details and broader contextual information. In the research paper "Intelligent Dragon Fruit Detection System using Optimized Hybrid Deep Learning Models," Inception V3 is used as a powerful feature extractor for dragon fruit images, generating high-level feature representations that are fed into a hybrid classifier **Fig 2**.

The network uses 299x299 pixels RGB input images for transfer learning, processing them through operations to generate a 2048-dimensional feature vector from its penultimate layer. Inception V3 is a deep learning model that uses multiple modules to extract features from dragon fruit images. It uses a 2048-dimensional feature vector and MLP classifiers for high-accuracy classification. The model has 48 layers, and 23.8 million parameters optimized via 1x1 convolutions. Grok 3 aids in its development.



layers	Inceptionv3	Input size
Convolution	3 × 3 conv, stride 2	299 × 299 × 3
Convolution	3 × 3 conv, stride 1	149 × 149 × 32
Convolution padded	3 × 3 conv, stride 1	147 × 147 × 32
Pooling	3 × 3 conv, stride 2	147 × 147 × 64
Convolution	3 × 3 conv, stride 1	73 × 73 × 64
Convolution	3 × 3 conv, stride 2	71 × 71 × 80
Convolution	3 × 3 conv, stride 1	35 × 35 × 192
3 × Inception	Module 1	35 × 35 × 288
5 × Inception	Module 2	17 × 17 × 768
2 × Inception	Module 3	8 × 8 × 1280
Pooling	8 × 8	8 × 8 × 2048
Linear	Logits	1 × 1 × 2048
Softmax	Classifier	1 × 1 × 1000

Fig 3. Inception V3 Model Description Model Description.

Performance Parameters

Accuracy is a measure of model performance, comparing the ratio of correct predictions to total predictions. Precision evaluates the accuracy of positive predictions, focusing on the proportion of true positives. Recall measures the model's ability to identify all relevant instances. The F1-Score offers a balanced harmonic mean of precision and recall, providing a single metric for evaluating performance when trade-offs between the two are significant. Equations (1) to (4)

$$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)} \tag{1}$$

$$\Pr ecision = \frac{TP}{(TP+FP)} \tag{2}$$

$$\operatorname{Re} call = \frac{TP}{(TP+FN)} \tag{3}$$

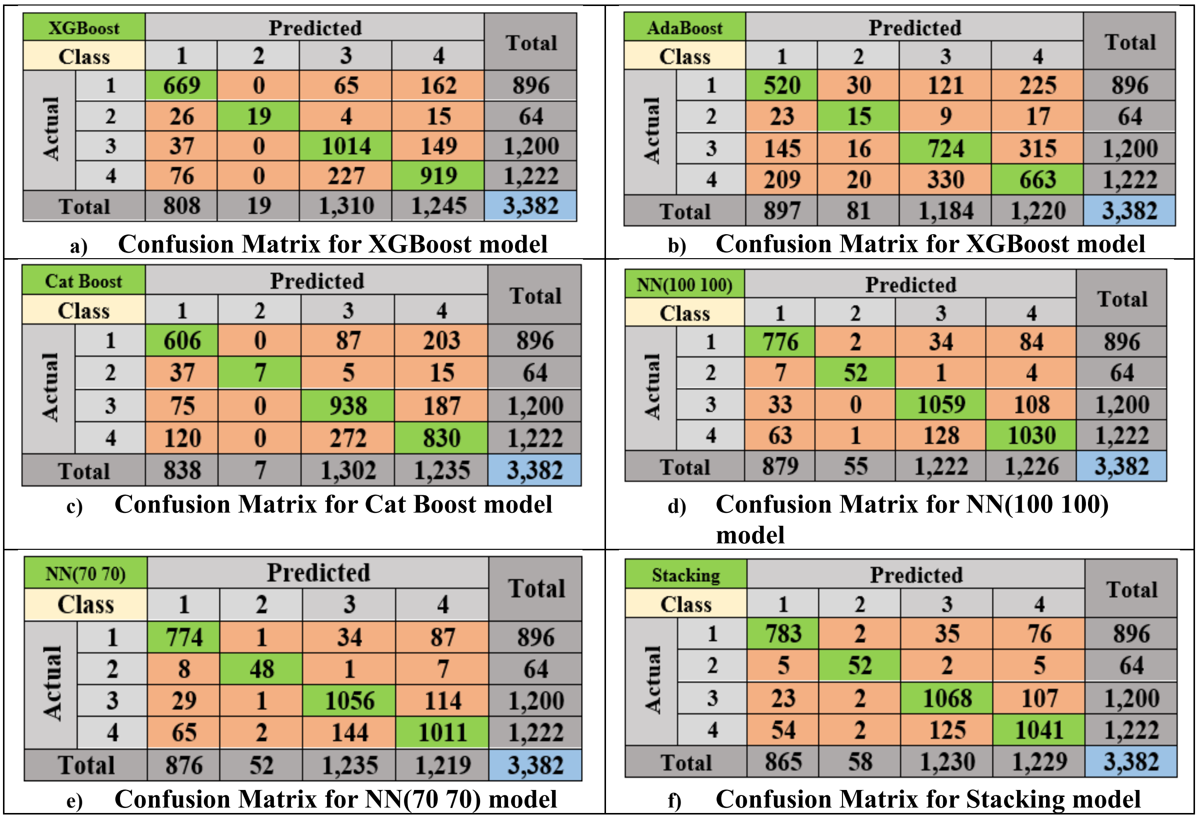
$$F1-Score = 2*\frac{(\operatorname{Re} call * \Pr ecision)}{(\operatorname{Re} call + \Pr ecision)} \tag{4}$$

IV. RESULT AND ANALYSIS

This section analyses a confusion matrix for six ML models. These models use Inception V3 features to classify Alzheimer's Disease. It compares their performance in identifying different stages of dementia. The analysis shows the strengths and weaknesses of each model.

Confusion Matrices Analysis for all ML Models for IV3 Features

The confusion matrix analysis for six ML models for classifying ADimages using Inception V3 features shows superior accuracy in Stacking, NN(100 100), and NN(70 70), particularly in recognizing Non-Demented and Mild Demented cases. XGBoost and CatBoost perform moderately well but struggle with overlapping classes like Moderate and Very Mild Dementia. AdaBoost, less accurate in earlier analyses, is likely outperformed by the ensemble and neural models.



XGBoost		Predicted				Total
Class		1	2	3	4	
Actual	1	669	0	65	162	896
	2	26	19	4	15	64
	3	37	0	1014	149	1,200
	4	76	0	227	919	1,222
Total		808	19	1,310	1,245	3,382

a) **Confusion Matrix for XGBoost model**

AdaBoost		Predicted				Total
Class		1	2	3	4	
Actual	1	520	30	121	225	896
	2	23	15	9	17	64
	3	145	16	724	315	1,200
	4	209	20	330	663	1,222
Total		897	81	1,184	1,220	3,382

b) **Confusion Matrix for XGBoost model**

Cat Boost		Predicted				Total
Class		1	2	3	4	
Actual	1	606	0	87	203	896
	2	37	7	5	15	64
	3	75	0	938	187	1,200
	4	120	0	272	830	1,222
Total		838	7	1,302	1,235	3,382

c) **Confusion Matrix for Cat Boost model**

NN(100 100)		Predicted				Total
Class		1	2	3	4	
Actual	1	776	2	34	84	896
	2	7	52	1	4	64
	3	33	0	1059	108	1,200
	4	63	1	128	1030	1,222
Total		879	55	1,222	1,226	3,382

d) **Confusion Matrix for NN(100 100) model**

NN(70 70)		Predicted				Total
Class		1	2	3	4	
Actual	1	774	1	34	87	896
	2	8	48	1	7	64
	3	29	1	1056	114	1,200
	4	65	2	144	1011	1,222
Total		876	52	1,235	1,219	3,382

e) **Confusion Matrix for NN(70 70) model**

Stacking		Predicted				Total
Class		1	2	3	4	
Actual	1	783	2	35	76	896
	2	5	52	2	5	64
	3	23	2	1068	107	1,200
	4	54	2	125	1041	1,222
Total		865	58	1,230	1,229	3,382

f) **Confusion Matrix for Stacking model**

Fig 4. Confusion Matrices ML Models for the Inception V3 Features of Alzheimer's Disease Images.

XGBoost ML Model for IV 3 Features of the AD Image Dataset

The model **(Fig 3(a))** had a high accuracy rate in Class 1 (639) and Class 2 (45), although it encountered difficulties in differentiating between early indicators of dementia and healthy patients. Class 3 (1014) had the best accuracy; nonetheless, it encountered difficulties in differentiating between early indicators of dementia and healthy patients. The model's performance was affected by class imbalance or nuanced variations. The model effectively identifies Mild Demented patients with an AUC of 0.949, indicating a balance between accuracy and recall. It has difficulties with mildly demented patients owing to class imbalance or feature overlap. The model excels at identifying non-demented persons, with few false negatives. The F1 score is 0.773, and the accuracy is 0.775. The model is conservative and has fewer false positives. It might miss some actual cases. Recommendations include fixing class imbalance. The stacking ensemble should be tuned to help underperforming classes. It is important to investigate confusion with ModerateDemented cases. Exploring more discriminatory features for ModerateDemented, possibly cognitive scores or structural imaging features **Fig 4**.

Table 3. Performance Parameters XGBoost ML Model for IV 3 Features AD Dataset Classes

Class	**AUC**	**CA**	**F1**	**Precision**	**Recall**
Mild_Demented	0.949	0.892	0.785	0.828	0.747
Moderate_Demented	0.989	0.987	0.458	1.000	0.297
Non_Demented	0.938	0.857	0.808	0.774	0.845
Very_Mild_Demented	0.898	0.814	0.745	0.738	0.752
All Over Classes	0.928	0.775	0.773	0.780	0.775

AdaBoost ML Model for IV 3 Features of The AD Image Dataset

Figure **(Fig 3(b))** indicates that categorizing Alzheimer's patients into three distinct classifications is a complicated endeavours. Class 1 is classified as slightly demented, exhibiting a significant incidence of misclassifications, mostly attributable to symptom overlap with early-stage Alzheimer's disease. Class 2 has mild dementia, characterized by a predominance of erroneous classifications **Table 3**. Class 3 is non-demented, exhibiting a significant number of misclassifications, which suggests challenges in differentiating healthy patients from those with very mild dementia. Class 4 exhibits mild dementia, with considerable overlap with Class 1 and Class 3, underscoring the persistent difficulty in differentiating early-stage symptoms from those of healthy persons. The AdaBoost classifier was tested for ADstages. It used metrics like AUC, Class Accuracy, F1 Score, Precision, and Recall. The model did well in identifying MildDemented cases. It had a high rate of correct predictions for this group. However, it struggled with ModerateDemented cases, showing low accuracy and precision. Non-demented cases had a fair accuracy of 72.3%. VeryMildDemented cases had a moderate success rate of 67.0%. Overall, the model performed only moderately in distinguishing dementia stages. It had a low-class Accuracy of 56.8%. Precision and recall were both around 0.570 **Table 4**.

Table 4. Performance Parameters AdaBoost ML Model for IV 3 Features AD Dataset Classes

Class	**AUC**	**CA**	**F1**	**Precision**	**Recall**
MildDemented	0.715	0.777	0.580	0.580	0.580
ModerateDemented	0.610	0.966	0.207	0.185	0.234
NonDemented	0.696	0.723	0.607	0.611	0.603
VeryMildDemented	0.644	0.670	0.543	0.543	0.543
All Classes	0.681	0.568	0.569	0.570	0.568

Cat Boost ML Model for IV 3 Features of the AD Image Dataset

The Cat Boost model performs well in identifying NonDemented and VeryMildDemented cases. It made 938 correct predictions for NonDemented and 830 for VeryMildDemented. MildDemented had 606 correct classifications, but many were misclassified as NonDemented or VeryMildDemented. ModerateDemented had very few correct predictions, only 7. The model struggles to detect ModerateDemented cases. The classification favors majority classes, suggesting a need for better handling of class imbalance. The classification model was tested across four ADcategories: MildDemented, ModerateDemented, NonDemented, and VeryMildDemented. The model effectively distinguished MildDemented cases with a high accuracy rate of 84.6%. However, its high accuracy may be misleading due to class imbalance. The model showed excellent separation ability but poor accuracy due to class imbalance. NonDemented showed good discriminatory power with a high accuracy rate of 81.5%, identifying 78.2% of actual samples. VeryMildDemented had a moderate capability with a high accuracy rate of 76.4%. Overall, the model's precision and recall were consistent across all classes **Table 5**.

Table 5. Performance Parameters Cat Boost ML Model for IV 3 Features AD Dataset Classes

Class	**AUC**	**CA**	**F1**	**Precision**	**Recall**
MildDemented	0.912	0.846	0.699	0.723	0.676
ModerateDemented	0.979	0.983	0.197	1.000	0.109
NonDemented	0.894	0.815	0.750	0.720	0.782
VeryMildDemented	0.837	0.764	0.676	0.672	0.679
All Classes	0.881	0.704	0.699	0.709	0.704

MLP(100 100) ML Model for IV 3 Features of The AD Image Dataset

The Neural Network (100 100) model accurately identified Non-Demented and Very Mild Demented cases, achieving 1,059 and 1,030 correct predictions, respectively. Mild Demented had 776 correctly classified cases, although some were confused with Class 4. Moderate Demented had 52 out of 64 accurately identified cases. Misclassifications mainly occurred between neighbouring dementia stages, particularly Classes 3 and 4. Overall, the model shows strong classification capability. The model performs very well in classifying different types of dementia. It has high AUC scores, which range from 0.952 to 0.996. This shows that it can distinguish between classes effectively. The ModerateDemented group has the best classification accuracy at 0.996. It also has a strong F1-score of 0.874, which balances precision and recall well. The MildDemented and Nondemented groups also have the same F1-score of 0.874, with high precision and recall values. VeryMildDemented performed slightly lower but maintained robust metrics with an F1-score of 0.842. The average performance across all classes (AUC: 0.967, CA: 0.863, F1: 0.863) demonstrates the model's strong and reliable prediction capability **Table 6**.

Table 6. Performance Parameters MLP(100 100) ML Model for IV 3 Features AD Dataset Classes

Class	AUC	CA	F1	Precision	Recall
MildDemented	0.981	0.934	0.874	0.883	0.866
ModerateDemented	0.996	0.996	0.874	0.945	0.812
NonDemented	0.970	0.910	0.874	0.867	0.882
VeryMildDemented	0.952	0.885	0.842	0.840	0.843
All Classes	0.967	0.863	0.863	0.863	0.863

MLP(70 70) ML Model for IV 3 Features of The AD Image Dataset

The confusion matrix for the NN(70 70) model shows promising results. Class 1, MildDemented, had 774 correct predictions out of 896. Class 3, NonDemented, also did well with 1,056 correct predictions and few mistakes. Class 4, VeryMildDemented, confused some instances with Class 3, misclassifying 144 times, but still had 1,011 correct predictions. Class 2 (ModerateDemented) had a small sample size and showed 48 correct predictions, with minor misclassification to other classes. The model performs very well in classifying ADcategories. For MildDemented, it has a high F1-score of 0.874. It shows good consistency and balance precision-recallll. Moderate Demented has a smaller class size but still performs impressively. It has a near-perfect AUC of 0.997 and an intense precision of 0.923. NonDemented and VeryMildDemented also perform well, each with F1-scores above 0.82. The average AUC across all classes is 0.962, indicating the model's strong discriminative ability. Overall, the model is reliable and well-generalized for multi-class classification **Table 7**.

Table 7. Performance Parameters MLP(70 70) ML Model for IV 3 Features AD Dataset Classes

Class	AUC	CA	F1	Precision	Recall
MildDemented	0.978	0.934	0.874	0.884	0.864
ModerateDemented	0.997	0.994	0.828	0.923	0.750
NonDemented	0.965	0.904	0.867	0.855	0.880
VeryMildDemented	0.943	0.876	0.828	0.829	0.827
All Classes	0.962	0.854	0.854	0.855	0.854

Stacking ML Model for IV 3 Features of the AD Image Dataset

The Stacking model shows good performance in classifying Alzheimer's disease. For Class 1, MildDemented, the model made 783 correct predictions out of 896. It indicates high precision and few mistakes. In Class 2, ModerateDemented, there were 52 correct predictions. Only a few were confused with other classes. Class 3, NonDemented, had 1,068 correct identifications. It shows that the model works well for this group. Class 4, VeryMildDemented, achieved 1,041 correct predictions out of 1,222, indicating strong class-wise recall. The model is very good at classifying Alzheimer's disease. It has a strong F1-score of 0.889. The AUC is also high at 0.972. The moderate-demented class has perfect accuracy and precision. The non-demented and very-demented classes perform consistently well. The model is robust in identifying Alzheimer's disease. The model's robustness in ADidentification is confirmed by its excellent average AUC of 0.959 and F1-score of 0.871 **Table 8**.

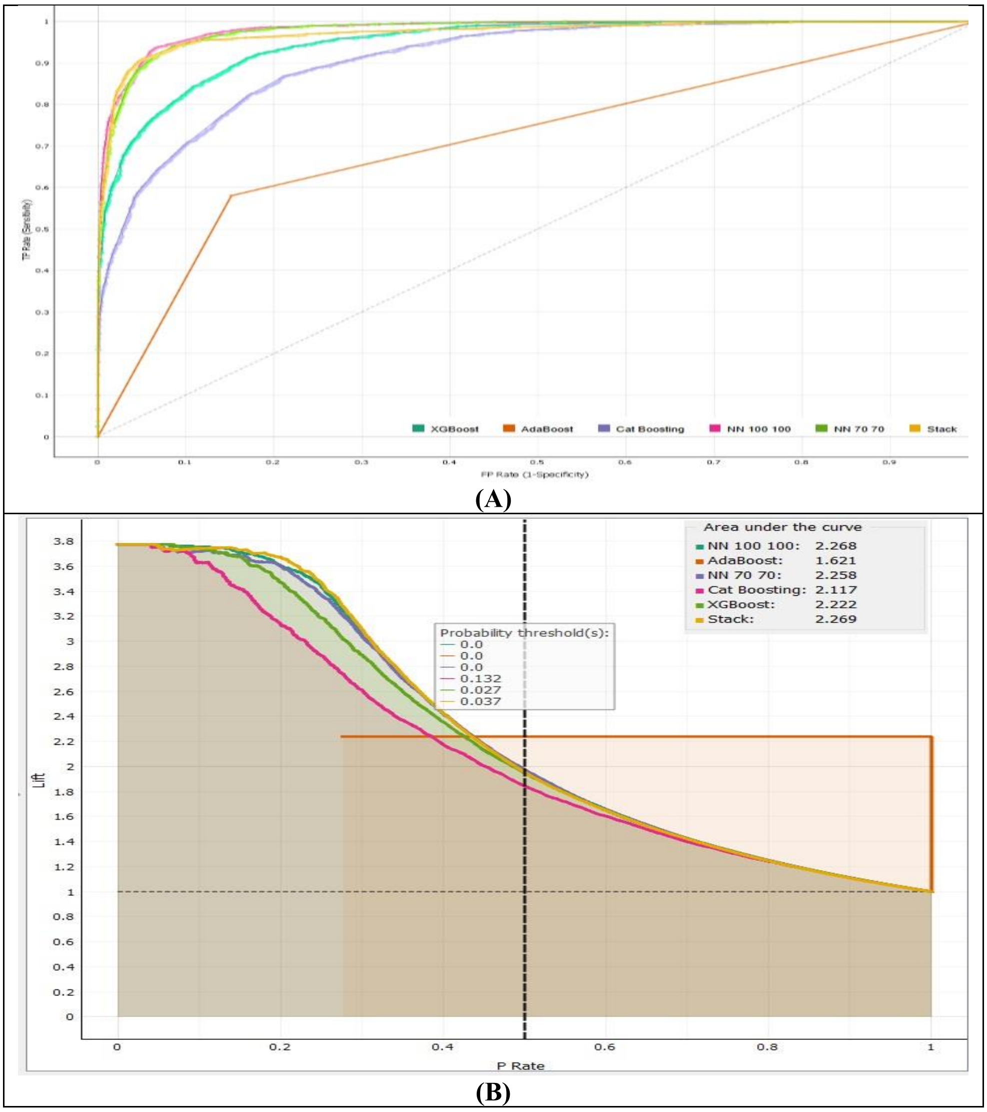
Table 8. Performance Parameters MLP(70 70) ML Model for IV 3 Features AD Dataset Classes

Class	AUC	CA	F1	Precision	Recall
MildDemented	0.972	0.942	0.889	0.905	0.874
ModerateDemented	0.996	0.996	0.874	0.945	0.812
NonDemented	0.963	0.913	0.879	0.868	0.890
VeryMildDemented	0.946	0.891	0.849	0.847	0.852
All Classes	0.959	0.870	0.871	0.871	0.870

ROC-AUC and Performance Curves Analysis for IV 3 Features of The AD Image Dataset

Fig 5 (A) shows several ROC curves. These curves represent how different ML models perform on an ADimage dataset. Each model has a specific color. Dark Green is for XGBoost, Light Brown for AdaBoost, Purple for CatBoost, Violet for a Neural Network with two hidden layers of 100 neurons, Light Green for a Neural Network with two hidden layers of 70 neurons, and Orange for the Stacking Ensemble model.

There is a legend to match colors with models. A dashed diagonal line indicates the performance of a random classifier with an AUC of 0.5. Models above this line perform better than random classifiers. The ROC-AUC values measure how well different models perform on a dataset. The largest neural network, with layers of 100 and 100, has the highest AUC of 0.967. This means that it is the best at distinguishing between classes. The smaller network, with layers of 70 and 70, also performs well. The stacking ensemble model has a high AUC of 0.959, showing it works well by combining predictions from multiple base models. XGBoost has a good AUC of 0.928, while CatBoost has the lowest AUC of 0.881.



(A)

(B)

Fig 5. (A) ML ROC Curves ML Model for IV 3 Features AD Dataset (B) ML Lift Curves ML Model for IV 3 Features AD Dataset.

Lift curves (**Fig 5 (B))** help evaluate ML models, especially with imbalanced data. They show how well a model finds positive cases compared to random guessing. Higher lift values mean better performance. The area under the curve measures this improvement. The probability threshold shows how confident the predictions are. The Stacking model has the highest lift value of 2.269. It finds more than twice the true positives compared to random selection. Its probability threshold is 0.037, showing it makes confident predictions at a low threshold. The Neural Network model has a lift of 2.268 and shows high confidence in positive predictions. Another Neural Network model has a lift of 2.258 and is efficient in early detection**.** XGBoost had a lift of 2.222 with a threshold of 0.027. This means it balanced early predictions and precision well. CatBoost achieved a lift of 2.117 with a higher threshold of 0.132. This suggests it predicts positive cases confidently only at higher probabilities, which may lower false positives. AdaBoost performed the worst with a lift of 1.621 and started at a threshold of 0.0. This shows it struggled to distinguish positives from random guessing. The analysis indicates that ensemble models like Stacking and deep neural networks perform better in detecting Alzheimer's stages when using Inception V3 features.

CRediT Author Statement

The authors confirm contribution to the paper as follows:
Conceptualization: Andhavarapu Tejtha, Ravi Kumar T, Panduranga Vital Terlapu, Ramkishor Pondreti and Suneel Gowtham Karudumpa; **Methodology:** Andhavarapu Tejtha and Ravi Kumar T; **Writing- Original Draft Preparation:** Andhavarapu Tejtha, Ravi Kumar T, Panduranga Vital Terlapu, Ramkishor Pondreti and Suneel Gowtham Karudumpa; **Supervision:** Panduranga Vital Terlapu, Ramkishor Pondreti and Suneel Gowtham Karudumpa; **Validation:** Andhavarapu Tejtha and Ravi Kumar T; **Writing- Reviewing and Editing:** Andhavarapu Tejtha, Ravi

Kumar T, Panduranga Vital Terlapu, Ramkishor Pondreti and Suneel Gowtham Karudumpa; All authors reviewed the results and approved the final version of the manuscript.

Data availability statement:

No data were used in this research.

Conflict of Interest:

There is no potential conflict of interest was reported by the authors.

Funding Statement:

This research is not funded by any government or private bodies.

Competing Interests

There are no competing interests

References

[1]. E. Mahamud, M. Assaduzzaman, J. Islam, N. Fahad, M. J. Hossen, and T. T. Ramanathan, “Enhancing Alzheimer’s disease detection: An explainable machine learning approach with ensemble techniques,” Intelligence-Based Medicine, vol. 11, p. 100240, 2025, doi: 10.1016/j.ibmed.2025.100240.

[2]. O. Topsakal and S. Lenkala, “Enhancing Alzheimer’s Disease Detection through Ensemble Learning of Fine-Tuned Pre-Trained Neural Networks,” Electronics, vol. 13, no. 17, p. 3452, Aug. 2024, doi: 10.3390/electronics13173452.

[3]. Jenber Belay, Y. M. Walle, and M. B. Haile, “Deep Ensemble learning and quantum machine learning approach for Alzheimer’s disease detection,” Scientific Reports, vol. 14, no. 1, Jun. 2024, doi: 10.1038/s41598-024-61452-1.

[4]. Nasir, N., Ahmed, M., Afreen, N., & Sameer, M. (2024). Alzheimer's Magnetic Resonance Imaging Classification Using Deep and Meta-Learning Models. arXiv preprint arXiv:2405.12126.

[5]. M. M. Rana et al., “A robust and clinically applicable deep learning model for early detection of Alzheimer’s,” IET Image Processing, vol. 17, no. 14, pp. 3959–3975, Sep. 2023, doi: 10.1049/ipr2.12910.

[6]. M. Mujahid, A. Rehman, T. Alam, F. S. Alamri, S. M. Fati, and T. Saba, “An Efficient Ensemble Approach for Alzheimer’s Disease Detection Using an Adaptive Synthetic Technique and Deep Learning,” Diagnostics, vol. 13, no. 15, p. 2489, Jul. 2023, doi: 10.3390/diagnostics13152489.

[7]. Kommu Bhushanm, T. Venkata Gayathri, S. Reshma, V. Nagi Reddy, and J. Naveen Reddy, “Novel Inception V3 Deep Learning Model Designing for Alzheimer’s Disease Detection,” International Journal of Scientific Research in Science, Engineering and Technology, pp. 08–15, Feb. 2023, doi: 10.32628/ijsrset2310154.

[8]. S. Alatrany, W. Khan, A. J. Hussain, J. Mustafina, and D. Al-Jumeily, “Transfer Learning for Classification of Alzheimer’s Disease Based on Genome Wide Data,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 20, no. 5, pp. 2700–2711, Sep. 2023, doi: 10.1109/tcbb.2022.3233869.

[9]. S. Sharma et al., “Transfer learning-based modified inception model for the diagnosis of Alzheimer’s disease,” Frontiers in Computational Neuroscience, vol. 16, Nov. 2022, doi: 10.3389/fncom.2022.1000435.

[10]. D. Agarwal, G. Marques, I. de la Torre-Díez, M. A. Franco Martin, B. García Zapiraín, and F. Martín Rodríguez, “Transfer Learning for Alzheimer’s Disease through Neuroimaging Biomarkers: A Systematic Review,” Sensors, vol. 21, no. 21, p. 7259, Oct. 2021, doi: 10.3390/s21217259.

[11]. H. A. Helaly, M. Badawy, and A. Y. Haikal, “Deep Learning Approach for Early Detection of Alzheimer’s Disease,” Cognitive Computation, vol. 14, no. 5, pp. 1711–1727, Nov. 2021, doi: 10.1007/s12559-021-09946-2.

[12]. S. U. Sadat, H. H. Shomee, A. Awwal, S. N. Amin, M. T. Reza, and M. Z. Parvez, “Alzheimer’s Disease Detection and Classification using Transfer Learning Technique and Ensemble on Convolutional Neural Networks,” 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 1478–1481, Oct. 2021, doi: 10.1109/smc52423.2021.9659179.

[13]. R. Jansi, N. Gowtham, S. Ramachandran, and V. S. Praneeth, “Revolutionizing Alzheimer’s Disease Prediction using InceptionV3 in Deep Learning,” 2023 7th International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 1155–1160, Nov. 2023, doi: 10.1109/iceca58529.2023.10395534.

[14]. M.-U.-I. Tamim, S. Malik, S. G. Sneha, S. M. H. Mahmud, K. O. M. Goh, and D. Nandi, “Predicting Different Classes of Alzheimer’s Disease using Transfer Learning and Ensemble Classifier,” JOIV : International Journal on Informatics Visualization, vol. 8, no. 4, p. 2452, Dec. 2024, doi: 10.62527/joiv.8.4.3038.