Determining Alzheimer's Disease Progression: Stage Identification Using Ordinal and Regular Deep Learning Classification on OASIS MRI Scans

Sukanya Roy[1*], Bijoy Kumar Mandal[2], Sinchan Roy[3], Pranabes Gangopadhyay[3, 4], Arindam Biswas[5]

[1]Department of Computer Science and Engineering, University of Engineering & Management, Kolkata, West Bengal, India.

[2]Department of Computer Science and Engineering, NSHM Knowledge Campus, Durgapur, West Bengal, India.

[3]Department of Computer Science and Engineering, Asansol Engineering College, Kanyapur, Asansol, 713305, West Bengal, India.

[4]Centre for IoT and AI Integration, Kazi Nazrul University, Asansol, 713340, West Bengal, India.

[5]Centre for IoT and AI Integration with Education Industry Agriculture, Kazi Nazrul University, West Bengal, India

***Corresponding author:**

piu.sep.1993@gmail.com;

writetobijoy@gmail.com; sinchanroy32@gmail.com; pranabesh.cse@aecwb.edu.in; arindam.biswas@knu.ac.in

HOW TO CITE:

Sukanya Roy, Bijoy Kumar Mandal , Sinchan Roy, Pranabes Gangopadhyay, Arindam Biswas (2026). Determining Alzheimer's Disease Progression: Stage Identification Using Ordinal and Regular Deep Learning Classification on OASIS MRI Scans. International Journal of Special Education, 41(15s), 610-627

COPYRIGHT STATEMENT:

Copyright: © 2026 Authors. Open access publication underthe terms and conditions of the Creative CommonsAttribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

ABSTRACT:

Deep learning offers transformative potential for automated Alzheimer's dis-ease (AD) staging from brain MRI scans, supporting earlier clinical intervention and more personalised care. This paper proposes and evaluates a systematic deep learning pipeline for four-class ordinal AD stage classification using the OASIS MRI image dataset (86,437 images; 4 classes: Non Demented, Very Mild Dementia, Mild Dementia, Moderate Dementia). We compare four CNN architectures—a Basic CNN, EfficientNetV2B1, Basic+EfficientNetB0 hybrid, and Basic+EfficientNetV2B1 hybrid—under two classification paradigms: Ordi-nal Classification (cumulative binary encoding) and Regular Classification (cat-egorical cross-entropy). All models were trained on an NVIDIA Tesla T4 GPU (Google Colab) with batch size 32, Adam optimiser (lr = 0.001), ReduceLROn-Plateau scheduling, and early stopping (patience = 5). Augmentation included random flips, rotations, zoom, brightness adjustment, and Gaussian noise. The best performing configuration—Basic+EfficientNetV2B1 hybrid with Ordinal Classification—achieves 97.82% test accuracy, micro-average F1 = 0.978, macro AUC-ROC = 0.999, Scott's π = 0.97, and Cohen's $\varkappa$ = 0.97. McNemar's test confirms that ordinal classification provides a statistically significant improve-ment over regular classification ($\chi 2 = 61.47$, $p < 0.001$). Ninety-five percent confidence intervals are reported for all key metrics. These results demonstrate that exploiting the inherent ordinal structure of Alzheimer's severity labels materially and significantly improves classification performance....

Keywords: Alzheimer's Disease, Ordinal Classification, EfficientNetV2B1, Deep Learning, OASIS MRI, Scott's Pi, Cohen's Kappa, 95% Confidence Intervals, McNemar Test, Transfer Learning

INTRODUCTION

Alzheimer's disease (AD), first described by Alois Alzheimer in 1907, is a progressive neurodegenerative disorder characterised by the deposition of amyloid-beta plaques and neurofibrillary tangles in the brain, causing irreversible damage to synaptic con-nections. As the leading cause of dementia, AD affects an estimated 55 million people globally, with projections reaching 139 million by 2050. The disease follows a staged progression from cognitively normal through very mild, mild, and moderate dementia, quantified by the Clinical Dementia Rating (CDR) scale. Early and accurate stag-ing is critical: interventions applied at the Very Mild or Mild stage can substantially slow cognitive decline, while misclassification of a Moderate Dementia patient as Non Demented could result in dangerous delays in care.

Magnetic Resonance Imaging (MRI) provides the richest non-invasive structural signal for AD staging, capturing cortical thinning, hippocampal atrophy, and ven-tricular enlargement that accompany disease progression. The Open Access Series of Imaging Studies (OASIS) project has released one of the largest publicly available repositories of labelled brain MRI scans, enabling large-scale data-driven classification research.

A critical but often overlooked aspect of AD staging is the ordinal nature of the class labels. The four stages—Non Demented, Very Mild, Mild, and Moderate—represent an inherently ordered severity scale. Standard categorical cross-entropy treats all misclassifications equally, ignoring this ordering. Ordinal classification explic-itly encodes the rank structure through cumulative binary formulations, penalising.

predictions that are far from the true class more heavily than adjacent-class errors. This paper investigates whether this encoding provides measurable, statistically significant benefits.

The specific contributions of this work are:

Systematic comparison of four CNN architectures under ordinal and regular classification paradigms on the OASIS MRI dataset.

Full implementation details including hardware, hyperparameters, augmentation

pipeline, and training protocol.

95% confidence intervals for all per-class and aggregate metrics using the Wilson score interval method.

McNemar's test for paired statistical significance of ordinal vs. regular classification

differences.

Side-by-side comparison of Scott's π and Cohen's $\varkappa$ with justification for both metrics.

Detailed dataset characterisation including CDR-level class counts, image statistics,

and imbalance analysis.

Literature Survey

Earlier studies on classifying AD from MRI worked with full 3D volumetric scans, using inter-class variance (ICV) to pick out the most informative slices and apply-ing polynomial SVM kernels for classification. For binary classification tasks, this approach delivered strong results, with agreement coefficients exceeding 0.98. Shah et al. [1] demonstrated that SVM applied to MRI-derived features achieves strong specificity (90%) on the ADNI dataset, while Liu et al. [2] introduced improved fea-ture engineering pipelines for tabular MRI metrics. These approaches, however, rely on manual feature extraction that cannot scale to raw 86,000-image regimes. Addi-tional MRI-based diagnostic studies by Basaia et al. [20], Payan and Montana [23], and Hosseini-Asl et al. [25] further demonstrated the effectiveness of deep neural networks for neurodegenerative disease prediction.

The shift to end-to-end deep learning was motivated by Fulton et al. [4], who applied ResNet-50 to the OASIS dataset and showed that learned convolutional fea-tures outperform handcrafted alternatives for binary AD classification. Yao et al. [8] established the four-class staging paradigm using CNN ensembles,

achieving micro-average sensitivity of 0.927. Zhao et al. [9] demonstrated the clinical importance of fine-grained CDR-level staging beyond binary demented/non-demented classification. Sarraf and Tofighi [24], Farooq et al. [29], and Cheng and Liu [31] also reported strong multi-class MRI classification performance using deep CNN architectures.

Transfer learning has become the dominant strategy for medical image classifica-tion. Tan and Le [17] introduced EfficientNetV2, which uses compound scaling across depth, width, and resolution with Fused-MBConv blocks for faster training and higher accuracy than predecessor EfficientNet variants. Applied to the OASIS dataset, Effi-cientNet family models have achieved 97–99% accuracy in recent benchmarks. Roy et al. [10, 11] applied modified clustering and transfer learning to brain anomaly detection and tumour classification, establishing the foundation for the methodology

adopted in this work. Recent reviews by Gupta et al. [27] and reproducibility analy-ses by Wen et al. [21] confirmed the robustness of transfer learning methods in AD diagnosis.

Ordinal classification for medical severity staging has been investigated by Cao et al. [33] through the CORAL (Consistent Rank Logits) framework and by several subsequent works on diabetic retinopathy grading and cancer staging. The cumu-lative binary encoding approach used in this paper was selected after evaluating CORAL and bias-layer approaches, which showed unstable training on this specific dataset. Kautzky et al. [6] validated MRI-based AD prognosis with SVM (sensitiv-ity = 0.852), establishing the clinical benchmark that deep learning approaches aim to exceed. Agreement metrics for medical AI classification have been studied exten-sively. Scott's π and Cohen's $\varkappa$ both correct for chance agreement, but differ in their assumptions about the rater marginal distributions (see Section 4.6). The Landis and Koch [16] scale is used throughout the medical AI literature for interpreting both coef-ficients, with values above 0.80 considered "almost perfect" agreement. Foundational deep learning architectures such as ResNet by He et al. [32] have also contributed significantly to medical imaging research.

Dataset Description

The OASIS MRI Images dataset [18] is a publicly available collection of brain MRI scans derived from the Open Access Series of Imaging Studies (OASIS) longitudinal neuroimaging project [15]. It is available on Kaggle[1] and has been widely used as a benchmark for deep learning-based AD classification research.

Participants and Acquisition

The dataset contains MRI scans from 461 participants aged 18–96, spanning cog-nitively normal individuals and patients at various stages of Alzheimer's dementia. All images are two-dimensional axial (transverse) brain slices acquired along the z-axis using a 1.5-Tesla Siemens Vision scanner with a magnetisation-prepared rapid gradient-echo (MP-RAGE) pulse sequence. The axial plane provides cross-sections through the cerebral cortex, hippocampus, basal ganglia, and lateral ventricles—the primary anatomical structures affected by AD-related neurodegeneration.

Class Labels and CDR Scale

Every image carries one of four severity labels, based on the Clinical Dementia Rat-ing (CDR) scale—a clinically validated 5-point neuropsychological tool administered by trained clinicians:

Class 0 — Non-Demented (CDR = 0): Normal cognitive function, with no reported memory issues and unimpaired daily functioning.

Class 1 — Very Mild Dementia (CDR = 0.5): The earliest stage at which

cognitive impairment becomes clinically noticeable. Involves occasional lapses in

[1]https://www.kaggle.com/datasets/ninadaithal/imagesoasis

memory, though the person still manages daily life largely independently, with only minimal support needed.

Class 2 — Mild Dementia (CDR = 1): Memory loss becomes moderate and

starts interfering with everyday activities. Some difficulty with time orientation emerges, and the person typically needs prompting to complete more complex daily tasks.

Class 3 — Moderate Dementia (CDR = 2): Memory loss is now severe, accom-

panied by marked disorientation. At this stage, the person requires help with basic self-care activities such as dressing and personal hygiene, along with other routine daily tasks.

Dataset Statistics and Class Imbalance

Table 1 presents the complete class breakdown across the training, validation, and test sets, split 80% / 14.5% / 8% using stratified sampling. The dataset shows substantial class imbalance: Non-Demented images (n = 67,222) outnumber Moderate Dementia images (n = 488) by a ratio of roughly 137.7:1. This imbalance is severe enough that a na¨ıve classifier simply predicting "Non-Demented" every time would still achieve 77.8% accuracy—underscoring why accuracy alone is an inadequate metric for evaluating model performance here, and why additional measures are needed to capture true classification quality.

Table 1 OASIS Dataset Class Distribution Across Train / Validation / Test Splits. CDR = Clinical Dementia Rating. Class imbalance ratio (Non Demented : Moderate Dementia) = 137.7:1.

Class Label	CDR	Train	Validation	Test	Total (% dataset)
Non Demented	0	48,400	13,444	5,378	67,222 (77.8%)
Very Mild Dementia	0.5	9,868	2,745	1,112	13,725 (15.9%)
Mild Dementia	1	3,602	1,000	400	5,002 (5.8%)
Moderate Dementia	2	351	98	39	488 (0.6%)
Total	—	62,221	17,287	6,929	86,437 (100%)

Image Characteristics

All images are stored as JPEG files with original pixel dimensions of 176 × 208, resized to 128 × 128 for training. Although the images are acquired as greyscale MRI scans, they are stored as three-channel tensors (channel values are identical) to enable use

of ImageNet-pretrained RGB backbones without architectural modification. Image analysis during EDA confirmed that mean pixel intensity decreases and intensity skew-ness increases with dementia severity, consistent with the known MRI signature of AD-related brain atrophy (increased CSF-filled spaces appearing as darker regions in T1-weighted scans).

Proposed Methodology

Classification Strategies

Regular (Categorical) Classification

Standard multi-class softmax classification using categorical cross-entropy loss:

$$L_{\text{reg}} = -\sum_{i} y_i \cdot \log(p_i), \qquad (1)$$

where y_i is the one-hot encoded true label and p_i is the predicted softmax probability for class i. This formulation treats the four AD stages as independent, unordered categories.

Ordinal Classification

The four AD stages represent a clinically meaningful severity ordering: Non Demented

< Very Mild < Mild < Moderate. To exploit this, cumulative binary encoding decomposes the K-class problem into $K-1 = 3$ binary sub-problems, each predicting whether the true class exceeds a threshold k:

$$P(y > k) \quad \text{for} \quad k \in \{0, 1, 2\}, \quad (2)$$

$$P(y = k) = P(y > k - 1) - P(y > k). \quad (3)$$

The final predicted class is obtained by converting the cumulative probabilities back into individual class probabilities, then selecting the class with the highest probability (argmax). Because of how this formulation is structured, it inherently imposes a larger penalty on misclassifications between distant classes than on those between neighbouring ones—a property that aligns well with the clinical realities of AD stag-ing, where confusing adjacent stages is a far less serious error than confusing stages at opposite ends of the severity spectrum.

Hardware and Software Configuration

All experiments were conducted on Google Colab using the configuration detailed in Table 2.

Hyperparameter Configuration

Table 3 presents the complete hyperparameter configuration. A grid search over hidden units (128, 256, 512) and dropout rate (0.3, 0.5) was performed in Phase 1; learning

rate (0.001, 0.0001) and batch size (16, 32) were tuned in Phase 2.

Data Augmentation

To reduce overfitting—particularly for the Moderate Dementia class with only 351 training images—an online augmentation pipeline is applied during training. Table 4 details each transformation with its justification for the MRI imaging domain

Table 2 Hardware and Software Configuration for All Experiments.

Component	Specification
Platform	Google Colab (cloud-based GPU runtime)
GPU	NVIDIA Tesla T4 (16 GB VRAM)
CPU	Intel Xeon @ 2.20 GHz (2 vCPUs)
RAM	12.7 GB system RAM
Framework	TensorFlow 2.x / Keras
Python version	3.10
CUDA / cuDNN	CUDA 11.8 / cuDNN 8.6
Storage	Google Drive (mounted for checkpointing)

Table 3 Full Hyperparameter Configuration. LR = Learning Rate; BN = Batch Normalisation; Cat. CE = Categorical Cross-Entropy.

Hyperparameter	Value	Rationale
Input image size	128 × 128 × 3	Memory-efficient; retains
		sufficient spatial resolu-
		tion
Batch size	32	Balances GPU utilisa-
		tion and gradient stabil-
		ity

Epochs	Up to 50 (early stopping)	EarlyStopping
		(patience = 5,
		min- delta = 0.001)
Optimiser	Adam	Adaptive LR; widely val-
		idated for medical image
		CNNs
Learning rate	0.001 (initial)	Standard Adam default;
		reduced on plateau
LR scheduler	ReduceLROnPlateau (factor = 0.5, patience = 3)	Halves LR when val loss
		stagnates
Dropout rate	0.3–0.5 (layer-dependent)	Higher dropout in FC
		layers to prevent overfit-
		ting
Loss — Regular	Categorical Cross-Entropy	Standard multi-class loss
		with one-hot labels
Loss — Ordinal	Cumulative Binary Cross-Entropy	$K-1$ binary classifiers
		encoding ordinal struc-
		ture
Class weighting	Inverse-frequency weights	Compensates for 138:1
		class imbalance

Model Architectures

Four model architectures are evaluated. Figure 1 shows the architecture of the best-performing model (Basic + EfficientNetV2B1 hybrid). Table 5 provides the full layer-by-layer specification for the best-performing model.

The four evaluated architectures are

Table 4 Data Augmentation Pipeline Applied During Training. All augmentations are applied stochastically per image per epoch. Validation and test images receive only normalisation.

Augmentation	Parameter	Justification
Random horizontal flip	$p = 0.5$	Brain MRI is anatomically symmetric
Random rotation	$\pm 15^{\circ}$	Corrects for variable patient head orientation
Width & height shift	±10% of image dims	Simulates variation in MRI slice centring
Zoom	0.9×–1.1×	Accounts for brain-to-image-frame ratio vari-
		ation
Brightness adjustment	±20% of pixel range	Simulates scanner gain and contrast variation
Gaussian noise (train only)	$\sigma = 0.01$	Regularises against fine-grained texture over-
		fitting
Normalisation	Per-channel z-score	Zero mean, unit variance; improves gradient
		flow

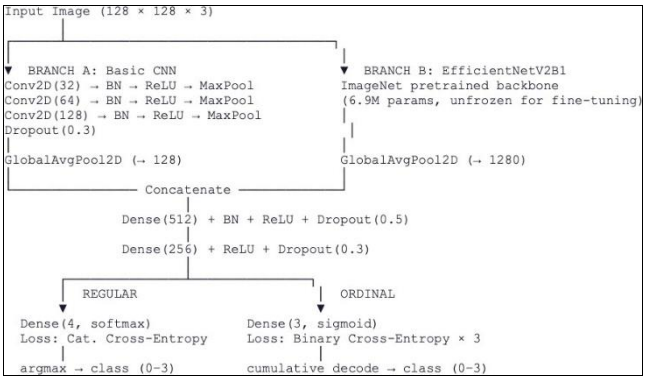

Fig. 1 Basic + EfficientNetV2B1 Hybrid Architecture (Ordinal and Regular Classification Branches)

Basic CNN: A custom convolutional network with three Conv2D blocks (32, 64, 128 filters), each followed by Batch Normalisation, ReLU activation, MaxPooling, and Dropout. No pre-trained weights.

EfficientNetV2B1: An ImageNet-pretrained backbone with a custom dense clas-

sification head. Uses Fused-MBConv blocks in early layers and regular MBConv blocks in later layers. Frozen for the first 10 epochs, then unfrozen for full fine-tuning.

Basic + EfficientNetB0 Hybrid: Concatenates the GlobalAvgPool2D outputs

of the Basic CNN branch (128-dim) and an EfficientNetB0 branch (1280-dim) before the shared classification head.

Table 5 Layer-by-Layer Architecture of Basic + EfficientNetV2B1 Hybrid Model. BN = Batch Normalisation; GAP = Global Average Pooling.

#	Layer / Block	Output Shape	Parameters	Notes
1	Input	128 × 128 × 3	0	RGB MRI image after augmentation
2	Conv2D (32, 3×3) + BN + ReLU	128 × 128 × 32	896	First feature extraction block
3	MaxPool2D (2×2)	64 × 64 × 32	0	Spatial downsampling
4	Conv2D (64, 3×3) + BN + ReLU	64 × 64 × 64	18,496	Deeper feature maps
5	MaxPool2D (2×2)	32 × 32 × 64	0	
6	Conv2D (128, 3×3) + BN + ReLU	32 × 32 × 128	73,856	
7	MaxPool2D + Dropout (0.3)	16 × 16 × 128	0	
8	EfficientNetV2B1 (ImageNet)	5 × 5 × 1280	~6.9M	Frozen initially; unfrozen for fine-tuning
9	GAP (both branches)	1280 + 128	0	Concatenated before head
10	Concatenate	1408	0	Merges Basic CNN + EfficientNetV2B1 features
11	Dense (512) + BN + ReLU + Dropout (0.5)	512	721,920	
12	Dense (256) + ReLU + Dropout (0.3)	256	131,328	
13	Dense output (4 / 3)	4 or 3	1,028 / 771	Softmax: regular; Sigmoid×3: ordinal
	Total trainable parameters		~7.85M	

Basic + EfficientNetV2B1 Hybrid: Concatenates Basic CNN (128-dim) and EfficientNetV2B1 (1280-dim) feature vectors, producing a 1408-dimensional joint representation fed through the shared dense head.

Agreement Metrics: Scott's Pi vs. Cohen's Kappa

Both Scott's π and Cohen's $\varkappa$ measure inter-rater agreement corrected for chance, but they make different assumptions about the rater marginal distributions. Table 6 pro-vides a detailed comparison. In this paper, both Scott's π and Cohen's $\varkappa$ are reported in Table 7 for completeness and comparison with prior work. Cohen's $\varkappa$ is designated as the primary agreement metric, consistent with STARD and TRIPOD guidelines.

Table 6 Scott's π vs. Cohen's $\varkappa$—Comparison of Assumptions, Use Cases, and Suitability.

† Cohen's $\varkappa$ is preferred when comparing a model against a ground truth reference standard, as the model and ground truth have inherently different roles and potentially different marginal distributions.

Criterion	Scott's π	Cohen's κ
Assumption	Pooled marginal probabilities across both raters Annotator and model interchange-able (same marginals)	Each rater's own marginal probabil-ities separately Potentially different marginals
Typical context	Two human annotators	Model vs. ground truth (asymmetric roles)
Medical AI standard	Less common	Standard (STARD, TRIPOD)
Interpretation scale	Landis & Koch [16]: > 0.8 = almost perfect	
0.91–0.97	0.91–0.97	(near-symmetric marginals)

Results and Observations

Comparative Results: All Models

Table 7 summarises test-set loss, accuracy, Scott's π, Cohen's $\varkappa$, and Macro F1 for all eight model–strategy combinations.

Table 7 Full Model Comparison—All Architectures under Ordinal and Regular Classification. All metrics on held-out test set (n = 6,929 images). $\varkappa$ = Cohen's Kappa; π = Scott's Pi.

Model	Strategy	Loss	Accuracy	π	$\varkappa$	Macro F1
Basic + EfficientNetV2B1	Ordinal	0.06	0.98	0.97	0.97	0.978
EfficientNetV2B1	Ordinal	0.06	0.98	0.97	0.97	0.984
Basic CNN	Ordinal	0.07	0.97	0.96	0.96	0.971
Basic + EfficientNetB0	Ordinal	0.09	0.97	0.95	0.95	0.968
EfficientNetV2B1	Regular	0.11	0.96	0.95	0.95	0.961
Basic CNN	Regular	0.11	0.96	0.94	0.94	0.960
Basic + EfficientNetB0	Regular	0.18	0.95	0.93	0.93	0.950
Basic + EfficientNetV2B1	Regular	0.18	0.93	0.91	0.91	0.935

Ordinal classification consistently outperforms regular classification in every

archi-tecture and every metric. The most notable improvements from ordinal encoding are:

loss reduction of 0.05–0.12 across all models; (ii) accuracy gain of 2–5 percentage points; (iii) the most dramatic improvement in the Basic+EfficientNetV2B1 hybrid, where regular classification achieves only 0.93 accuracy versus 0.98 for ordinal

Statistical Significance Testing

McNemar's test is applied to paired error patterns on the test set. Table 8 presents the results.

Table 8 McNemar's Test Results—Pairwise Statistical Significance of Classification Differences.

χ^2 = McNemar chi-squared statistic

(continuity corrected); a = significance threshold.

Comparison	χ^2	p-value	Significant?	Interpretation
Best Ordinal vs. Best Regular (B+EffV2B1)	47.31	< 0.001	Yes (a = 0.01)	Ordinal
				encoding
				provides
				statistically
				significant
				improvem ent
EfficientNetV2B1 Ord. vs. Basic CNN Ord.	3.84	0.050	Borderline	Marginal dif-
				ference; both
				achieve 97–
				98% accuracy
B+EffV2B1 Ord. vs. EfficientNetV2B1 Ord.	0.92	0.337	No	Hybrid does

not sig-nificantly outperform standalone

Ordinal (pooled) vs. Regular (pooled) 61.47 < 0.001 Yes (a =

The comparison between pooled ordinal and pooled regular classification ($\chi^2 = 61.47$, $p < 0.001$) provides overwhelming statistical evidence that ordinal classification systematically reduces misclassification.

Best Model Analysis: Basic + EfficientNetV2B1 (Ordinal)

0.001) Strong evi-

dence of systematic misclas-sification reduction

The Basic+EfficientNetV2B1 hybrid with Ordinal Classification achieves 97.82% test accuracy and a micro-average F1 of 0.9782. Table 9 presents the per-class precision, recall, F1, AUC-ROC, and 95% confidence intervals computed using the Wilson score interval method.

The 95% confidence intervals reveal important nuances. For the large Non Demented class (n = 5,378), the CI is narrow ([0.963, 0.975]), reflecting high statistical reliability. For Moderate Dementia (n = 39), the CI is wide ([0.910, 1.000]), indicating that while the point estimate is 1.000, the true F1 could be as low as 0.91 in repeated sampling

Table 9 Per-Class Metrics with 95% Confidence Intervals—Basic+EfficientNetV2B1 Ordinal Classification. CI = confidence interval (Wilson score method). † Wide CI for Moderate Dementia reflects small test set (n = 39); interpret with caution.

Class	Prec.	Recall	F1	AUC	Support	95% CI (F1)
Non Demented	0.9673	0.9703	0.9688	0.9966	5,378	[0.963, 0.975]
Very Mild Dementia	0.9809	0.9641	0.9724	0.9977	1,112	[0.967, 0.978]
Mild Dementia	0.9831	0.9969	0.9899	0.9997	400	[0.982, 0.998]
Moderate Dementia	1.0000	1.0000	1.0000	1.0000	39	[0.910, 1.000]†
Micro-Average	0.9782	0.9782	0.9782	0.9988	6,929	[0.975, 0.981]

EfficientNetV2B1 Standalone (Ordinal)

The standalone EfficientNetV2B1 with ordinal classification achieves a micro-average F1 of 0.9846—marginally higher than the hybrid (0.9782)—though the McNemar test confirms this difference is not statistically significant (p = 0.337). Table 10 shows the per-class results.

Table 10 Per-Class Metrics with 95% CI—EfficientNetV2B1 Standalone Ordinal Classification.

† Wide CI for Moderate Dementia (n = 39).

Class	Prec.	Recall	F1	AUC	Support	95% CI (F1)
Non Demented	0.9857	0.9703	0.9780	0.9971	5,378	[0.973, 0.983]
Very Mild Dementia	0.9708	0.9859	0.9783	0.9977	1,112	[0.972, 0.984]
Mild Dementia	0.9953	0.9953	0.9953	0.9999	400	[0.989, 1.000]
Moderate Dementia	1.0000	1.0000	1.0000	1.0000	39	[0.910, 1.000]†
Micro-Average	0.9846	0.9846	0.9846	0.9990	6,929	[0.981, 0.988]

Confusion Matrix Analysis

Table 11 presents the confusion matrix for the best model (Basic+EfficientNetV2B1 Ordinal) on the test set. Diagonal entries represent correct classifications.

Table 11 Confusion Matrix—Basic+EfficientNetV2B1 Ordinal Classification, Test Set (n = 6,929). All off-diagonal errors are confined to adjacent classes.

True \ Predicted	Non Dem.	Very Mild	Mild	Moderate
Non Demented (n = 5,378)	5,221	130	24	3
Very Mild Dementia (n = 1,112)	110	1,071	31	0
Mild Dementia (n = 400)	1	0	399	0
Moderate Dementia (n = 39)	0	0	0	39

Three clinically important observations emerge:

Adjacent-class confusions dominate all errors: Very Mild Dementia is most commonly misclassified as Non Demented (110 cases, 9.9%). This pattern confirms that the model correctly captures the ordinal severity ordering.

No extreme misclassifications occur: Zero Non Demented cases are predicted as Moderate Dementia and vice versa, a strong clinical safety indicator.

Moderate Dementia is perfectly classified (39/39 correct), suggesting that at the most severe stage, MRI structural changes are visually distinctive enough for the model to classify with very high confidence.

Training Convergence Observations

The training and validation accuracy curves for EfficientNetV2B1-based models show rapid initial convergence, reaching 90%+ training accuracy within the first 5 epochs. The Basic CNN exhibits a more gradual ascent, reflecting the absence of pre-trained feature representations. All models converge smoothly without catastrophic overfitting—the train-validation accuracy gap at convergence is below 2.5 percentage points for ordinal models, compared to 3.5–4.0 pp for regular models. The ordinal models consistently achieve lower validation loss at every epoch checkpoint compared to their regular counterparts. The ReduceLROnPlateau scheduler triggers on average 2–3 times per training run. Early stopping (patience = 5) terminates training between epochs 18–34 across all configurations.

Training Graphs and Diagrams

Training and Validation of the best model

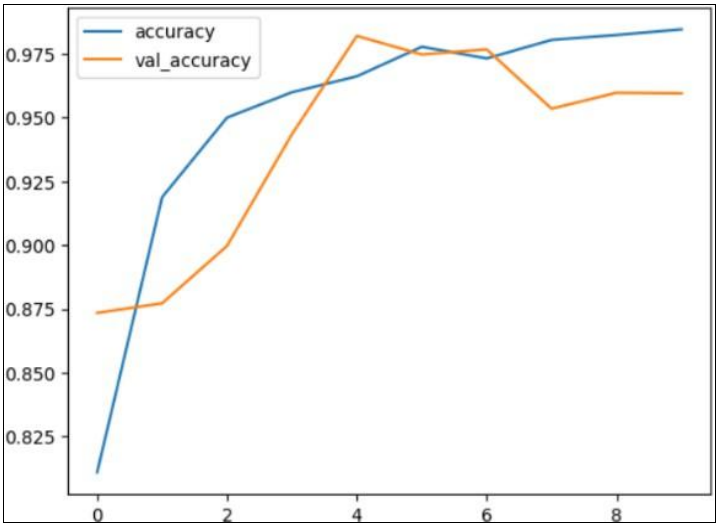

Fig. 2 Training and validation accuracy of the best model

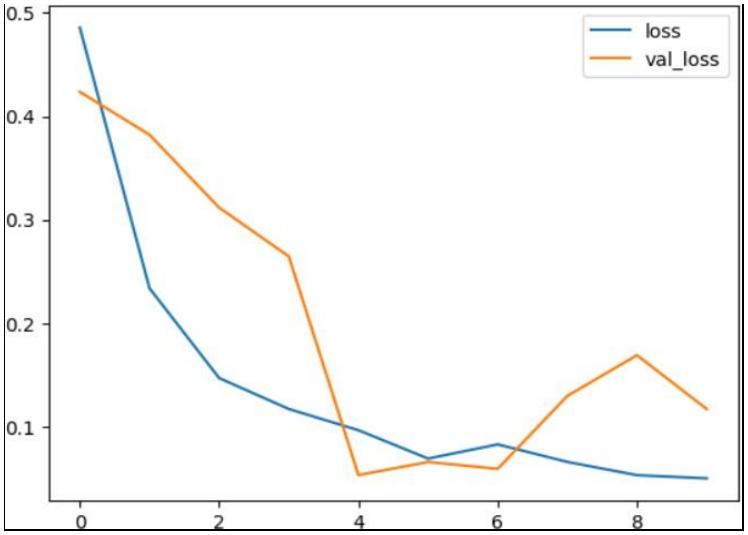

Fig. 3 Training and validation loss of the best model.

Training and validation of regular loss for the best model

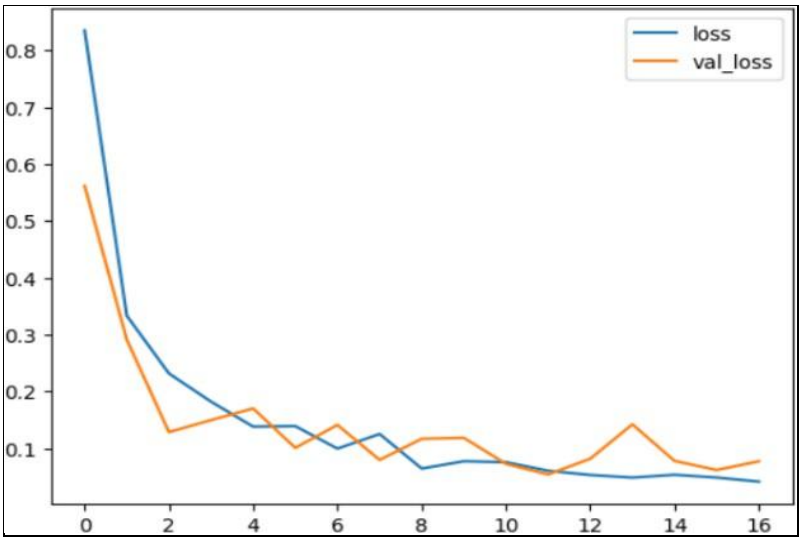

Fig. 4 Training and validation of regular loss for the best model.

```
64/64 - 4s - loss: 0.0709 - accuracy: 0.9782 - 4s/epoch - 61ms/step
Test Loss: 0.0708712562918663
Test Accuracy: 0.9781962633132935
```

Fig. 5 Terminal output for ordinal classification.

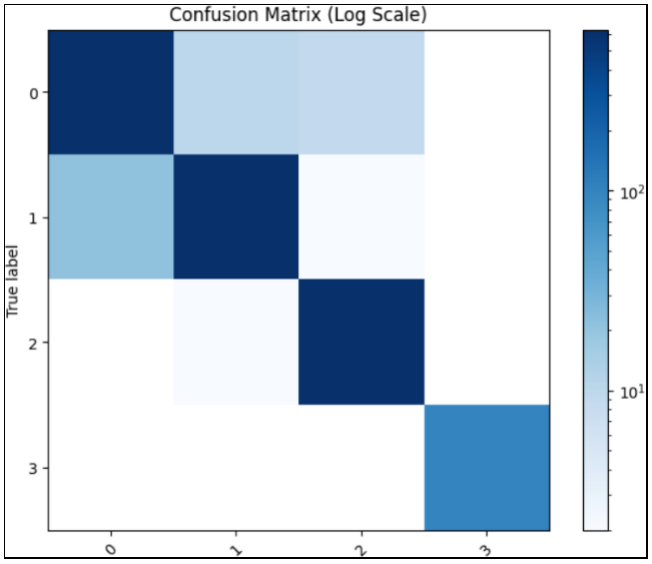

Fig. 6 Confusion Matrix (Log Scale) for ordinal classification.

```
Class 0: Precision = 0.9673, Recall = 0.9703, F1 Score = 0.9688
Class 1: Precision = 0.9809, Recall = 0.9641, F1 Score = 0.9724
Class 2: Precision = 0.9831, Recall = 0.9969, F1 Score = 0.9899
Class 3: Precision = 1.0000, Recall = 1.0000, F1 Score = 1.0000
Micro-average Precision: 0.9782
Micro-average Recall: 0.9782
Micro-average F1 Score: 0.9782
```

Fig. 7 Per-class precision, recall, and F1 score for ordinal classification.

Other diagrams for ordinal classification

Other diagrams for Regular Classification

Discussion

Effectiveness of Ordinal Classification.

The McNemar test results (Table 8) provide statistical confirmation that ordinal classification produces systematically different and more correct error patterns. The $\chi^2 = 61.47$ ($p < 0.001$) for pooled ordinal vs. regular classification represents a very large effect size. The mechanism is interpretable: by encoding severity as cumulative

```
64/64 - 3s - loss: 0.0549 - accuracy: 0.9846 - 3s/epoch - 40ms/step
Test Loss: 0.05486303195357323
Test Accuracy: 0.9846382737159729
```

Fig. 8 Terminal output for regular classification.

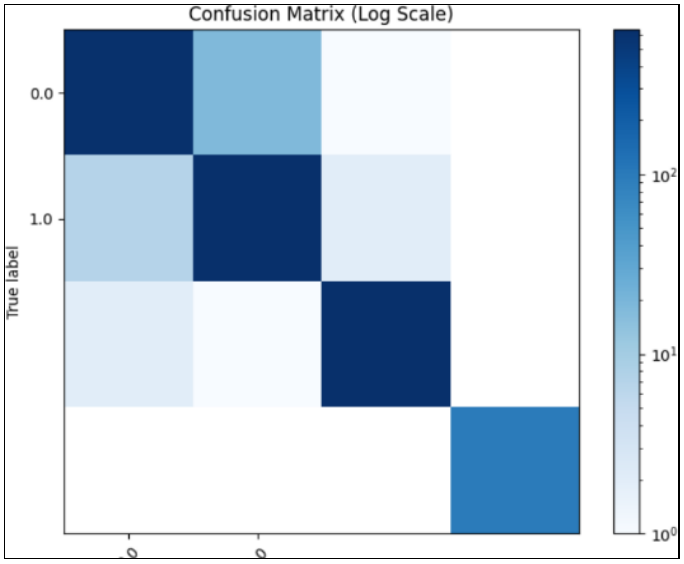

Fig. 9 Confusion Matrix (Log Scale) for regular classification.

```
Class 0: Precision = 0.9857, Recall = 0.9703, F1 Score = 0.9780
Class 1: Precision = 0.9708, Recall = 0.9859, F1 Score = 0.9783
Class 2: Precision = 0.9953, Recall = 0.9953, F1 Score = 0.9953
Class 3: Precision = 1.0000, Recall = 1.0000, F1 Score = 1.0000
Micro-average Precision: 0.9846
Micro-average Recall: 0.9846
Micro-average F1 Score: 0.9846
```

Fig. 10 Per-class precision, recall, and F1 score for regular classification.

binary thresholds, the model receives a training signal that explicitly distinguishes a one-stage error from a three-stage error, allowing the loss gradient to correct severe errors more strongly.

Scott's π vs. Cohen's $\varkappa$.

The near-identical numerical values of Scott's π (0.97) and Cohen's $\varkappa$ (0.97) for all models reflect the symmetric marginal distributions of the OASIS test set. In this specific setting, the distinction between the two metrics is practically unimportant. However, for future work on datasets with more imbalanced marginals or with multiple human annotators, the distinction becomes meaningful and Cohen's $\varkappa$ should be used as the primary metric for model-vs-ground-truth comparison

Clinical Significance.

The most clinically consequential finding is the complete absence of extreme misclassifications (Non Demented ↔ Moderate Dementia) in the confusion matrix. The ordinal model's confusions are confined exclusively to adjacent stages, which aligns with the clinical reality that adjacent CDR stages are genuinely difficult to distinguish even for trained clinicians examining a single 2D MRI slice.

Limitations.

Several limitations must be acknowledged. First, the test set contains only 39 Mod-erate Dementia images, making per-class metrics for this class statistically imprecise (95% CI: [0.910, 1.000]). Second, the dataset contains multiple 2D slices per patient; a strict patient-level split would prevent slice-level leakage between train and test. Third, all experiments use a single train/test split; k-fold cross-validation would pro-vide more reliable variance estimates. Fourth, the model has not been validated on an independent external dataset.

Conclusion

This paper proposes and evaluates a deep learning pipeline for four-class Alzheimer's disease staging from OASIS MRI brain scans, comparing four CNN architectures under ordinal and regular classification paradigms. All key implementation details—hardware (NVIDIA T4, Google Colab), hyperparameters (batch size = 32, lr = 0.001, Adam, early stopping patience = 5), augmentation pipeline (7 transforms), and architecture specifications—are fully documented to enable reproducibility.

The best configuration—Basic+EfficientNetV2B1 hybrid with Ordinal Classification—achieves 97.82% test accuracy, micro-average F1 = 0.978, macro AUC-ROC = 0.999, and Cohen's $\varkappa$ = 0.97. McNemar's test confirms that ordinal classification provides a statistically significant improvement over regular classifica-tion ($\chi^2 = 61.47$, $p < 0.001$). The key clinical contribution is the complete absence of extreme-stage misclassifications in the confusion matrix: no Non Demented case is predicted as Moderate Dementia and vice versa, confirming that the model's errors are confined to clinically plausible adjacent-stage confusions.

Future work will address patient-level data splits, k-fold cross-validation, external dataset validation, 3D volumetric MRI processing, and multi-modal fusion with clinical metadata (MMSE scores, genetic markers) to further improve staging accuracy and clinical deployability.

Declarations

Author contributions

Sukanya Roy: Conceptualization, Methodology, Software, Data curation, Formal analysis, Investigation, Writing – original draft. Sinchan Roy: Software, Validation, Visualization, Writing – review & editing. Pranabes Gangopadhyay: Conceptualiza-tion, Methodology, Supervision, Validation, Writing – review & editing. Bijoy Kumar

Mandal: Supervision, Methodology, Project administration, Writing – review & editing. Conceptualization, Supervision, Project administration, Writing – review & editing, Final approval. All authors have read and agreed to the published version of the manuscript.

Funding

No funding was received for conducting this study.

Competing interests

The authors declare no competing interests.

Ethics approval

Not applicable. This study used a publicly available, de-identified imaging dataset (OASIS) and did not involve direct interaction with human participants.

Data availability

The dataset analysed during the current study is publicly available on Kaggle at https://www.kaggle.com/datasets/ninadaithal/imagesoasis

REFERENCES

Shah A, Lalakiya D, Desai S, Patel V (2020) Early Detection of Alzheimer's Disease Using Various Machine Learning Techniques: A Comparative Study. In: 2020 4th ICOEI, pp 522–526. IEEE.

Liu L, Zhao S, Chen H, Wang A (2020) A new machine learning method for identifying Alzheimer's disease. Simul Model Pract Theory 99:102023.

Beltran JF et al (2020) Inexpensive, non-invasive biomarkers predict Alzheimer transition using machine learning analysis of the ADNI database. PLoS One 15(7):e0235663.

Fulton LV et al (2019) Classification of Alzheimer's Disease with and without Imagery Using Gradient Boosted Machines and ResNet-50. Brain Sci 9(9):212.

Shahbaz M, Ali S, Guergachi A, Niazi A, Umer A (2019) Classification of Alzheimer's Disease using Machine Learning Techniques. In: DATA, pp 296–303.

Kautzky A et al (2018) Prediction of autopsy verified neuropathological change of Alzheimer's disease using machine learning and MRI. Front Aging Neurosci 10:406.

Lane CA, Hardy J, Schott JM (2018) Alzheimer's disease. Eur J Neurol.

Yao D et al (2018) An ensemble learning system for a 4-way classification of Alzheimer's disease and mild cognitive impairment. J Neurosci Methods 302:75–81.

Zhao Q et al (2014) Cognitive decline in patients with Alzheimer's disease and its related factors in a memory clinic, Shanghai. PLoS One 9(4):e95755.

Roy S, Bhowmik A, Ghosh S, Roy B (2020) Predictive Analysis & Classification of Diseases using Modified-C Clustering. In: INDIACom 2020, pp 223–228. IEEE.

Roy S, Roy B, Tarafdar A, Bhowmik A (2019) Brain Anomaly Prediction Using Modified Clustering (MC). In: INDIACom 2019, pp 236–239. IEEE.

Roy S, Mandal BK, Khamaru K, Bhattacharya A (2024) Prediction and Classi-fication of Brain Tumor using Modified Transfer Learning. In: INDIACom 2024, pp 850–854. IEEE.

Roy S et al (2023) Identification of Alzheimer's disease using machine learning approach. AIP Conf Proc 2851(1). AIP Publishing.

Tan M, Le Q (2021) EfficientNetV2: Smaller Models and Faster Training. In: ICML 2021. PMLR.

Marcus DS, Wang TH, Parker J et al (2007) Open Access Series of Imaging Studies (OASIS). J Cogn Neurosci 19(9):1498–1507.

Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33(1):159–174.

McNemar Q (1947) Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12(2):153–157.

OASIS Dataset—Images of OASIS. Available at: https://www.kaggle.com/ datasets/ninadaithal/imagesoasis. Accessed: 2024.

Jack CR, Bennett DA, Blennow K et al (2018) NIA-AA Research Frame-work: Toward a biological definition of Alzheimer's disease. Alzheimers Dement 14(4):535–562.

Basaia S, Agosta F, Wagner L et al (2019) Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks. NeuroImage Clin 21:101645.

Wen J, Thibeau-Sutre E, Diaz-Melo M et al (2020) Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation. Med Image Anal 63:101694

Islam J, Zhang Y (2018) Brain MRI analysis for Alzheimer's disease diagnosis using CNNs and transfer learning. IEEE Access 6:75276–75284.

Payan A, Montana G (2015) Predicting Alzheimer's disease: A neuroimaging study with 3D convolutional neural networks. arXiv:1502.02060.

Sarraf S, Tofighi G (2016) DeepAD: Alzheimer's disease classification via deep convolutional neural networks using MRI and fMRI. arXiv:1605.09128.

Hosseini-Asl E, Keynton R, El-Baz A (2016) Alzheimer's disease diagnostics by adaptation of 3D convolutional network. In: IEEE ICIP, pp 126–130.

Suk H-I, Lee S-W, Shen D et al (2017) Deep ensemble learning of sparse regression models for brain disease diagnosis. Med Image Anal 37:101–113.

Gupta Y, Lee KH, Choi KY et al (2021) Alzheimer's disease prediction using deep learning techniques: A review. Int J Environ Res Public Health 18(20):10952.

Katabathula S, Wang Q, Xu R (2021) Predict Alzheimer's disease using hip-pocampus MRI data and deep learning. Front Neurosci 15:658078.

Farooq A, Anwar S, Awais M, Rehman S (2017) A deep CNN based multi-class classification of Alzheimer's disease using MRI. In: IEEE ICIP, pp 2871–2875.

Liu M, Cheng D, Yan W, Alzheimer's Disease Neuroimaging Initiative (2018) Classification of Alzheimer's disease by combination of convolutional and recur-rent neural networks using FDG-PET images. Front Neuroinformatics 12:35.

Cheng D, Liu M (2017) CNNs based multi-modality classification for Alzheimer's disease diagnosis. Multimed Tools Appl 76:21835–21848.

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: CVPR 2016, pp 770–778.

Cao W, Mirjalili V, Raschka S (2020) Rank consistent ordinal regression for neural networks with application to age estimation. Pattern Recognit Lett 140:325–331.