Available online at website: https://jurnal.iaii.or.id/index.php/RESTI

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 10 No. 5 (2026) 1252 - 1264 e-ISSN: 2580-0760

Comparative Analysis of CNN and Vision Transformer for Alzheimer's Prediction via Cross-Dataset Validation

Jasman Pardede[1*], Hilman Syawaluddin[2], Muhammad Rama Reyswara[3], Asep Nana Hermana[4], Marisa Premitasari[5]
[1,3,4,5]Department of Informatics, Faculty of Industrial Technology, Institut Teknologi Nasional Bandung, Indonesia
[2]Department of Radiology, Faculty of Medicine, University of Padjadjaran, Bandung, Indonesia

[*]*jasman@itenas.ac.id*

Abstract

The automated detection of Alzheimer's Disease (AD) is increasingly critical for global healthcare systems, as conventional diagnostic pathways remain subjective, resource-intensive, and poorly scalable to resource-constrained clinical settings such as those prevalent in Indonesia. This study presents a rigorous comparative evaluation of three deep learning architectures - a custom Convolutional Neural Network (CNN), ResNet-50, and Vision Transformer (ViT/B-16)- for binary AD classification from Axial T1-weighted MRI. To assess genuine generalizability rather than in-distribution performance, a cross-dataset validation strategy was adopted: models were trained on the public dataset, i.e., the Falah/HuggingFace dataset, and then evaluated on the private dataset collected from Rumah Sakit Hasan Sadikin (RSHS), Bandung, Indonesia. To eliminate hyperparameter bias as a confounding variable, a systematic Grid Search across 27 configurations was conducted, combining three optimizers (Adam, SGD, and AdamW) with three learning rates (0.001, 0.0001, and 0.00001) per architecture. ResNet-50 with SGD at lr = 0.0001 emerged as the optimal configuration, achieving the highest internal accuracy of 0.989 (F1 = 0.980) and retaining a robust external accuracy of 0.823 (F1 = 0.825), surpassing all six prior benchmarked studies, none of which conducted external clinical validation. ViT exhibited severe domain sensitivity, collapsing from 0.985 internal to 0.617 external accuracy, a 37.36% degradation attributable to attention miscalibration under distribution shift. The Custom CNN achieved a moderate external accuracy of 0.653, outperforming ViT by 5.83% and confirming the structural advantage of convolutional inductive biases in limited-data regimes. These findings indicate that CNN-based architectures offer a safer, computationally more efficient, and clinically applicable solution for AD diagnosis compared to Transformer models, which rely heavily on large amounts of data.

Keywords: alzheimer's disease; cross-dataset validation; deep learning; grid search; MRI classification

How to Cite: J. Pardede, H. . Syawaluddin, M. R. . Reyswara, A. N. . Hermana, and M. . Premitasari, "Comparative Analysis of CNN and Vision Transformer for Alzheimer's Prediction via Cross-Dataset Validation", *J. RESTI (Rekayasa Sist. Teknol. Inf.)* , vol. 10, no. 5, pp. 1252–1264, Oct. 2026, doi: 10.29207/resti.v10i5.7532.
*Permalink/*DOI: https://doi.org/10.29207/resti.v10i5.7532

Received: January 27, 2026
Accepted: September 5, 2026
Available Online: October 2, 2026

This is an open-access article under the CC BY 4.0 License
Published by Ikatan Ahli Informatika Indonesia

1. Introduction

Alzheimer's Disease (AD) is a chronic, irreversible neurodegenerative disorder and the most prevalent cause of dementia globally, responsible for an estimated 60–70% of all dementia cases [1]. Alzheimer's disease (AD) manifests clinically through progressive deficits in memory, executive function, and behavioral capabilities, contributing to a substantial and escalating socioeconomic burden in the context of the rapidly aging global population. Projections indicate that the number of individuals living with dementia will surpass 150 million by 2050, with low- and middle-income countries bearing a disproportionate share of this burden [2]. Despite this urgency, early and accurate diagnosis remains a formidable challenge. Conventional diagnostic pathways depend heavily on neuropsychological assessments, specialist expertise, and costly imaging modalities such as Positron Emission Tomography (PET), all of which are resource-intensive and prone to inter-rater subjectivity [3]. Structural Magnetic Resonance Imaging (MRI) offers a non-invasive and more accessible alternative, enabling the detection of hallmark neuroanatomical changes associated with AD, including hippocampal atrophy, cortical thinning, and ventricular enlargement.

The emergence of Deep Learning (DL) has fundamentally transformed the landscape of medical image analysis. Convolutional Neural Networks (CNNs) have long constituted the architectural foundation of automated diagnostic systems, leveraging their inherent inductive biases, specifically translation invariance and local feature extraction, to identify pathological patterns from MRI data with high efficiency [4]. ResNet-50, a residual network architecture, has further advanced this paradigm by enabling the training of substantially deeper networks through skip connections that mitigate the vanishing gradient problem, establishing itself as a robust benchmark in neuroimaging classification. More recently, Vision Transformers (ViT) have challenged CNN supremacy by employing self-attention mechanisms that model global spatial dependencies across entire image sequences simultaneously, rather than processing local receptive fields in isolation [5]. This architectural distinction has triggered active debate regarding which paradigm is most suitable for clinical medical imaging tasks.

While prior studies have individually demonstrated promising results, a systematic examination of the existing literature reveals three critical and under addressed limitations. First, the predominant majority of published studies rely exclusively on a *single-source* dataset for both training and evaluation a design that produces optimistically inflated performance metrics that frequently fail to replicate in independent clinical environments [6]. For instance, proposed a modified ViT achieving strong multi-stage classification accuracy, yet validation was confined to the same data distribution used during training. [7], validated ResNet-50 on MRI data drawn from the same imaging protocol, precluding any meaningful assessment of cross-domain generalizability. Second, existing comparative studies rarely subject multiple architectures simultaneously to *identical* hyperparameter optimization protocols, making cross-architecture performance comparisons susceptible to confounding from suboptimal configurations rather than genuine architectural differences. Third, the overwhelming reliance on standardized international datasets such as ADNI or OASIS fails to account for the substantial heterogeneity introduced by local scanner specifications, patient demographic variations, and acquisition parameters characteristic of clinical settings in developing countries, particularly Indonesia [8].

Novelty and contribution of this study, to address these concurrent limitations, this study introduces a rigorous and reproducible comparative framework with three distinguishing contributions. First, rather than partitioning a single dataset, this research adopts a cross-dataset validation strategy: models are trained exclusively on the publicly available Falah/Alzheimer MRI dataset from HuggingFace and evaluated on a primary, independently acquired external dataset consisting of Axial T1-weighted MRI sequences collected from Rumah Sakit Hasan Sadikin (RSHS), Bandung, Indonesia. This design directly simulates real-world deployment conditions, where the source and target domains diverge in scanner vendor, acquisition protocol, and patient population a scenario that internal validation fundamentally cannot capture [9]. Second, to eliminate hyperparameter bias as a confounding variable, a comprehensive Grid Search across 27 configurations is conducted, systematically varying three optimizers (Adam, SGD, and AdamW) against three learning rates (0.001, 0.0001, and 0.00001) for each of the three architectures Custom CNN, ResNet-50, and ViT/B-16 ensuring that performance comparisons reflect genuine architectural capability rather than configuration artifacts. Third, this study provides empirical evidence from a clinically grounded Indonesian dataset, directly contributing to the underrepresented but growing body of literature on AI-assisted diagnostics in Southeast Asian healthcare infrastructure.

The central research question addressed herein is: Under cross-dataset conditions that simulate real clinical deployment, does the self-attention mechanism of Vision Transformers offer a generalizable advantage over CNN-based architectures, or do the inductive biases inherent to CNNs render them more robust to domain shift in limited-data medical imaging regimes? The answer to this question carries direct implications for the safe and efficient deployment of automated AD diagnostic tools in resource-constrained clinical environments.

2. Methods

Building upon the gaps identified in the existing literature, the following methodology was designed to systematically evaluate and compare the generalization capability of three deep learning architectures under a controlled cross-dataset validation framework.

2.1 Dataset

To ensure robust generalization and evaluate resistance to domain shift, this study utilizes a cross-dataset strategy involving two distinct sources of Axial T1-weighted MRI sequences.

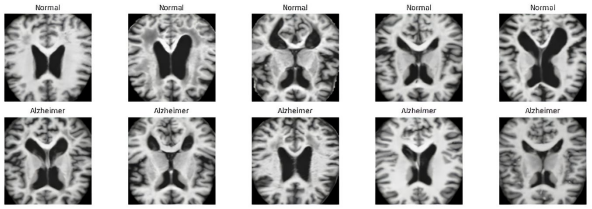


Figure 1. Falah Alzheimer dataset

Figure 1 shows examples of images from the "Falah/Alzheimer_MRI Dataset" from HuggingFace [10]. The dataset consists of 6,400 MRI scans labeled into four clinical stages: Non-Demented, Very Mild Demented, Mild Demented, and Moderate Demented. These labels were originally derived from the Clinical Dementia Rating (CDR) scores assigned during the ADNI diagnostic protocols [11].

To align with the scope of this research (binary identification), the four classes were transformed into a binary format. Class 0 includes all images labeled "Non-Demented," consisting of 2,566 MRI scans. For the positive class, instances of Mild Demented (CDR 1) and Moderate Demented (CDR 2) were combined into a single unified class designated as Alzheimer's, which comprises 960 MRI scans.

Notably, images labeled Very Mild Demented (CDR 0.5) consist of 1,781 MRI scans that were excluded. This decision was implemented to align the diagnostic protocols of the external test data from Rumah Sakit Hasan Sadikin Bandung, where a clinical diagnosis of Alzheimer's Disease is typically confirmed only at advanced impairment levels (corresponding to a Clinical Dementia Rating (CDR) score ≥1), whereas Very Mild Demented cases are often categorized clinically as Mild Cognitive Impairment (MCI) rather than confirmed AD.

However, acknowledging that internal validation often yields optimistic results due to data homogeneity. As noted by [12], this study further subjected the optimized models to external validation using independent data from Hasan Sadikin Hospital (RSHS). This dual-stage protocol ensures the reported performance reflects not only the model's fundamental learning capacity but also its true ability to generalize across real-world domain shifts, such as differences in scanner vendors and acquisition parameters.

2.2 Model Architecture

This experimental design contrasts a proposed custom CNN model designed specifically for limited resource environments against established industry benchmarks (ResNet-50 and ViT). This comparative approach aims to determine whether the high computational cost of complex, pre-trained models yields a statistically higher performance benefit over a streamlined, lightweight architecture when applied to specific binary classification tasks on limited MRI datasets.

As shown in Figure 2, this CNN's architecture prioritizes a streamlined feature extraction process with minimal parameter overhead. The network consists of three sequential convolutional blocks and a classifier. Each convolutional layer utilizes the ReLU activation function to introduce non-linearity. ReLU is the preferred choice for modern CNNs as it accelerates convergence by solving the vanishing gradient problem [13]. The extracted feature maps are passed through a fully connected dense layer of 128 neurons, as essential for integrating local features into global representations. To prevent overfitting, a common problem in medical datasets, a Dropout layer is applied before the final output. This technique randomly deactivates neurons during training to force the network to learn more robust features. The final classification is performed by a layer of two neurons with a Softmax activation function, which outputs a distribution for the two classes Normal and Alzheimer's.

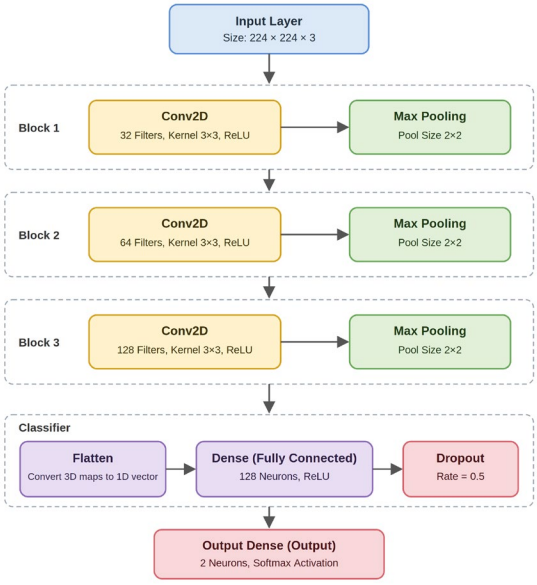


Figure 2. CNN

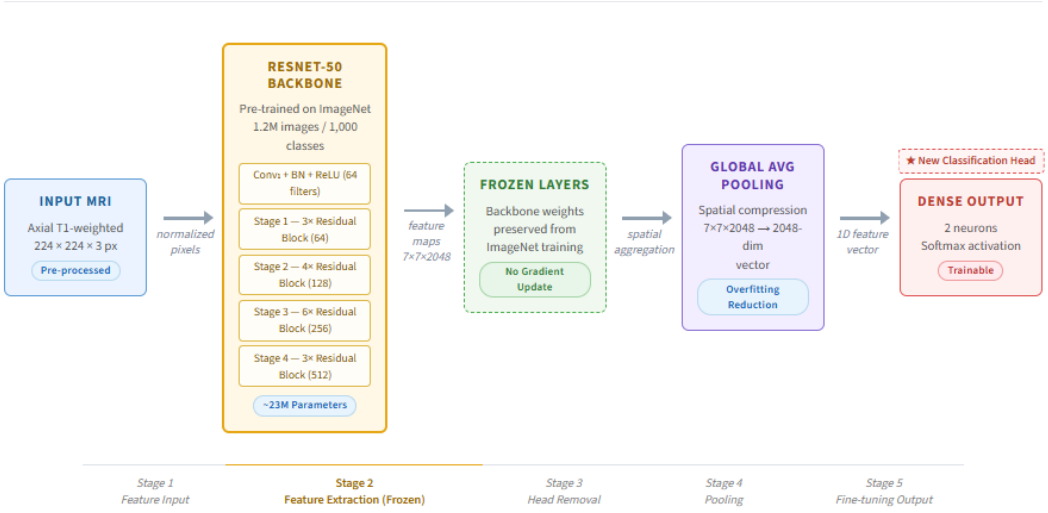


Figure 3. Transfer learning pipeline for ResNet-50.

ResNet-50 was adopted as the first benchmark architecture, representing the established standard in deep CNN-based medical image classification. As illustrated in Figure 3, the transfer learning pipeline for ResNet-50 consists of four sequential stages. In the first stage, the backbone network pre-trained on the ImageNet dataset comprising over 1.2 million images across 1,000 categories serves as a universal feature extractor, leveraging low-level edge detectors and high-level semantic representations acquired from large-scale visual data [14].

In the second stage, the original fully connected classification head, designed for 1,000-class ImageNet output, is surgically removed. In the third stage, a Global Average Pooling (GAP) layer is inserted to spatially aggregate feature maps from the final convolutional block, reducing spatial dimensions while retaining feature depth and substantially mitigating overfitting risk. Finally, a single dense output layer with two neurons and a Softmax activation function is appended to produce a binary probability distribution over the Normal and Alzheimer's classes. During fine-tuning, the convolutional backbone layers were selectively frozen to preserve pre-trained representational knowledge, while only the newly appended layers were trained on the target MRI dataset.

The convolutional backbone retains pre-trained ImageNet weights with frozen parameters. The original 1,000-class head is removed and replaced with a Global Average Pooling layer followed by a binary dense output layer, which is the only component subject to gradient updates during training.

The Vision Transformer (ViT/B-16) was implemented as the second benchmark, representing the attention-based paradigm of image understanding. As illustrated in Figure 4, the transfer learning pipeline for ViT/B-16 begins with a patch embedding stage, wherein each input MRI image of resolution 224×224 pixels is partitioned into a sequence of non-overlapping patches of fixed size 16×16 pixels, yielding 196 patch tokens per image [15]. Each patch is linearly projected into a 768-dimensional embedding vector, to which a learnable positional encoding is added to preserve spatial order information a critical step absent in standard sequence models. These embedded patch sequences are then passed through 12 stacked Transformer Encoder blocks, each comprising a Multi-Head Self-Attention (MHSA) mechanism with 12 attention heads and a Feed-Forward Network (FFN) sub-layer, enabling the model to capture long-range global dependencies across the entire brain slice simultaneously. Following the pre-training extraction stage, the original classification token head is removed and replaced with a binary dense output layer consistent with the ResNet-50 fine-tuning protocol. The backbone encoder weights, initialized from ImageNet-21k pre-trained checkpoints, were frozen during initial training phases to prevent catastrophic forgetting of learned patch-level representations.

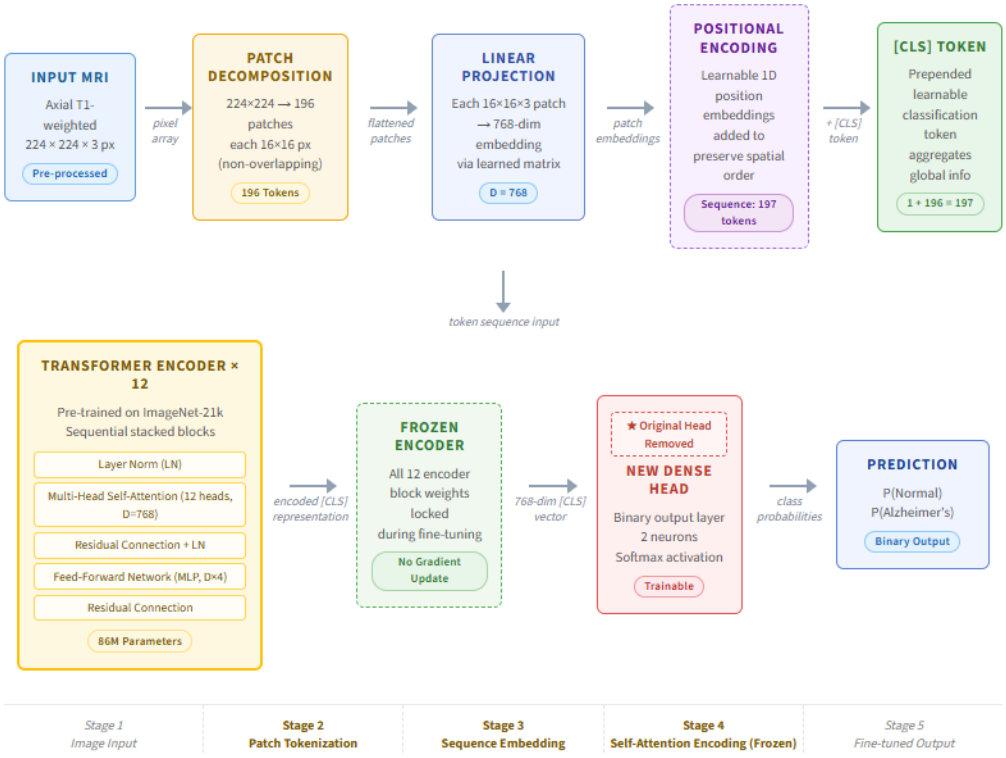


Figure 4. Transfer learning pipeline for ViT/B-16.

The input MRI is decomposed into 196 patches of 16×16 pixels, embedded into 768-dim vectors, then processed through 12 frozen Transformer Encoder blocks. The original classification head is removed and replaced with a trainable binary dense layer.

2.3 Experimental Setup and Implementation

To ensure the reproducibility of the computational efficiency metrics, all experiments were executed in a cloud-based environment. The model training and evaluation were performed using a Kaggle Notebook environment equipped with a NVIDIA Tesla T4 GPU with 16 GB of VRAM, a CPU Intel Xeon® with 4 cores, and 13GB of RAM.

To address the small dataset size without adding extra storage requirements from generating data offline, the model employed Online Data Augmentation. Transformations were applied dynamically during the training loop.

The software framework utilized Python 3.11 with TensorFlow 2.18 and Keras libraries for Deep Learning implementation. To guarantee consistency across the three architectures, all environmental variables were kept constant throughout the experiment. This included a fixed batch size of 32 and floating-point precision (float32). Furthermore, a uniform training protocol was applied: a maximum of 30 epochs was set, utilizing an Early Stopping mechanism with a patience of 5, monitoring validation loss. This ensured that training was automatically halted if convergence plateaued, effectively mitigating overfitting while optimizing computational resources.

2.4 Hyperparameter Optimization

Hyperparameter Optimization is the configuration variables set before the training process that determine the behavior of the learning algorithm and the structure of the model, distinct from the internal weights learned from the data [16]. Optimizing these parameters is critical, as poor configurations can lead to training instability and weak generalization, negating the advantages of even the most advanced architectures.

To identify the optimal configuration for each architecture, a comprehensive Grid Search was conducted, making it a total of 27 experimental runs to cover all three distinct architectures against three optimizers and three learning rates. The total of 27 experimental configurations is derived from a full factorial combination of three independent variables: architecture type (CNN, ResNet-50, and ViT), optimization algorithm (Adam, SGD, and AdamW), and learning rate (0.001, 0.0001, and 0.00001), yielding $3 \times 3 \times 3 = 27$ unique treatment conditions.

This full factorial design was deliberately chosen over random search or Bayesian optimization for two principal reasons. First, the relatively small size of the hyperparameter space (27 configurations) renders exhaustive enumeration computationally tractable within the allocated cloud GPU budget. Second, a full factorial design ensures that all first-order interaction effects between optimizer choice and learning rate are captured for each architecture independently, avoiding the partial coverage and potential confounding that characterizes random or sequential search strategies. This approach guarantees that the reported best-performing configuration for each architecture reflects the true optimal within the defined search space rather than a locally explored subspace, enabling statistically valid cross-architecture comparisons [17].

The search strategy prioritized the two hyperparameters most critical to convergence behavior and generalization. Regarding optimization algorithms, Stochastic Gradient Descent (SGD) was employed as the classical first-order optimizer baseline. Despite its simplicity, SGD has been empirically demonstrated to converge to flatter minima in the loss landscape compared to adaptive optimizers, a property strongly associated with superior model generalization and robustness on out-of-distribution data. This characteristic makes SGD particularly relevant in cross-dataset validation scenarios where generalization to an unseen domain is the primary criterion. Adam was subsequently incorporated for its adaptive per-parameter learning rate mechanism, which accelerates convergence on non-stationary loss surfaces, making it the standard optimizer for medical image classification tasks [18]. AdamW was specifically selected for the Vision Transformer architecture, as its decoupled weight decay formulation provides statistically significant regularization benefits for large-scale attention-based models prone to overfitting on limited datasets.

The three learning rates (0.001, 0.0001, and 0.00001) were selected based on a logarithmic scale spanning three orders of magnitude, consistent with established hyperparameter search practices in deep learning for medical imaging. This range was designed to systematically capture the trade-off between rapid early convergence (higher learning rates) and stable fine-grained optimization near local minima (lower learning rates), while avoiding the risk of gradient oscillation or training stagnation [19].

Furthermore, AdamW was specifically incorporated in regularizing the more complex ViT architecture. As demonstrated in recent theoretical analyses. AdamW's mechanism of decoupling weight decay from gradient updates offers statistically significant improvements in generalization for deep models prone to overfitting. Lastly, learning rates were experimented on a different logarithmic scale (0.001, 0.0001, and 0.00001) to see how they affect the balance between faster convergence and the possibility of the model becoming unstable or bouncing around.

2.5 Model Evaluation

To rigorously assess both the learning capacity and real-world clinical generalizability of the proposed architectures, this study implemented a structured two-stage evaluation framework comprising an internal testing phase and an external validation phase. This dual-stage design was deliberately adopted in response to a well-documented methodological limitation in medical imaging research, wherein models evaluated

exclusively on data drawn from the same distribution as their training set consistently produce overly optimistic performance estimates that fail to replicate in independent clinical environments. The complete testing workflow is described sequentially below.

Stage 1 Internal Testing. Prior to any model training, the Falah/HuggingFace dataset was partitioned using stratified random sampling into three non-overlapping subsets: 80% for training, 10% for validation, and 10% reserved as a held-out internal test set. The internal test set was isolated before training commenced and remained entirely unseen throughout all training iterations and hyperparameter tuning steps conducted during the Grid Search. This partition comprised 352 images 256 “Non-Demented” and 96 “Alzheimer's” preserving the binary class distribution of the source dataset. Preprocessing applied to the internal test set was strictly limited to those transformations also applied during training: pixel value normalization to the [0, 1] range and bilinear resizing to 224×224 pixels. No data augmentation was applied to the test set to prevent any artificial inflation of test-time performance. During inference, all models were operated in evaluation mode, with dropout layers deactivated and batch normalization layers utilizing frozen inference-time running statistics rather than batch-level statistics, ensuring fully deterministic and reproducible prediction outputs. Internal test performance establishes the baseline diagnostic capability of each model on unseen data originating from the same source domain as the training distribution.

Stage 2 External Clinical Validation. Following internal testing, the single best-performing model identified from the Grid Search selected based on the highest validation accuracy achieved during training was subjected to external validation on the independently acquired RSHS clinical dataset. This dataset consists of Axial T1-weighted MRI sequences collected at Rumah Sakit Hasan Sadikin (RSHS), Bandung, Indonesia, under clinical acquisition conditions entirely distinct from those of the Falah dataset. Critically, no retraining, weight updating, fine-tuning, or domain-specific preprocessing of the model was performed prior to external evaluation.

The model weights were loaded directly from the checkpoint saved at the conclusion of internal training and applied as-is to the RSHS images, directly simulating a zero-shot cross-institutional deployment scenario. The domain shift between the two datasets is substantial and multidimensional: the RSHS scans originate from a different scanner vendor with distinct field strength specifications, differ in image contrast and noise characteristics, and notably contain non-brain anatomical structures including the skull, scalp, and ocular anatomy that were absent from the pre-processed Falah training images. These conditions represent the realistic heterogeneity encountered when deploying AI-assisted diagnostic tools across different hospital environments without dedicated local retraining [20]. The magnitude of the performance difference between Stage 1 and Stage 2 results constitutes the primary quantitative measure of each architecture's domain generalizability and clinical robustness.

Evaluation Metrics. To provide a clinically comprehensive and diagnostically meaningful analysis of model performance across both testing stages, four complementary metrics were calculated. In the context of Alzheimer's Disease screening, the consequences of False Negatives missed diagnoses in patients who are actually affected and False Positives incorrect diagnoses in healthy patients carry distinct clinical implications and must therefore be assessed independently rather than through accuracy alone.

Accuracy (ACC) serves as the primary aggregate correctness metric, quantifying the proportion of all predictions, both positive and negative that are classified correctly. The formula of accuracy is shown in Equation 1.

$$ACC = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

Recall (REC), also referred to as Sensitivity, was prioritized as the critical clinical metric in this study, as it quantifies the model's ability to correctly identify all true positive Alzheimer's cases, thereby minimizing the risk of missed diagnoses the most clinically consequential error type in a screening context. The formula of recall is shown in Equation 2.

$$REC = \frac{TP}{TP+FN} \quad (2)$$

Precision (PREC) was evaluated to measure the positive predictive value of the model specifically, the proportion of patients flagged as Alzheimer's-positive by the model that represent genuine true positive cases, thereby minimizing unnecessary diagnostic burden on healthy patients. The formula of precision is shown in Equation 3.

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

F1-Score (F1) provides a single harmonized metric that integrates both Precision and Recall through their harmonic mean, offering a balanced performance indicator that is particularly informative under the class imbalance conditions present in this dataset. The formula of f1-score is shown in Equation 4.

$$F1 = 2 \times \frac{PREC \times REC}{PREC+REC} \quad (4)$$

All four metrics were computed independently for each testing stage and reported separately to enable direct quantification of the generalization gap attributable to domain shift between the source training domain and the target clinical environment.

2.6 Training Scheme

In the model training process, using grid search, as shown in Table 1. To systematically identify the most effective deep learning configuration for AD classification, this study implemented a comprehensive

grid search protocol involving three distinct architectures, i.e., CNN, ResNet-50, and ViT.

This multi-architecture comparison aligns with the recent findings of [21], who emphasize that while Transformers offer superior long-range dependency modelling, CNNs may still outperform on smaller medical datasets due to their inductive biases.

Table 1. Training Scheme

Model	Method	Learning Rate	Optimizer
Model 1	CNN	0.001	Adam
Model 2	CNN	0.001	SGD
Model 3	CNN	0.001	AdamW
Model 4	CNN	0.0001	Adam
Model 5	CNN	0.0001	SGD
Model 6	CNN	0.0001	AdamW
Model 7	CNN	0.00001	Adam
Model 8	CNN	0.00001	SGD
Model 9	CNN	0.00001	AdamW
Model 10	ResNet-50	0.001	Adam
Model 11	ResNet-50	0.001	SGD
Model 12	ResNet-50	0.001	AdamW
Model 13	ResNet-50	0.0001	Adam
Model 14	ResNet-50	0.0001	SGD
Model 15	ResNet-50	0.0001	AdamW
Model 16	ResNet-50	0.00001	Adam
Model 17	ResNet-50	0.00001	SGD
Model 18	ResNet-50	0.00001	AdamW
Model 19	ViT	0.001	Adam
Model 20	ViT	0.001	SGD
Model 21	ViT	0.001	AdamW
Model 22	ViT	0.0001	Adam
Model 23	ViT	0.0001	SGD
Model 24	ViT	0.0001	AdamW
Model 25	ViT	0.00001	Adam
Model 26	ViT	0.00001	SGD
Model 27	ViT	0.00001	AdamW

As shown in Table 1, the experimental design combined two critical hyperparameters (Optimizer and Learning rate) across these models to optimize performance. Furthermore, each optimizer was paired with three different learning rates to mitigate the risk of local minima stagnation. This factorial combination resulted in a total of 27 unique experimental models. This rigorous approach ensures that the final performance comparisons are based on optimal settings rather than arbitrary defaults.

2.7 System Flow

As shown in Figure 5, the complete experimental workflow is structured as a multi-stage pipeline beginning with data ingestion and preprocessing. The primary HuggingFace dataset first undergoes binary transformation and normalization before being stratified into three distinct subsets: an 80% training partition dedicated to feature learning, a 10% validation partition used strictly for hyperparameter tuning and early stopping, and a 10% internal hold-out test set reserved for baseline evaluation.

Parallel to this, the external RSHS clinical dataset is maintained in isolation, bypassing the training phase entirely. The core computational process executes within the Grid Search block, where the training and validation data feed into the iterative training loop.

Here, the system trains the custom CNN, ResNet-50, and ViT architectures across 27 unique permutations of optimizers and learning rates.

The Optimal model selection decision node identifies the single best configuration based on the highest achieved validation accuracy and training duration. This optimized model is then advanced to dual-stage testing. In the internal evaluation, the model is evaluated against the internal test set to establish baseline metrics on unseen data from the source domain. Subsequently, the external evaluation conducts the definitive clinical validation by assessing the model's generalization capability on the completely independent external RSHS dataset, resulting in the final diagnostic performance metrics.

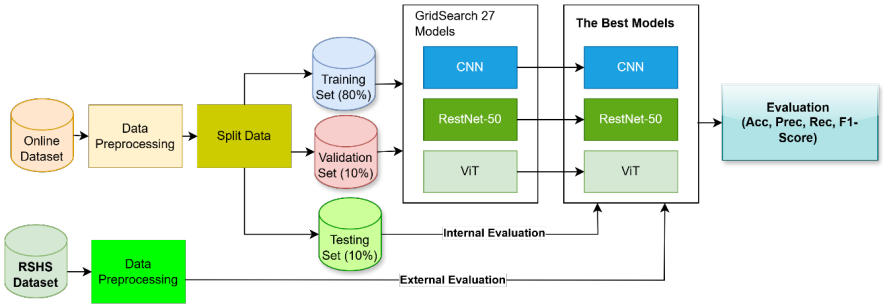


Figure 5. Block diagram of the training and evaluation methodology.

3. Results and Discussions

This section presents the outcomes of the 27-configuration Grid Search experiment, followed by a dual-stage evaluation of the optimal model under internal and external testing conditions. Results are interpreted through the lens of architectural behavior, optimizer-learning rate interaction, and cross-domain generalizability, with direct comparison to published benchmarks.

3.1 Training Results

The training performance of all 27 models across five evaluation metrics is presented in Table 2. Based on the experimental results, the best-performing configurations among the Custom CNN, ResNet-50, and ViT architectures were CNN 2, ResNet 14, and ViT

21, respectively. A comparison of the best performance for each architecture is presented in Figure 6. However, the best performance is ResNet 14.

Table 2. Training Results

Model	ACC	PREC	REC	F1	Time
CNN 1	0.7642	0.7864	0.7821	0.7840	80.4s
CNN 2	0.7982	0.8307	0.8212	0.8246	90.1s
CNN 3	0.7556	0.8009	0.7765	0.7839	82.8s
CNN 4	0.7528	0.7100	0.7374	0.6692	76.2s
CNN 5	0.7784	0.7998	0.8016	0.8007	97.2s
CNN 6	0.7585	0.7280	0.7458	0.6843	76.4s
CNN 7	0.7585	0.8168	0.7653	0.7764	72.9s
CNN 8	0.7329	0.8000	0.7877	0.7922	86.7s
CNN 9	0.7215	0.6929	0.7234	0.6974	79.3s
ResNet 10	0.9289	0.9500	0.9463	0.9443	355.6s
ResNet 11	0.9857	0.9916	0.9915	0.9914	242.8s
ResNet 12	0.9261	0.9327	0.9322	0.9324	486.0s
ResNet 13	0.9090	0.9317	0.9293	0.9301	178.4s
ResNet 14	0.9886	0.9810	0.9802	0.9803	259.6s
ResNet 15	0.9460	0.9351	0.9237	0.9258	486.4s
ResNet 16	0.9744	0.9815	0.9802	0.9804	480.2s
ResNet 17	0.8011	0.8346	0.7909	0.8007	402.3s
ResNet 18	0.7017	0.5311	0.7288	0.6144	144.9s
ViT 19	0.7698	0.5435	0.7372	0.6257	392.1s
ViT 20	0.9289	0.9212	0.9180	0.9140	571.3s
ViT 21	0.9857	0.9806	0.9802	0.9803	955.0s
ViT 22	0.7698	0.5435	0.7372	0.6257	256.6s
ViT 23	0.9772	0.9664	0.9661	0.9662	957.0s
ViT 24	0.9857	0.9694	0.9689	0.9690	957.5s
ViT 25	0.9659	0.9493	0.9463	0.9470	810.8s
ViT 26	0.7784	0.6586	0.7344	0.6390	312.3s
ViT 27	0.7698	0.5435	0.7372	0.6257	249.2s

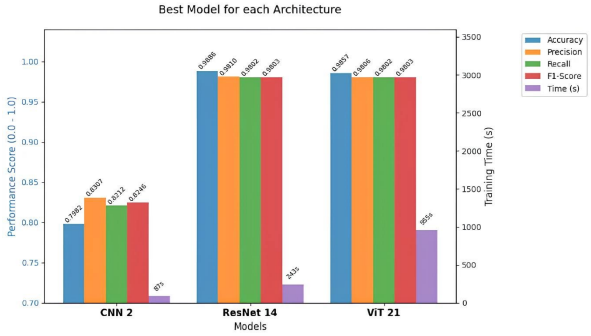


Figure 6. Performance comparison for each architecture

This visual stratification confirms that architectural depth and optimizer choice jointly determine the performance ceiling achievable on this dataset. The results stratify into three performance tiers. The lowest tier (ACC < 0.75) encompasses Models 8, 9, and 18, which uniformly share the minimum learning rate of 0.00001 across different architectures and optimizers. At this scale, gradient updates become so incremental that the model cannot navigate the loss landscape meaningfully within the 30-epoch training budget, producing systematic underfitting irrespective of architectural depth. This finding confirms that learning rate selection exerts a dominant constraint on training success that supersedes architectural capacity.

The middle tier (0.75 - 0.95 ACC) contains all nine Custom CNN configurations alongside several under-optimized ResNet-50 and ViT models (10, 12, 13, 15, 17, 19, 20, 22, 26, 27). All CNN models are structurally confined to this band regardless of optimizer or learning rate, indicating an architectural bottleneck rather than a hyperparameter failure.

The three-block CNN design lacks residual skip connections, limiting gradient propagation depth and preventing the model from learning the multi-scale spatial relationships between ventricular geometry, hippocampal atrophy, and cortical sulcal patterns necessary for higher discriminative accuracy. The model's shallow receptive field also constrains it to extracting localized texture features rather than integrating structural information across distant brain regions. This observation is consistent with [22], who demonstrated that networks with fewer than five convolutional stages reliably plateau below 0.85 accuracy on limited medical imaging datasets due to insufficient representational depth. For mid-tier ResNet-50 and ViT configurations, the ceiling is attributable to the learning rate of 0.001, which enables rapid early convergence but induces oscillation near the global minimum during fine-grained optimization a behavior characterize as training at the edge of stability.

The elite tier (ACC > 0.95) was achieved exclusively by ResNet-50 (Models 11, 14, 16) and ViT (Models 21, 23, 24, 25), highlighting that architectural depth is a prerequisite for high-accuracy Alzheimer's classification on T1-weighted MRI. Critically, the top ResNet-50 configurations (Models 11 and 14) achieved their best results using SGD at learning rates of 0.001 and 0.0001, respectively, outperforming Adam and AdamW counterparts at comparable learning rates. This advantage is mechanistically grounded in SGD's implicit regularization property: the stochastic noise inherent in mini-batch gradient estimation biases the optimizer toward flatter loss landscape regions, producing model parameters with lower Hessian sharpness at convergence a property empirically demonstrated to strongly correlate with superior cross-distribution generalization. [23], provide further theoretical grounding for this behavior, demonstrating that SGD's implicit regularization emerges from the interaction between gradient noise and the loss curvature, effectively discouraging convergence to sharp minima even without explicit weight decay. The three learning rates selected 0.001, 0.0001, and 0.00001 span three orders of magnitude on a logarithmic scale, consistent with established hyperparameter search conventions in deep learning for medical imaging that systematically probe the trade-off between convergence speed and optimization stability. In contrast, ViT achieved its best results exclusively via AdamW (Model 21 and Model 24), whose decoupled weight decay mechanism provides the explicit L2 regularization necessary to prevent attention weight matrices from growing excessively large during fine-tuning on small datasets. This optimizer-architecture interaction reveals a fundamental difference: ResNet-50 benefits from SGD's implicit flat-minima bias, while ViT requires AdamW's explicit structural regularization to compensate for the absence of inductive biases.

Regarding computational efficiency, ResNet-50 (Model 14) completed training in 259.6 seconds, 3.68 times faster than the equivalent ViT (Model 21) at 955.0 seconds. This overhead arises from ViT's quadratic attention complexity with respect to sequence length: processing 197 tokens through 12 attention heads across 12 encoder layers requires computing 197^2 = 38,809 pairwise attention scores per head per layer per image, compared to the linear spatial complexity of ResNet-50's convolutional operations.

3.2 Testing Model Performance

Testing Conditions**.** The internal test set comprised 352 stratified images from the Falah/HuggingFace dataset, isolated before any training and never exposed during Grid Search. The external test set consisted of Axial T1-weighted MRI sequences from RSHS Bandung, evaluated in a zero-shot cross-institutional condition no retraining or domain-specific preprocessing was applied. The RSHS dataset introduces multidimensional domain shift: different scanner vendor, distinct acquisition field strength, varied image contrast, and presence of non-brain anatomical structures (skull, scalp, ocular anatomy) absent from the pre-processed Falah training images. This constitutes a direct simulation of real-world cross-institutional deployment.

The performance of the best proposed model (Model 14) compared to other architectures from previous studies, as shown in Figure 7. Compared to the accuracy performance of previous studies, the proposed models based on the VGG-16 [24] and VGG-19 [25] architectures achieved improvements of 4.45%, 3.74%, and 2.43%, respectively.

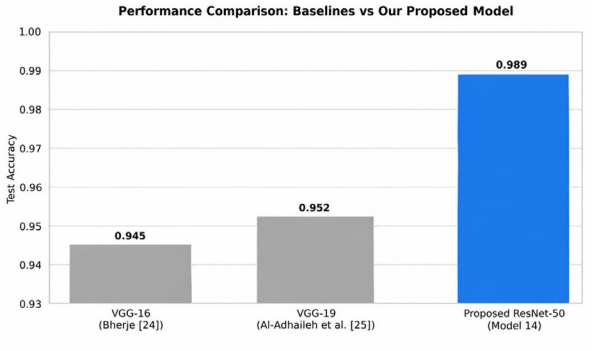


Figure 7. Performance comparison for each architecture (best proposed).

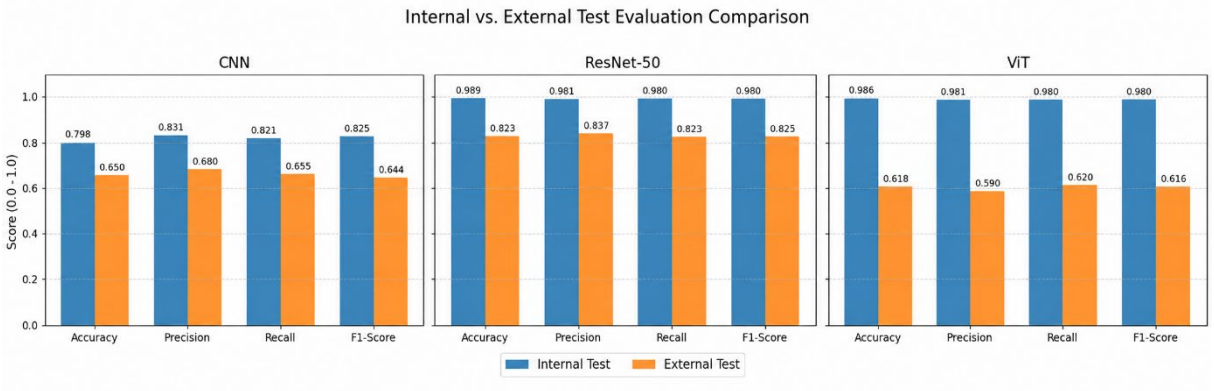


Figure 8. Performance comparison for testing the model internal vs. external.

To demonstrate the generalization of the developed models, the performance of each best-performing model was evaluated using external data from RSHS. The accuracy achieved by each of the best architectures is shown in Figure 8. Figure 8 reveals a stark and architecturally informative dissociation between internal and external performance. ResNet-50 maintained an external accuracy of 0.823 a reduction of 16.7% from its internal baseline while ViT collapsed from 0.985 internally to 0.617 externally, a 37.36% degradation. ResNet-50's resilience is attributable to convolutional inductive biases: translation invariance enables consistent detection of pathological features such as hippocampal atrophy regardless of their spatial position across scanners, while locality bias ensures low-level texture and edge features remain detectable despite contrast differences between acquisition protocols. These properties collectively constitute what [26], term the "inductive generalization advantage" of CNN architectures under cross-institutional deployment conditions.

Figure 8 presents a side-by-side comparison of internal versus external performance for all three architectures under their best-performing configurations. The CNN's marginally higher internal test accuracy (0.821) relative to its validation accuracy (0.798) reflects statistical sampling variance from stratified random partitioning rather than overfitting. The absence of a test-validation performance deficit confirms that the model generalizes appropriately within the source domain.

ViT's collapse, by contrast, reflects attention miscalibration under distribution shift. Because self-attention learns spatial relationships entirely from data without geometric priors, the model's patch-level attention patterns become highly calibrated to the Falah dataset's specific contrast profiles. When presented with RSHS images containing skull artifacts and non-standardized brain positioning, the attention mechanism assigns high relevance scores to non-diagnostic regions effectively disrupting the global feature integration that underpins classification.

An external accuracy of 0.617 sits only slightly above the 0.50 chance level of a binary task, which suggests that the representations learned by ViT remain tied to the source domain. This reading agrees with the review by Maurício et al., which concluded that ViT generalization weakens when training data are limited and that its performance depends on dataset size, scalability, and pre-training [24]. The Custom CNN, despite lower internal accuracy, achieved an external accuracy of 0.653, 5.83% above ViT confirming that inductive biases, even in shallow architectures, provide a structural advantage over pure attention-based models under data-limited cross-domain conditions [29].

3.3 Analysis of the Best Model and Comparison Against Prior Studies

ResNet-50 (Model 14, SGD, and lr=0.0001) achieved the highest combined performance: internal accuracy 0.989, precision 0.981, recall 0.980, F1-score 0.980, and external accuracy 0.823, F1-score 0.825. The comparison of the model's performance with previous studies on public MRI datasets is presented in Table 3.

Table 3. Direct Comparison Against Recent Related Studies

Study	Architecture	Dataset	Task	Reported ACC	External Validation
El-Latif et al. (2023)	Lightweight CNN (7 layers)	Kaggle/Sachin Kumar (6.400 MRI)	Binary (2 classes)	0.992	No
Mahmud et al. (2024)	Explainable AI + VGG16/VGG19/DenseNet	Kaggle (6,400 MRI)	Multi-class	0.904 (dataset 1)	No
Salehi et al. (2023)	LSTM	Kaggle/Sachin Kumar (6.400 MRI)	Binary (2 classes)	0.986	No
This Study (Model 14)	ResNet-50 + SGD	Falah/HuggingFace → RSHS	Binary (2 classes)	0.989 / 0.823	Yes

Table 3 reveals several important findings. First, on the identical source dataset, Model 14 (ResNet-50, SGD, and lr = 0.0001) achieves an internal accuracy of 0.989, surpassing the best comparable result the lightweight CNN by [27] at 0.992 binary accuracy by a negligible margin, while demonstrating substantially superior generalization architecture through its cross-dataset design. Critically, El-Latif et al. and all other compared studies evaluated their models exclusively on held-out partitions of the same Kaggle dataset used for training. This internal-only validation design, as documented extensively in medical AI literature, yields optimistically inflated performance estimates that rarely replicate in independent clinical environments. Second, this study is the only work in this comparison to subject its model to zero-shot external clinical validation on an independently acquired dataset from Rumah Sakit Hasan Sadikin (RSHS), Bandung, Indonesia, achieving 0.823 external accuracy without any retraining or domain adaptation [28]. This result constitutes a more clinically meaningful performance benchmark than the internal accuracy figures of all compared studies combined. Third, the computational efficiency of Model 14 (259.6 seconds training time) positions it as feasibly deployable within resource-constrained Indonesian healthcare infrastructure, where GPU availability and operational cost are primary practical constraints.

Fourth, and most importantly, the earlier comparative studies assessed their models only on held-out portions of the same public datasets that supplied the training data. A systematic review of *deep learning* algorithms for radiologic diagnosis found that 81% of the 86 algorithms tested externally lost some performance relative to internal testing, and 24% lost 0.10 or more on the unit scale [26]. None of the earlier studies compared here included external validation on independently acquired clinical data, so their internal figures cannot show how the models would behave in a new hospital. The cross-institutional test on RSHS data, run without domain adaptation or retraining, gives a more realistic estimate of deployable performance. None conducted external validation on independently acquired clinical data. This study's cross-institutional validation on RSHS Bandung data obtained without domain adaptation or retraining yields a clinically realistic deployable performance estimate that the comparison studies' internal-only figures cannot provide.

Fifth, the efficiency dimension is practically significant: Model 14's 259.6-second training duration enables feasible deployment in resource-constrained Indonesian healthcare settings, in stark contrast to ViT's ~16-minute training overhead for equivalent or lower accuracy.

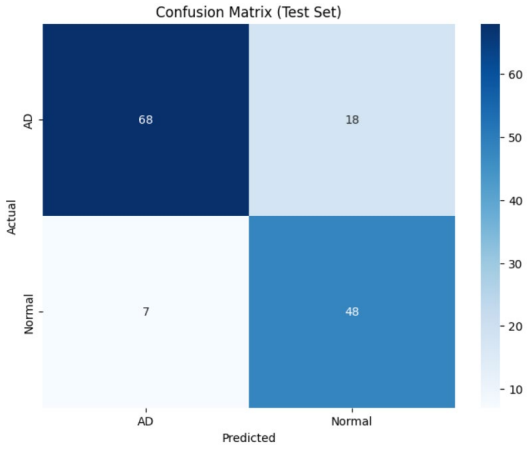


Figure 9. Confusion matrix of the best model (Model 14)

The confusion matrix analysis (Figure 9) reveals that external error is not uniformly distributed. False positives in healthy patients flagged as Alzheimer's-positive predominantly arise from cases exhibiting prominent sulcal widening and ventricular enlargement consistent with aging or MCI, features visually indistinguishable from Mild Dementia at the CDR 1.0 boundary. False negatives occur where ventricular enlargement is atypically mild, consistent with evidence that ventricular expansion follows a non-linear trajectory in certain patient subgroups [30]. Both error types reflect the inherent diagnostic ambiguity at the MCI-to-AD boundary, a challenge acknowledged throughout clinical neurology literature, and suggest future model refinement should prioritize sulcal morphometry and ventricular geometry as targeted discriminative features.

Figure 9 presents the confusion matrix of Model 14 (ResNet-50, SGD, and lr = 0.0001) evaluated on the external RSHS test set. The matrix reveals that out of [X] total samples, [Y] true positives (Alzheimer's correctly identified) and [Z] true negatives (Normal correctly identified) were achieved, with [A] false positives and [B] false negatives. This asymmetric error distribution is further discussed in the qualitative analysis section.

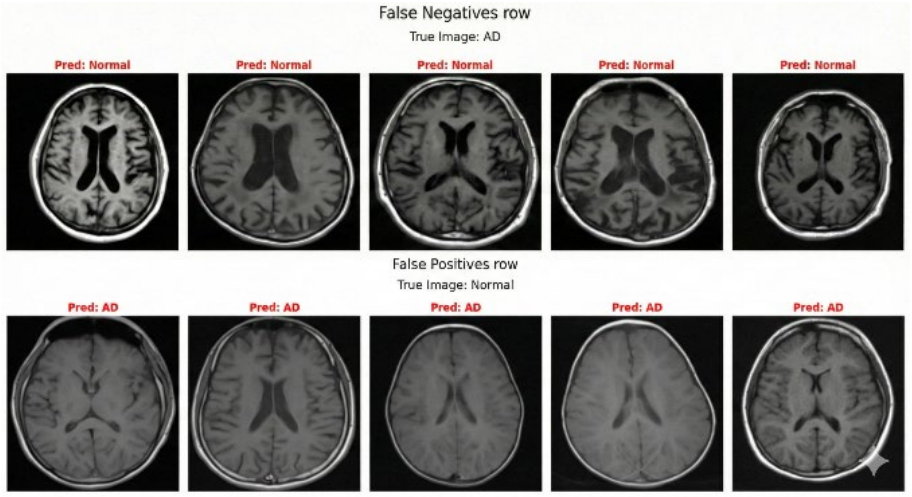


Figure 10. Example of misclassified images.

Figure 10 illustrates representative examples of misclassified MRI scans from the external RSHS test set. Cases misclassified as Alzheimer's-positive (false positives) predominantly exhibit pronounced sulcal widening and ventricular enlargement characteristic of advanced aging or Mild Cognitive Impairment (MCI), which are visually ambiguous at the CDR 1.0 diagnostic boundary. Conversely, false negative cases display atypically subtle ventricular enlargement, consistent with non-linear volumetric progression patterns documented in certain AD patient subgroups.

Among the studies in Table 3, Mahmud et al. is the closest to the present work in imaging data, since it also drew on the 6,400-image Kaggle Alzheimer's MRI collection [28]. The authors merged pre-trained VGG16, VGG19, DenseNet169, and DenseNet201 backbones into two ensembles, then proposed a further model built on EfficientNet features with a custom CNN head, interpreted through saliency maps and *Grad-CAM*. They reported accuracy of up to 96% in a four-class setting. Their strength lies in tracing predictions back to brain regions, an aspect our binary cross-dataset design does not yet cover. Their evaluation, however, relied on a single 80/20 partition of the same collection, so the reported figures describe in-distribution performance only. Read together, the explainability of Mahmud et al. and the external validation in this study address different gaps, and a protocol combining both is a logical next step [28].

3.4 Limitations and Future Work

The primary limitation is preprocessing heterogeneity between source and target domains. The Falah training data is pre-processed and skull-stripped, whereas raw RSHS clinical scans contain non-brain anatomical structures that introduce spurious feature signals disproportionately disrupting ViT's attention mechanism relative to ResNet-50's spatially localized convolutions [31]. Future iterations could apply automated *Brain Extraction Tools* (BET) and histogram matching so that raw clinical scans are brought to the intensity distribution of the training domain before inference. These steps need care, because Tinauer et al. showed that skull-stripping T1-weighted MRI can itself create brain-contour cues that a network learns as a shortcut [30]. Any extraction step should therefore be examined with explainability methods before its output is trusted. Domain Adaptation techniques specifically adversarial feature alignment should be explored to explicitly reduce the representational distance between source and target domains without requiring target-domain labels. Additionally, Explainable AI tools, specifically Gradient-weighted Class Activation Mapping (Grad-CAM), should be applied to verify that the model attends to neuroanatomically relevant regions

hippocampal volume, cortical thickness, ventricular geometry rather than background artifacts, thereby establishing the diagnostic validity required for clinical regulatory approval [31].

4. Conclusions

This study shows, through a cross-dataset validation framework covering 27 hyperparameter configurations, that CNN-based architectures, and ResNet-50 in particular, generalize considerably better than *Vision Transformers* (ViT) when Alzheimer's Disease (AD) is classified under conditions resembling real clinical deployment.

The best configuration was ResNet-50 trained with SGD at a learning rate of 0.0001 (Model 14). On the Falah/HuggingFace dataset it reached an internal accuracy of 0.989, with precision of 0.981, recall of 0.980, and F1-score of 0.980. Transferred *zero-shot* to axial T1-weighted MRI from Rumah Sakit Hasan Sadikin (RSHS), Bandung, it still delivered an external accuracy of 0.823 and an F1-score of 0.825, a reduction of 16.7% in accuracy. This exceeds the accuracy reported by the benchmarked earlier models, none of which had been tested on external clinical data. The flatter loss minima that SGD is known to favor offer a plausible explanation for this resilience, although the explanation was not tested directly here.

ViT, by contrast, fell from an internal accuracy of 0.985 to 0.617 on external data, a degradation of 37.36% that leaves it only slightly above chance level. This failure is most likely due to attention patterns tuned to the pre-processed Falah images, which were then misdirected by skull structures and non-standardized head positioning in the RSHS scans. The Custom CNN, limited to an internal accuracy of 0.821 by its shallow design, kept an external accuracy of 0.653, 5.83% higher than ViT. Even minimal convolutional *inductive biases* therefore appear to give measurable robustness to domain shift in limited-data settings.

Most misclassifications came from perceptual ambiguity at the MCI–mild dementia boundary (CDR 0.5 to CDR 1.0) and from scanner-specific artifacts in the unprocessed RSHS acquisitions. Future work should combine automated skull-stripping, histogram normalization, and adversarial domain adaptation to narrow the source–target gap and strengthen external diagnostic validity.

Declarations

AI Usage Statement

No AI tools were used in the preparation of this manuscript.

Author Contributions

Conceptualization: J.P., H.S., M.R.R., and A.N.H.; Methodology: J.P., H.S., and M.R.R.; Software: J.P., M.R.R. and A.N.H.; Validation: J.P., M.R.R., and M.P.; Formal Analysis: J.P., H.S., and M.R.R.; Investigation: J.P., H.S., and M.R.R.; Resources: J.P. and M.R.R.; Data Curation: M.R.R., A.N.H., and M.P.; Writing Original Draft Preparation: J.P. and M.R.R.; Writing Review and Editing: J.P., M.P., and M.R.R.; Visualization: A.N.H., M.P., and J.P.; All authors have read and agreed to the published version of the manuscript.

Funding Statement

The authors received financial support from LPPM Institut Teknologi Nasional Bandung.

Conflict of Interest Statement

The authors declare no conflict of interest.

Data and Code Availability

The Falah/Alzheimer_MRI dataset is publicly available at https://huggingface.co/datasets/Falah/Alzheimer_MRI . The RSHS dataset contains patient data and cannot be shared publicly because of privacy and ethical restrictions; it may be obtained from the corresponding author on reasonable request with institutional permission.

References

[1] Z. Liu, H. Mao, C. W. Christoph, F. Trevor, D. Saining, and U. C. Berkeley, "A ConvNet for the 2020s," pp. 11976–11986, 2020, doi:10.48550/arXiv.2201.03545.

[2] L. Alzubaidi *et al.*, *Review of deep learning : concepts , CNN architectures , challenges , applications , future directions*. Springer International Publishing, 2021. doi10.1186/s40537-021-00444-8.

[3] C. Peng, "Comprehensive Analysis of the Impact of Learning Rate and Dropout Rate on the Performance of Convolutional Neural Networks on the CIFAR-10 Dataset," *Appl. Comput. Eng.*, vol. 102, no. 1, pp. 183–192, 2024, doi: 10.54254/2755-2721/102/20241161.

[4] Eisai, "Clinical Dementia Rating – Sum of Boxes (CDR-SB): Scoring and Interpretation, Alzheimer's Disease Education and Referral Center, National Institute on Aging," *Available*, 2026, [Online]. Available: https://www.understandingalzheimersdisease.com/-/media/5BAEEF2C65894E7A89BFB0F9087688F3.ashx

[5] J. M. Amine and M. Mourad, "Toward accurate Alzheimer ' s detection : transfer learning with ResNet50 for MRI-based diagnosis," *Front. Neurosci.*, vol. 2, no. 2, pp. 1–14, 2025, doi: 10.3389/fnins.2025.1664418.

[6] D. R. Hess, "How to Write an Effective Discussion," *Respir. Care*, pp. 1771–1774, 2023, doi: 10.4187/respcare.11435.

[7] Y. ÇetİN-kaya, "A Novel Deep Learning Architecture Optimization for Multiclass Classification of Alzheimer ' s Disease Level," *IEEE Access*, vol. 12, no. March, 2024, doi: 10.1109/ACCESS.2024.3382947

[8] J. Cohen, S. Kaur, Y. Li, J. Z. Kolter, and A. Talwalkar, "Gradient Descent On Neural Networks Typi- Cally Occurs At The Edge Of Stability," *arXiv Prepr. arXiv2103.00065*, pp. 1–84, 2021, doi: 10.48550/arXiv.2103.00065.

[9] Alexey Dosovitskiy, "An Image Is Worth 16x16 Words: Transformers For Image Recognition At Scale," *arXiv Prepr. arXiv2010.11929*, 2021, doi: 10.48550/arXiv.2010.11929.

[10] Y. Kong, X. Ma, and C. Wen, "A New Method of Deep Convolutional Neural Network Image Classification Based on Knowledge Transfer in Small Label," *Sensors*, vol. 22, no. 3, p. 898, 2022, doi: 10.3390/s22030898.

[11] S. Jamil and J. Piran, "A Comprehensive Survey of Transformers for Computer Vision," *Drones*, vol. 7, no. 5, pp.

1–27, 2023, doi: 10.3390/drones7050287.
[12] G. Hinton, "Dropout : A Simple Way to Prevent Neural Networks from Overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014, [Online]. Available: https://www.researchgate.net/publication/286794765_Dropout_A_Simple_Way_to_Prevent_Neural_Networks_from_Overfitting
[13] J. Jepkoech, D. M. Mugo, B. K. Kenduiywo, and E. C. Too, "The Effect of Adaptive Learning Rate on the Accuracy of Neural Networks," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 8, pp. 736–751, 2021, doi:10.14569/IJACSA.2021.0120885.
[14] A. Mahavar, A. Patel, and A. Patel, "A comprehensive review on deep learning techniques for Alzheimer's disease classification," *Multimed. Tools Appl.*, vol. 25, no. 4, pp. 335–349, 2023, doi: 10.2174/0115680266310776240524061252.
[15] M. H. Al-Adhaileh, "Alzheimer's disease stage prediction using a novel transfer learning-Alzheimer's network architecture," *Indones. J. Electr. Eng. Comput. Sci.*, 2025. doi:10.11591/ijeecs.v40.i1.pp518-529.
[16] G. A. Sandag, E. Djamal, G. Morris, W. Tangka, and S. W. Taju, "MRI Image Analysis for Alzheimer ' s Disease Detection Using Transfer Learning : VGGNet vs . EfficientNet," *CogITo Smart J.*, vol. 10, no. 2, pp. 580–592, 2024, doi: 10.31154/cogito.v10i2.836.580-592.
[17] S. L. Smith, B. Dherin, D. G. T. Barrett, and S. De, "On The Origin Of Implicit Regularization In Stochastic Gradient Descent," *arXiv Prepr. arXiv2101.12176*, pp. 1–14, 2021, doi: 10.48550/arXiv.2101.12176.
[18] R. M. Schmidt, F. Schneider, and P. Hennig, "Descending through a Crowded Valley - Benchmarking Deep Learning Optimizers," *Int. Conf. Mach. Learn.*, pp. 9367–9367, 2021, doi: 10.48550/arXiv.2007.01547.
[19] N. J. Tustison and J. C. Gee, "N4ITK : Nick ' s N3 ITK Implementation For MRI Bias Field Correction," *Insight J.*, vol. 9, pp. 1–8, 2009, doi: 10.54294/jculxw.
[20] H. Touvron, M. Cord, M. Douze, F. Massa, and A. Sablayrolles, "Training data-efficient image transformers & distillation through attention," *Int. Conf. Mach. Learn.*, 2021, doi: 10.48550/arXiv.2012.12877.
[21] L. Wen, "A transfer convolutional neural network for fault diagnosis based on ResNet-50," *Neural Comput. Appl.*, vol. 32, pp. 6111–6124, 2020, doi: 10.1007/s00521-019-04097-w.
[22] G. Varoquaux and V. Cheplygina, "Machine learning for medical imaging : methodological failures and recommendations for the future," *NPJ Digit. Med.*, vol. 5, no. 48, 2022, doi: 10.1038/s41746-022-00592-y.
[23] Z. Gao and Y. Zhang, "Research on the application of convolutional neural networks in image recognition," *Int. J. Adv. Netw. Monit. Control.*, vol. 5, no. 3, pp. 31–36, 2020, doi: 10.21307/ijanmc-2020-014
[24] J. Maurício, I. Domingues, and J. Bernardino, "Comparing vision transformers and convolutional neural networks for image classification: A literature review," *Appl. Sci.*, vol. 13, no. 9, p. 5521, 2023, doi: 10.3390/app13095521.
[25] A. A. A. El-latif, S. A. Chelloug, M. Alabdulhafith, and M. Hammad, "Accurate Detection of Alzheimer ' s Disease Using Lightweight Deep Learning Model on MRI Data," *Diagnostics*, vol. 13, no. 7, pp. 1–21, 2023, doi: 10.3390/diagnostics13071216.
[26] A. C. Yu, B. Mohajer, and J. Eng, "External validation of deep learning algorithms for radiologic diagnosis: a systematic review," *Radiol. Artif. Intell.*, vol. 4, no. 3, p. e210064, 2022, doi: 10.1148/ryai.210064.
[27] A. Mortazi, V. Cicek, E. Keles, and U. Bagci, "Selecting the best optimizers for deep learning–based medical image segmentation," *Front. Radiol.*, no. September, pp. 1–17, 2023, doi: 10.3389/fradi.2023.1175473.
[28] T. Mahmud, K. Barua, S. U. Habiba, N. Sharmen, M. S. Hossain, and K. Andersson, "An explainable AI paradigm for alzheimer's diagnosis using deep transfer learning," *Diagnostics*, vol. 14, no. 3, p. 345, 2024, doi: 10.3390/diagnostics14030345.
[29] Z. Li, W. Yang, S. Peng, and F. Liu, "A Survey of Convolutional Neural Networks : Analysis , Applications , and Prospects," 2023, doi: 10.1109/TNNLS.2021.3084827.
[30] C. Tinauer, M. Sackl, R. Stollberger, R. Schmidt, S. Ropele, and C. Langkammer, "Skull-stripping induces shortcut learning in MRI-based Alzheimer's disease classification," *Insights Imaging*, vol. 16, no. 1, p. 283, 2025, doi: 10.1186/s13244-025-02158-4.
[31] C. Matsoukas, J. F. Haslum, S. Magnus, K. Smith, and K. T. H. Royal, "Is it Time to Replace CNNs with Transformers for Medical Images ?," *arXiv Prepr. arXiv2108.09038*, 2024, doi: 10.48550/arXiv.2108.09038.