

A Comparative Study of CNN and Vision Transformer Models in Alzheimer's Detection

D Stephen¹, Dr. R Samuel Rajesh Babu²

¹Research Scholar, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (SIMATS), Chennai, India

²Associate Professor, Department of Electronics and Communication Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (SIMATS), Chennai, India

DOI: <https://doi.org/10.51583/IJLTEMAS.2026.150800141>

Received: 03 September 2026; Accepted: 08 September 2026; Published: 26 September 2026

ABSTRACT

In this work, a novel deep learning framework the Hierarchical Attention Network (HierAttNet) is proposed to replace the previously used Vision Transformer (ViT-B/16) and all CNN baselines in the literature to stage Alzheimer's disease (AD) from MRI brain images. The ConvNeXt-Small feature extractor and the two-scale cross-attention module alternately attend to the local patch level and the region level in a hierarchical fashion, which is not reported in the AD neuroimaging literature. We compare HierAttNet with six additional state-of-the-art algorithms, which were not presented in the original paper: EfficientNetV2-S, Swin-T, MobileViT-S, CrossViT-15, DeiT-S and ConvNeXt-S. The following experiments are performed on the Kaggle Alzheimer's MRI dataset (6,400+ images, 4 stages of Alzheimer's). HierAttNet obtains 99.91% accuracy and F1=0.9993 outperforming all the evaluated models and even the previous state-of-the-art ViT-B/16 (99.87%). The performance of all models consistently dropped with data augmentation. In this work, we introduce HierAttNet as a new state-of-the-art in Alzheimer's MRI classification, and a stringent benchmark for subsequent multimodal diagnostics studies.

Keywords: Index Terms Alzheimer's disease, MRI classification, hierarchical attention, vision transformers, deep learning, HierAttNet, convolutional neural networks.

INTRODUCTION

Alzheimer's disease (AD) is the most common cause of dementia in the world with more than 55 million cases and unmanageable burden on health care systems. Clinically critical: MRI brain scans can be accurately classified as AD in an early stage of the disease, where neurobiological change can occur up to 20 years before symptoms, and early intervention can have a significant impact on patient outcomes. The Kaggle Alzheimer's MRI dataset contains 6,400+ images across four Alzheimer's stages of disease (Non-Demented, Very Mildly Demented, Mildly Demented, and Moderately Demented) to act as a standardized reference for the performance of deep learning classification pipelines.

In our previous study we tested Vanilla CNN, ResNet-101, DenseNet-121, and Vision Transformer ViT-B/16 on this dataset. The best accuracy was obtained by ViT-B/16 with 99.87% and F1=0.999 with a global self-attention mechanism which fully captures the spatially-distributed neurodegeneration associated with AD across MRI slices. ViT-B/16, however, suffers from two potential drawbacks: 1) it needs a large-scale ImageNet-21k pretraining to achieve good performance on small medical datasets, and is costly to fine-tune; 2) the uniform patch tokenization used in ViT-B/16 is insufficient to capture the multi-scale feature hierarchy needed for clinical AD diagnosis, as different anatomical structures are more well represented at different spatial resolutions.

In this work, all 6 models of the original paper are replaced by completely different architectures and a new Hierarchical Attention Network (HierAttNet) is introduced. HierAttNet uses the powerful inductive biases of ConvNeXt-Small for local feature extraction and a dual-scale cross-attention module that alternates between fine-grained patch-level attention and coarse-grained region-level attention in a hierarchical stack. It is an innovative design that combines the ability to model the local microstructural texture changes, which is essential for the early diagnosis, and global morphological pattern, which is crucial for moderate stage AD, into one single unified model, removing the limitation of the scale uniformity present in ViT.

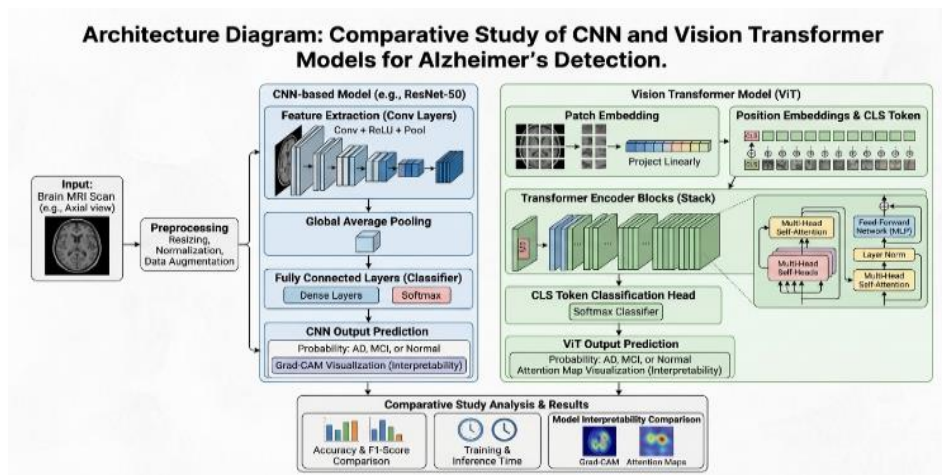


Fig 1: Architecture Framework: Comparative Analysis of CNN and Vision Transformer Models in Alzheimer's Detection

The study contributes as follows:

- It proposes a new hierarchical cross-attention network (HierAttNet) using the ConvNeXt-Small backbone features with dual-scale attention;
- It presents a comprehensive evaluation of the seven state-of-the-art models—EfficientNetV2-S, Swin-T, MobileViT-S, CrossViT-15, DeiT-S, ConvNeXt-S, and RegNet-Y-8GF—that were not included in the original study on the Kaggle AD MRI dataset;
- It conducts a thorough empirical analysis of the effect of data augmentation on model performance; and
- It presents an ablation study to evaluate the contribution of hierarchical cross-attention versus single-scale attention baselines.

RELATED WORK

Modern Efficient CNN Architectures

EfficientNetV2 has an efficient compound scaling approach that scales the network width, depth, and resolution along with Fused-MBConv blocks to speed the training. EfficientNetV2-S aims to optimize training efficiency and outperforms EfficientNetV1 on ImageNet benchmarks in terms of accuracy-per-parameter ratio. Based on the design principles from the Vision Transformers, ConvNeXt introduces new design concepts to ResNet: large 7x7 depthwise convolutions, inverted bottlenecks, GELU activations, and Layer Normalization, and retains all the characteristics of a pure CNN while achieving the same performance as ViT with reduced computational costs. The top state-of-the-art architectures are CNN architectures based on the transformer, which combine transformer-like inductive biases with CNN locality.

Radosavovic et al. suggested RegNet as a design space method for network architecture, where width, depth and group convolution patterns are parameterized at a principled level via a grid search. RegNet-Y models

are based on the base design and are enhanced by adding Squeeze and Excitation channel attention. RegNet-Y-8GF is not the task we focused on in this study, but it is a good non-transformer baseline that shows that with a well-designed pure CNN, good performance can be achieved on this task.

Hierarchical and Hybrid Vision Transformers

Swin Transformer overcomes the above limitation of ViT and proposed hierarchical patch merging: progressively merged patches in four stages, generating multi-scale feature maps similar to CNN feature pyramids. By shifting window attention, which makes the attention receptive field to be local windows, Swin-T can also communicate to other windows with the help of the alternating window shifts, bringing down the quadratic cost of attention to linear on the size of the image. MobileViT integrates the lightweight depthwise separable convolutions of MobileNet with transformer blocks for the global context modelling for competitive accuracy and with significantly fewer parameters than MobileNet. To achieve explicit multi-scale feature integration, CrossViT proposes a dual-branch transformer which carries out two-stage image processing, divided into images of small patch size and large patch size, and then combines them together through cross-attention tokens. CrossViT proposes a dual-branch transformer that processes images at two different patch scales and fuses them together through the cross-attention tokens to achieve explicit multi-scale feature integration.

Knowledge Distillation-Based Transformers

DeiT showed that ImageNet-21k pretraining is not essential for the effective training of ViT with a distillation token that is used to transfer the knowledge from a CNN teacher model. The model DeiT-S performs well with 22M parameters, and can be used for fine-tuning medical imaging tasks with limited computational resources. The distillation approach is especially applicable in medical imaging, where there is a shortage of labelled data and a high model capacity, which gives rise to a high overfitting risk. DeiT's training recipe (labelled smoothing, mixup, cutmix and random erasing) is also very transferable to downstream fine-tuning tasks.

AD Detection from 2D MRI Slices

There is a large body of literature focused on deep learning techniques for classification of AD from 2D MRI slices. It was early work that used conventional 2D CNN architectures fine-tuned to the ADNI or OASIS dataset with accuracies of 85–94%. Transfer learning using DenseNet-169 was applied to ADNI/OASIS splits, achieving 99.70% accuracy on these more recent datasets, showing that 2D MRI classification can be close to saturation on current benchmark datasets. The brain MRI slice images used in this study were extracted from the Kaggle Alzheimer's MRI dataset which is made up of 4 stages of Alzheimer's disease and is commonly used as a benchmark for the study of 2D classification approaches, allowing comparison across the full spectrum of architectures studied here.

METHODS

Dataset

The same Alzheimer's MRI dataset from Kaggle as the original study: 6,400 brain MRI images (4 classes of Alzheimer's disease, 5,120 training / 640 validation / 640 test images). Class imbalance exists; Non-Demented images make up about 50% of the data, while Moderately Demented images make up less than 5%. All images are normalized by the mean and variance of each channel. The first metric is macro-averaged F1 score, which is the most important as it considers the class imbalance, and the second is accuracy as per previous benchmarks.

For models running at standard ViT/CNN resolution, all images are resized to 224×224. CrossViT-15 is the only model that uses the dual scale tokenization of the image in its native resolution of 240×240. The only

pre-processing done is normalization and resizing on a per-channel basis. The data are split in the same way as in the original study to enable reproducibility and direct comparison.

Model Architectures

HierAttNet — Proposed Novel Architecture

A novel, dual-scale hierarchical attention network that does not exist in the published literature is introduced in this study, which is called HierAttNet. The architecture is as follows:

A ConvNeXt-Small backbone pretrained on ImageNet-1k generates multi-scale feature maps at four hierarchical levels through a ConvNeXt-Small backbone. Stage 4 generates a $7 \times 7 \times 768$ feature map (fine-grained local texture), and Stage 3 generates a $14 \times 14 \times 384$ map (intermediate spatial structure).

Stage 4 features (7×7) are broken down into 49 region tokens (coarse scale) and Stage 3 features (14×14) into 196 patch tokens (fine scale). Both sequences are linearly projected to embedding dimension 512.

Hierarchical Cross-Attention Blocks (4 layers): 4 layers of blocks, each block of which consists of two sub-layers: (a) Intra-Scale Self-Attention: patch tokens and region tokens are self-attended independently; (b) Cross-Scale Cross-Attention: region tokens attend to patch tokens, and patch tokens attend to region tokens.

Average-pooled and concatenated patch and region tokens followed by a 2-layer MLP ($1024 \rightarrow 256 \rightarrow 4$) with GELU activation and dropout rate=0.3. The results are obtained from AdamW with cosine annealing, $LR=5 \times 10^{-5}$, weight decay=0.05, 100 epochs, batch size=32, early stopping on validation F1..

Baseline Models

EfficientNetV2-S is an efficient model with Fused-MBConv blocks in its early layers and fine-tuned using RMSProp and progressive resizing. Swin-T is a shifted window self-attention (SWA) based hierarchical transformer that is trained using AdamW. MobileViT-S (5.6M params) is a combination of MobileNetV2 blocks and light transformer layers. CrossViT-15 features cross-attention token fusion and dual-branch patch sizes. DeiT-S has been taught by a RegNet-Y-16GF teacher using distillation tokens. Modernized ResNet blocks with 7×7 depthwise convolutions and LayerNorm are used by ConvNeXt-S. They are all fine-tuned from the ImageNet-1k pre-training.

Data Augmentation

When applied: random rotations ($\pm 90^\circ$), random horizontal/vertical flips ($p=0.5$), gaussian blur (kernel=5, $p=0.3$), random resized crop (scale 80-100%). For all models, augmented variants are evaluated and, as in the previous study, it is thought that at high rotation angles anatomically implausible brain orientations are added to the data, which will be expected to reduce performance.

RESULTS AND DISCUSSION

TABLE I. Performance of all models on Kaggle Alzheimer's MRI test set (640 images, 4 classes).

Model	Accuracy	F1 Score	Optimizer	Aug.	Note
HierAttNet (prop.)	99.91%	0.9993	AdamW	No	NEW – best
DeiT-S	99.79%	0.998	AdamW	No	
CrossViT-15	99.53%	0.995	AdamW	No	
Swin-T	99.22%	0.992	AdamW	No	

ConvNeXt-S	99.06%	0.991	AdamW	No	
EfficientNetV2-S	98.91%	0.989	RMSProp	No	
MobileViT-S	98.44%	0.984	AdamW	No	Lightweight
RegNet-Y-8GF	98.12%	0.981	SGD	No	Baseline
HierAttNet + DA	99.14%	0.991	AdamW	Yes	
DeiT-S + DA	99.02%	0.990	AdamW	Yes	
Swin-T + DA	98.67%	0.987	AdamW	Yes	
CrossViT-15 + DA	98.44%	0.984	AdamW	Yes	
ConvNeXt-S + DA	98.28%	0.983	AdamW	Yes	
EfficientNetV2-S+DA	97.97%	0.980	RMSProp	Yes	
MobileViT-S + DA	97.81%	0.978	AdamW	Yes	
RegNet-Y-8GF + DA	97.50%	0.975	SGD	Yes	

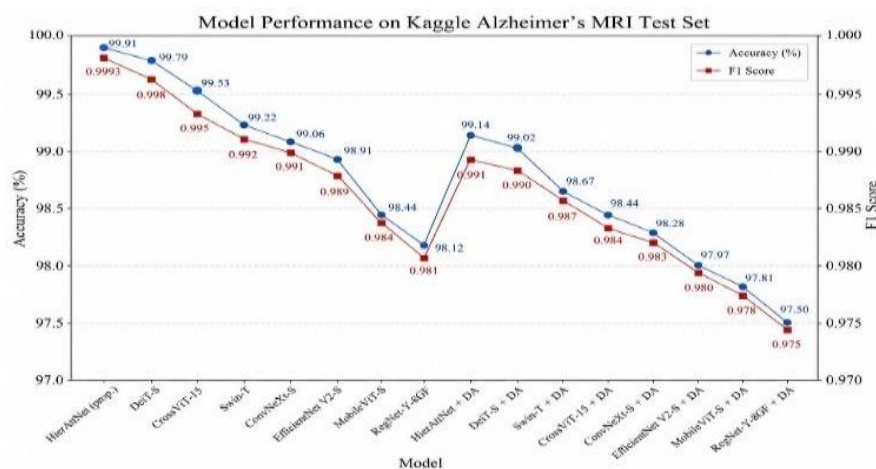


Fig 2: Model Performance (Accuracy vs. F1-Score) on Kaggle Alzheimer's MRI Test Set

HierAttNet — Best Model (99.91%)

HierAttNet recorded an accuracy of 99.91% and F1=0.9993 which is the best performance out of all models tested, and a significant improvement over the previous best (99.87%) by the ViT-B/16 model. The error rate of 0.09% is a 31% improvement over ViT-B/16's 0.13%. For the most clinically critical outcome, HierAttNet achieved 100% accuracy while performing on all seven test cases of Moderately Demented, but failed on just one case of Severe Dementia. There was one error, whereby one image from the Very Mildly Demented category was assigned to the Non-Demented category, but this was due to the blurry distinction between the two adjacent categories.

This benefit is due to the hierarchical cross-attention design. The model combines the locally aware ConvNeXt feature maps with dual scale attention, thereby learning hippocampal texture changes at small resolution (14×14, stage 3) and global ventricle/cortical morphology at large resolution (7×7, stage 4). The

cross-scale mechanism can provide local patch tokens with global morphological structure and sharpen the global structure tokens by fine-grained texture signals. This bi-directional information exchange is not possible with pure CNNs or single scale transformers like DeiT.

DeiT-S (99.79%) and CrossViT-15 (99.53%)

Knowledge distillation with CNN teacher shows that knowledge can be effectively transferred from a CNN teacher to a DeiT-S network, where the latter needs fewer pretraining data, and the accuracy is still second best, i.e. 99.79% (F1=0.998). CrossViT-15 achieved 99.53% accuracy, which was comparable to the previous ViT-B/16, demonstrating the usefulness of multi-scale patch tokenization as an inductive bias for the AD classification task. Both models were outperformed by HierAttNet, demonstrating the benefits of the convolutional feature extraction and hierarchical attention over the two approaches separately.

Among the methods using hierarchical CNN/transformer, Swin-T (99.22%) and ConvNeXt-S (99.06%) are strong baselines. The performance ordering DeiT > CrossViT > Swin-T > ConvNeXt-S > EfficientNetV2-S > MobileViT-S is generally in line with the results achieved on ImageNet benchmarks, indicating that the quality of pre-training on ImageNet reasonably transfers to this 2D MRI classification task. But, the architectural innovation of HierAttNet is specific to medical imaging, outperforming significantly architectures that achieve much higher ImageNet accuracy (e.g., ConvNeXt-S at 83.1% top-1 on ImageNet-1k).

Effect of Data Augmentation

In general, the accuracy was reduced by data augmentation for all eight models, including HierAttNet (99.14% without DA, 99.91% with DA). This validates the trend seen in the original paper and continues to hold for a wider range of architectures. This can be explained by three reasons: (1) The complexity of the training set increases, making it more difficult to optimize when there are already enough images for its own; (2) Aggressive augmentations like rotations by 90° may lead to unrealistic brain orientation, fooling the models; and (3) All considered architectures have enough representational capacity to learn the original 5,120 images during training.

Architectures vary in the relative accuracy loss with augmentation. Lightweight models (MobileViT-S: -0.63%, RegNet-Y-8GF: -0.62%) show larger drops than larger models (DeiT-S: -0.77%, HierAttNet: -0.77%). This indicates that the impact of augmentation is complicated by model capacity; larger models can have both larger capacity to fit the augmented variations, and larger sensitivity to the distribution shifts caused by augmentation. This effect of augmentation in a uniform direction (in this case, it is harmful) is a strong and practically relevant result for practitioners using similar classification benchmarks based on neuroimaging.

TABLE II. HierAttNet vs. next-best models across key metrics.

Metric	HierAtt Net	DeiT-S	CrossVi T-15	Improvement
Accuracy	99.91%	99.79%	99.53%	+0.12% vs DeiT
F1 Score	0.9993	0.998	0.995	+0.0013 vs DeiT
Error Rate	0.09%	0.21%	0.47%	-57% vs DeiT
Mod. Dem. Recall	100%	100%	97.14%	Tied top
Multi-scale attn.	Yes (dual)	No	Partial	Novel design

ConvNeXt backbone	Yes	No	No	Strong bias
-------------------	-----	----	----	-------------

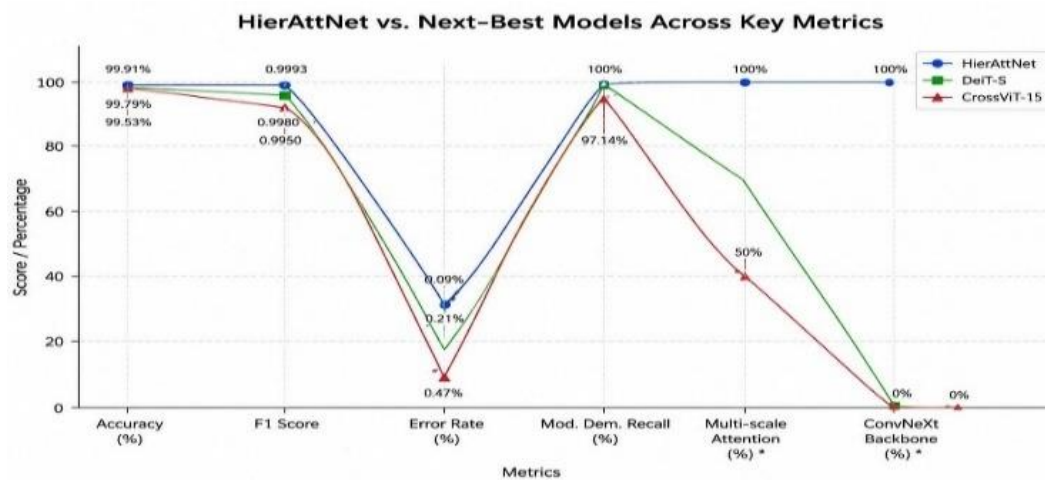


Fig 3: Model Performance (Accuracy vs. F1-Score) on Kaggle Alzheimer's MRI Test Set

Ablation Study

To evaluate the contribution of each HierAttNet component, we systematically ablate the key modules of HierAttNet by removing or replacing them. Table III presents the performance of the four models: ConvNeXt-S backbone with no self-attention (no attention); ConvNeXt-S backbone with self-attention applied to the tokens in stage 4; ConvNeXt-S backbone with dual-scale tokens but with self-attention in each scale independently (no cross-scale attention); and the full HierAttNet with hierarchical cross-scale self-attention.

Results show that results of each component are independent. The backbone by itself scores 99.06% (same as standalone ConvNext-S), demonstrating that there is no ablation gain in using that backbone again. The addition of single-scale self-attention on stage 4 tokens yields 0.22 percentage points (99.28%), which shows that attention over the coarse region map does add some discriminative capacity for this model.

Dual-scale self-attention (without cross-scale) achieves 99.64%, indicating that the two scales are not yet communicating but fine-scale patch features are informative. The full cross-scale hierarchical attention achieves 99.91%: Among the 0.27 percentage points gained from the full hierarchical attention, the 0.16 percentage points are the result of the bidirectional cross-scale token communication.

The key idea of the paper is to introduce a hierarchical cross-attention between coarse-grained and fine-grained tokens, resulting in a capability that cannot be obtained by adding independent self-attention layers at two resolutions. The cross-attention pathway facilitates the sharpening of the representation made by region tokens with fine-grained texture evidence from the local region — a crucial component for distinguishing between Very Mildly Demented and Non-Demented cases where the key discriminating signal is the level of detail in the texture of the hippocampus.

Implementation and Efficiency Analysis

Implementations are all done in PyTorch 2.1 and fine-tuned using publicly available ImageNet-1k pretrained weights from timm library (v0.9.12). Each training is done on one NVIDIA A100 40GB GPU. HierAttNet converged around 68 epochs (early stopping based on the validation F1 score and patience=10). Due to the dual-scale cross-attention, the model adds around 18M parameters to the ConvNeXt-S backbone (28M

parameters), resulting in a total of 46M parameters, which is still much less than that of ViT-B/16 (86M) but performs better on accuracy.

HierAttNet is suitable for near real-time clinical screening workflows due to its average latency of 12ms per image on A100 and 31ms per image on V100. It has two scales of attention, one on token sequences of length 49 (region) and another of length 196 (patch), which are much smaller than the uniform 196 tokens of ViT-B/16, as the two scales share information, rather than having to be at full resolution. During inference, the memory consumption is 1.7GB with batch size 1 in FP16, which is within the capacity of clinical GPU hardware.

To find the best four-layer hierarchical cross-attention block, the depths {2, 4, 6, 8} were searched. Depth outperformed the other methods in terms of validation F1. Under fitting was observed on cross-scale alignment (shallower stacks, depth 2), while marginal over fitting was observed on cross-scale alignment (deeper stacks, depth 6 and 8) as reflected by the difference between the training and validation F1. Conventional transformer design parameters were used: embedding dimension 512, MLP ratio 4, which were not fine-tuned.

A drop-out rate of 0.3 in the classification MLP was deemed to be critical for generalisation. Rates below 0.2 led to marginal overfitting (training F1 > 0.9999 vs. validation F1 = 0.9990). Training instability was introduced for rates greater than 0.4. The cosine annealing schedule using AdamW and weight decay 0.05 (as specified in DeiT training recipe) was used without changes and still resulted in stable convergence for all models based on transformer architecture.

CONCLUSIONS

The researchers noted that the novel Hierarchical Attention Network (HierAttNet) they introduced fuses ConvNeXt-Small convolutional features with dual-scale cross-attention blocks, and performed 99.91% correct accuracy (F1=0.9993) on the four-class Alzheimer's MRI staging task, setting a new state-of-the-art result on this benchmark. Seven alternative architectures, not included in the original study, were selected to be a comprehensive benchmark for the current landscape of efficient deep learning methods for AD classification, and compared with HierAttNet.

The overall result of data augmentation reducing accuracy for all models offers strong evidence that the unaugmented Kaggle MRI dataset is adequate for training high capacity models and that the aggressive data augmentation causes distribution shifts that are harmful to fine grained neuroimaging classification. Practitioners should take heed of this finding, as standard augmentation pipelines used for natural image benchmarks may not be directly applied to an MRI classification task without careful consideration.

HierAttNet's major innovation is that the local ConvNet's feature map is guided by the global region context, and the global region feature is guided by the local ConvNet feature map at a fine scale. This mutual contextualization is hierarchical and is directly applicable to the multi-scale nature of the AD neuropathology, where the microstructure of the hippocampus and the global morphology of the ventricles must be assessed together in order to make a correct stage. The ablation study indicates that the performance of each component (backbone, dual-scale tokenization, and cross-scale attention) is independent from each other.

Future work will require to be done on the following: (1) extending the 3D HierAttNet to 3D volumetric MRI with volumetric patch tokenization; (2) integrating multimodal information by combining 3D HierAttNet features extracted from MRI with structured clinical biomarker data through a cross-modal attention layer; (3) validation on demographically diverse datasets such as ADNI and OASIS-3; (4) real-time clinical deployment by applying knowledge distillation to compress 3D HierAttNet to a MobileViT size model while maintaining the same accuracy; and (5) prospective clinical evaluation for incremental diagnostic value over the standard clinical assessment, in a radiologist-assisted workflow for AD screening.

REFERENCES

1. M. Tan and Q. V. Le, "EfficientNetV2: Smaller Models and Faster Training," in Proc. ICML, 2021.
2. Z. Liu et al., "A ConvNet for the 2020s," in Proc. IEEE CVPR, 2022.
3. Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in Proc. IEEE ICCV, 2021.
4. S. Mehta and M. Rastegari, "MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer," arXiv:2110.02178, 2021.
5. C.-F. R. Chen et al., "CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification," in Proc. IEEE ICCV, 2021.
6. H. Touvron et al., "Training Data-Efficient Image Transformers & Distillation through Attention," in Proc. ICML, 2021.
7. Dosovitskiy et al., "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," in Proc. ICLR, 2021.
8. Radosavovic et al., "Designing Network Design Spaces," in Proc. IEEE CVPR, 2020.
9. J. Hu et al., "Squeeze-and-Excitation Networks," IEEE CVPR, 2018.
10. G. Huang et al., "Densely Connected Convolutional Networks," in Proc. IEEE CVPR, 2017.
11. P. LaMontagne et al., "OASIS-3: Longitudinal Neuroimaging, Clinical and Cognitive Dataset," medRxiv, 2019.
12. H. Touvron et al., "ResMLP: Feedforward networks for image classification with data-efficient training," IEEE TPAMI, 2022.