A Comparative Analysis of Statistical, Regional CNN, and Sequential Transformer Approaches for Alzheimer's Disease Classification

Tri Huynh[1], Xuan Hoc Pham[1], Nhu Nguyen[1], Thi Thu Nguyen[2] , Huong Ha[1,*] and Lua Ngo[1,**]
[1] School of Biomedical Engineering, International University – Vietnam National University – Ho Chi Minh City, Vietnam
[2] School of Electrical and Electronic Engineering, Hanoi University of Industry, Ha Noi, Viet Nam
[*]E-mail: htthuong@hcmiu.edu.vn
[**]E-mail: ntlua@hcmiu.edu.vn

***Abstract*— Alzheimer's disease classification from neuroimaging data represents a critical challenge in computational neuroscience, with significant implications for early diagnosis and intervention. This study presents a systematic comparison of three representative computational paradigms for AD classification using the ADNI dataset comprising 1,056 T1-weighted MRI images from 370 AD patients and 686 cognitively normal controls. We implemented and evaluated three approaches: (1) a statistically-based method using FreeSurfer-derived morphometric features with machine learning classifiers, (2) a region-based approach employing ResNet18 architecture focused on hippocampal segmentation, and (3) a sequence-based method utilizing DeiT with Group Query Attention for processing axial slice sequences. The statistically-based approach achieved the highest performance with 91% accuracy using Random Forest after comprehensive feature selection, reducing 274 initial features to 138 through variance filtering, ANOVA testing, and correlation analysis. The region-based ResNet18 method demonstrated 73-74% accuracy on hippocampal regions, while the sequence-based transformer approach achieved 76% accuracy processing 50-slice axial sequences. Results demonstrate that traditional statistical methods remain competitive with modern deep learning approaches on limited medical datasets, while highlighting the need for larger datasets and architectural optimization to fully realize the potential of transformer-based sequence modeling in neuroimaging applications.**

I. INTRODUCTION

Alzheimer's disease (AD) classification from neuroimaging data represents a critical challenge in computational neuroscience, with significant implications for early diagnosis and intervention. Multiple computational paradigms have emerged to address this challenge. Traditional machine learning approaches using statistical features have shown promise, with Shaker et al. achieving 93.95% accuracy using multimodal feature fusion (cognitive scores, genetics, MRI, PET, neuropsychological battery) with Random Forest classifiers [1]. CNN have demonstrated competitive or superior performance on imaging data alone, as Qi et al. reported 90%, 95%, and 95% classification accuracy using SVM, 3D-VGGNet, and 3D-ResNet respectively, with ResNet providing superior classification performance and more accurate localization of disease-associated brain regions through Grad-CAM visualization [2]. Later studies improved deep learning models. Nithya et al. reported 95% accuracy with ResNet-50, enhanced by preprocessing methods such as CLAHE and BADF. This highlights the importance of preprocessing in AD classification [3]. Sequence modeling is an emerging direction. Aliyu et al. reached 98% accuracy with a Vision Transformer (ViTB32), and Sait et al. reported 99% by combining ViT with a time-series transformer [4], [5]. However, systematic comparison of these paradigms using identical datasets and evaluation protocols remains limited, particularly for approaches that can capture temporal relationships within 3D neuroimaging sequences.

To address this gap, we implement and systematically compare three representative approaches with practical modifications tailored for AD classification. We selected ResNet18 as a representative CNN baseline due to its widespread use in medical imaging, and DeiT as a data-efficient transformer variant shown to perform well on moderate-sized datasets [6], [7]. Together with the statistical approach, these choices provide a balanced benchmarking framework across traditional, convolutional, and transformer paradigms. The first approach extracts morphometric measurements through FreeSurfer processing with advanced feature selection techniques; the second implements ResNet18 focused on hippocampal, thalamic, and amygdala regions; and the third applies DeiT with Group Query Attention to process axial slice sequences from 3D MRI volumes. Through a comprehensive evaluation on the ADNI dataset, we aim to understand performance versus complexity trade-offs across these paradigms, providing practical insights for method selection in real-world scenarios. This comparative analysis provides valuable guidance for choosing computational approaches while highlighting each methodology's strengths and limitations for AD classification.

979-8-3315-7206-8/25$31.00 ©2025 IEEE

II. MATERIAL AND METHOD

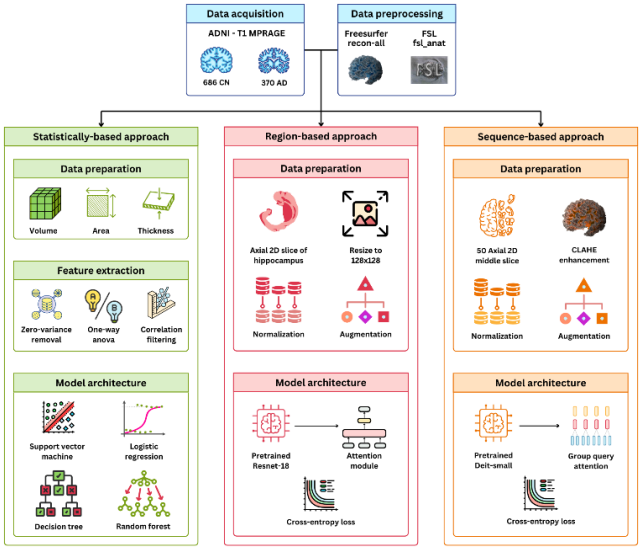

Figure 1: Study diagram

1. Data acquisition

Data used in this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org.

This study utilized 1,056 T1-weighted MP-RAGE MRI images from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, comprising 370 subjects diagnosed with Alzheimer's disease and 686 cognitively normal controls. The ADNI dataset was selected for its standardized acquisition protocols, rigorous quality control measures, and extensive clinical validation, making it the gold standard for Alzheimer's disease neuroimaging research. All images were acquired using 1.5T or 3T MRI scanners with consistent MP-RAGE sequences optimized for structural brain analysis. Since each paradigm processes fundamentally different input representations, distinct preprocessing pipelines were required: morphometric features for the statistical approach, hippocampal slices for the CNN, and axial sequences for the Transformer.

2. Data preprocessing

All MRI volumes underwent standardized preprocessing using established pipelines tailored for each analytical approach. For statistical and sequence-based methods, the FreeSurfer recon-all pipeline was applied to extract cortical and subcortical morphometric measurements, including regional volumes, cortical thickness, and surface area calculations [8]. For the regional CNN approach, FSL's fsl_anat pipeline was utilized for structural preprocessing, including bias field correction, brain extraction, and tissue-type segmentation, followed by specialized hippocampus, thalamus, and amygdala segmentation using FIRST [9]. To ensure anatomical consistency all MRI volumes were registered to a MNI152 template using FSL's linear registration tool (FLIRT).

3. Statistically-based approach

Data preparation

There are a total of 274 features collected from the output stats folder of FreeSurfer's recon-all pipeline. Statistics features, including subcortical and other volumetric measurements, cortical area, thickness, and volume measurements, which are based on the Desikan-Killiany atlas.

Zero variance feature removal

Any feature with a variance less than or equal to zero was eliminated from the dataset. This process removes features that are constant across all subjects.

ANOVA F-test between CN and AD groups

After removing constant variables, a one-way Analysis of Variance (ANOVA) was performed to determine features that are significantly different between the CN and AD groups. The result of the ANOVA test included the F-score and p-value of each feature. Features with a p-value greater than 0.05 were considered not significant and were removed.

High correlation filtering

After the ANOVA test, a correlation analysis was conducted on the remaining features to reduce multicollinearity. A Pearson correlation matrix was generated for the features identified as significant in the previous step. For any group of features where the absolute correlation coefficient was greater than or equal to 0.8, only the feature with the lowest p-value from the ANOVA test was retained. Features that were not highly correlated with any others were automatically kept. This process resulted in the final set of features used for model training.

Model training

70% of the dataset is allocated for training and 30% for testing. The target variable, representing the subject's diagnosis (CN or AD), was encoded into numerical values. To ensure that all features contributed equally to the model performance, the feature values were standardized to have a mean of 0 and a standard deviation of 1. The scale was fitted on the training data and then used to transform both the training and testing sets.

To optimize the performance of the machine learning models, hyperparameter tuning was performed using GridSearchCV. A stratified 5-fold cross-validation strategy was employed to ensure that each fold was representative of the overall class distribution. The models that underwent this tuning process were Support Vector Machine (SVM), Logistic Regression, Decision Tree, and Random Forest. The best-performing version of each model was then selected for the final evaluation.

4. Region-based approach

Data preparation

The segmented 3D hippocampus volumes were sliced along the axial plane into 2D images. Each axial slice represents a cross-section through the hippocampus region. To remove non-informative slices, the mean intensity of each slice was computed. Only slices with a mean intensity greater than a fixed threshold of 50 were retained.

The selected 2D slices were resized to 128x128 pixels for uniform input size. Each image was then normalized using a fixed mean and standard deviation of [0.5, 0.5, 0.5], effectively scaling pixel values to the range [-1, 1]. This normalization strategy helps stabilize training, especially in the absence of precise dataset-level intensity statistics, which is common in medical imaging. To enhance the robustness and generalization ability of the model, data augmentation was applied during training. These augmentations included random horizontal flips, rotations up to ± 15 degrees, light affine transformations, and mild color jittering, which was applied conservatively given the grayscale characteristics of MRI data.

Model architecture

The model utilizes transfer learning with a pretrained ResNet-18 architecture by initializing with ImageNet pretrained weights, fine-tuning only the final convolutional block (layer 4) while freezing earlier layers. To improve focus on subtle anatomical changes in the hippocampus, a SimpleAttention module is integrated into the network. This module combines channel attention (which reweights feature maps based on importance) and spatial attention (which emphasizes informative regions in space), helping the model better localize early Alzheimer's pathology in 2D MRI slices. The classifier consists of a fully connected layer with 512 input features and 2 outputs, with a dropout layer (rate = 0.5) to reduce overfitting. The training objective uses Cross Entropy Loss with softmax activation function, quantifies the difference between predicted class probabilities and the true labels. It is particularly effective when modeling probabilistic outputs in mutually exclusive class settings. A ReduceLROnPlateau scheduler adjusts the learning rate based on validation loss to improve convergence.

Experimental Setups

Table 1: Model configuration of the region-based approach

Parameter	Value
Optimizer	AdamW
Learning rate	1e-3
Weight Decay	1e-4
Loss function	Categorical cross-entropy
Batch size	32
Epochs	20

5. Sequence-based approach

Data preparation

The sequence-based approach processes 3D MRI volumes by extracting axial slice sequences for transformer-based analysis. Each MRI volume is stored as NPZ files containing preprocessed 3D arrays, yielding 50 middle axial slices at 224×224 pixel resolution, normalized to the [0,1] range through division by 255 for consistent input scaling.

The data pipeline utilizes TensorFlow's optimized data loading framework with parallel processing and prefetching capabilities. Optional Contrast Limited Adaptive Histogram Equalization (CLAHE) enhancement is applied to improve image contrast across all slices simultaneously. For training augmentation, transformations are applied across the entire slice sequence, including random rotation (±15°), translation (±5%), shear (±5°), and horizontal flipping. Each slice sequence is expanded with a channel dimension, resulting in tensors of shape (50, 1, 224, 224) for model compatibility. The dataset construction incorporates shuffle buffering for training randomization, configurable batch processing, and binary classification labeling (CN=0, AD=1), with memory optimization for efficient GPU utilization.

Model architecture

Our sequence-based approach combines DeiT with an attention module for temporal modeling. Each axial slice is processed by a pre-trained DeiT-small model, and the sequence of features is then analyzed by a transformer encoder. The model uses positional encoding to preserve slice order and produces the final classification through a softmax activation.

Experimental Setups

Table 2: Model configuration of the sequence-based approach

Parameter	Value
Optimizer	AdamW
Learning rate	1e-4
Weight Decay	1e-4
Loss function	Sparse categorical cross-entropy
Batch size	32
Epochs	300
Training framework	Distributed training with multiple GPUs

6. Evaluation metric

All three approaches were evaluated using six key classification metrics to ensure a comprehensive performance assessment. Accuracy measured overall correct predictions, while precision and recall evaluated positive class prediction quality and detection capability, respectively. F1-score provided a balanced assessment by combining precision and recall. Specificity quantified true negative identification rates, and AUC-ROC curves assessed discriminative performance across different threshold values. These metrics collectively enabled a thorough evaluation of each approach's diagnostic performance while accounting for potential dataset imbalances.

III. RESULTS

1. *Statistically-based approach*

In the first step, 11 features with zero variance were detected and removed from the dataset. After this step, 263 out of 274 initial features remained and went through the ANOVA test. According to the result of the ANOVA test, 215 features were significantly different between the two groups (p-value < 0.05) **(Fig. 2)**. The remaining 48 features were excluded from further steps since they did not show a significant difference. After the high correlation filtering step, the number of features was reduced from 215 to a final set of 138.

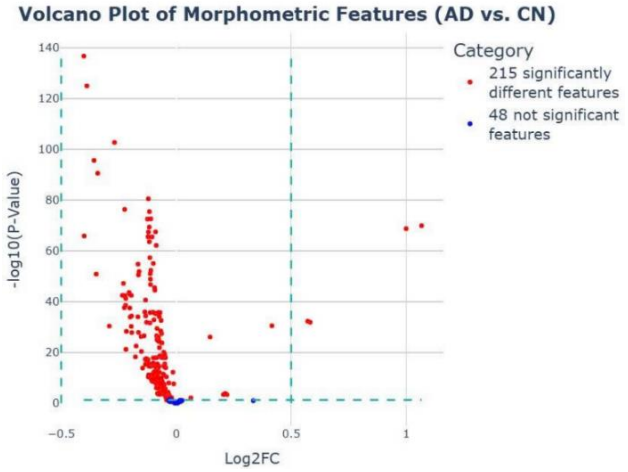

Figure 2: Selection of 215 discriminative features based on the ANOVA test's p-value

The Random Forest model demonstrated the best overall performance with the fewest misclassifications, totaling 30 errors, comprised of 14 false negatives and 16 false positives. In contrast, the Decision Tree model showed the weakest performance, making 49 incorrect predictions, which included 24 false negatives and 25 false positives. Both SVM and Logistic Regression models show strong and comparable performance (31 misclassifications), which were close to Random Forest in overall accuracy **(Fig. 3)**.

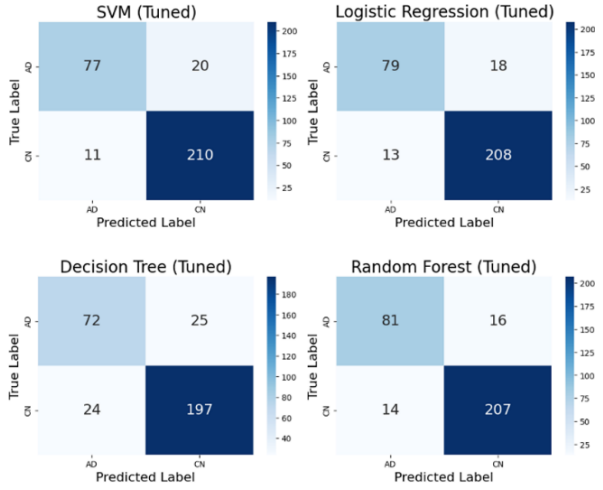

Figure 3: Confusion matrix of all model

The statistical approach achieved strong performance across all classifiers, though the four primary metrics were not identical as initially reported. Support Vector Machine (SVM) reached 90.25% accuracy, 0.7938 precision, 0.875 recall, and 0.8324 F1. Logistic Regression achieved 90.25% accuracy, 0.8144 precision, 0.8587 recall, and 0.8360 F1. The Decision Tree performed slightly lower with 84.59% accuracy, 0.7423 precision, 0.7500 recall, and 0.7461 F1. Random Forest yielded the best overall performance, with 90.57% accuracy, 0.835 precision, 0.8526 recall, and 0.8438 F1. These corrected values confirm that the statistical approach remains a competitive baseline for AD classification. Furthermore, SVM, Logistic Regression, and Random Forest all achieve an impressive AUC of 0.95, which outperforms the Decision Tree's AUC of 0.86 **(Fig. 4)**.

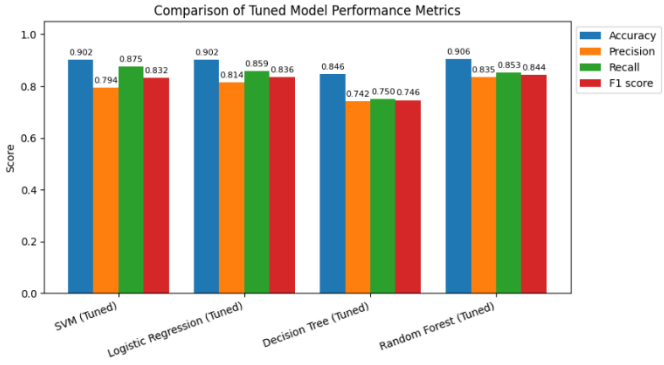

Figure 4: Comparison of tuned model performance

2. *Regional-based approach*

Based on a preprocessing pipeline that applied a threshold of 50 (corresponding to the 20th-25th percentile) to select informative 2D slices from 3D MRI volumes, the ResNet18-2DMRI model was trained and evaluated on training, and validation set. It achieved an overall accuracy of 74% on the validation. However, a consistent tendency to misclassify AD slices as CN was observed. Specifically, the sensitivity for CN was significantly higher (83% validation) compared to AD (55% validation), likely due to class imbalance, as CN slices were nearly twice as numerous as AD slices **(Fig. 5)**.

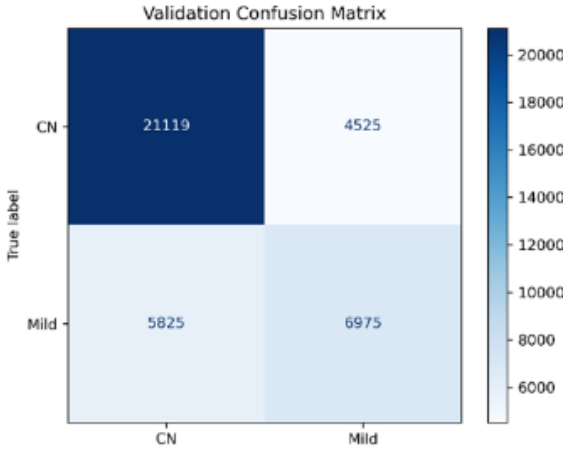

Figure 5: Confusion matrix of validation set

The fact that a large proportion of AD slices are misclassified as CN is further reflected in the validation loss and recall **(Fig. 6)**. The left plot shows that the training loss decreases steadily across epochs, indicating an effective learning process. Validation loss fluctuated and spiked at epoch 12, and after epoch 10 the model showed signs of overfitting. This effect was reduced by automatically adjusting the learning rate and saving the best-performing model during training.

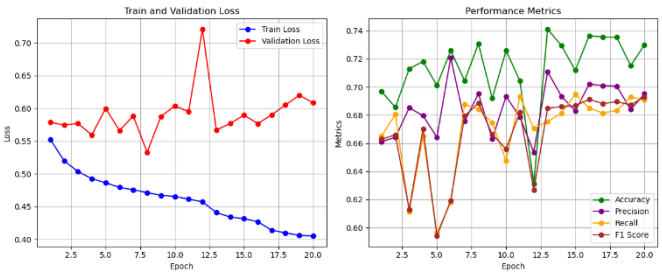

Figure 6: Model performance result

The performance metrics in **Fig. 6** illustrate detailed evaluation metrics on the validation set. Accuracy fluctuates within the range of 69 - 74%, indicating a relatively stable overall classification performance across epochs. Precision and F1 score exhibit more consistency than recall, with slight variations around 0.7. In contrast, recall varies more significantly, ranging from approximately 0.5 to 0.7, indicating instability in the model to detect AD samples. The consistently lower recall compared to precision, particularly in epochs such as 5, 11, and 13, highlights a high false negative rate, where many AD cases are misclassified as CN.

3. Sequence-based approach

The sequence-based approach demonstrated moderate performance in classifying Alzheimer's disease from axial MRI slice sequences. The hybrid architecture, comprising 5.2 million parameters, achieved a final test accuracy of 76.4% and a best validation accuracy of 79.2% after 152 training epochs.

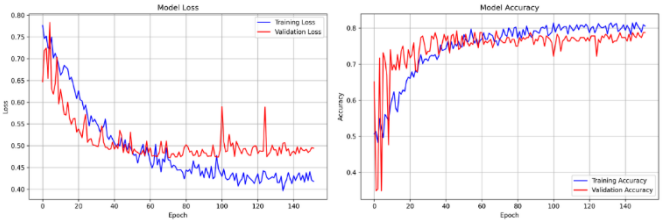

Figure 7: Model training performance

Fig. 7 illustrates the comprehensive training dynamics throughout the learning process. The model exhibited good convergence characteristics, with training loss decreasing gradually from an initial value of 0.77 to a final value of 0.43. The validation loss demonstrated fluctuating behavior, but still gradually decreased and converged to 0.5, indicating minimal overfitting tendencies. Training accuracy progressed smoothly to reach 82%, while validation accuracy achieved 76% with consistent performance in the final epochs.

Fig. 8 presents the confusion matrix analysis, demonstrating the model's classification performance across cognitive states. The model successfully identified 118 out of 138 CN cases (85.5% sensitivity) and 51 out of 74 AD cases (68.9% sensitivity). The model achieved precision rates of 85.5% for CN classification and 71.8% for AD classification, with an overall accuracy of 79.7%.

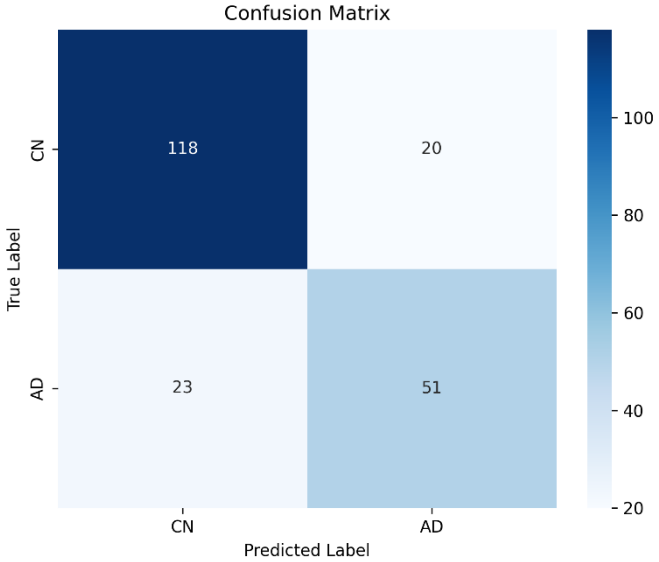

Figure 8: Confusion matrix of sequence-based model

IV. DISCUSSIONS

Our systematic comparison of three computational paradigms provides valuable insights into different approaches for AD classification. The statistical approach achieved 91% accuracy with Random Forest, confirming that traditional machine learning methods remain effective on neuroimaging features. While it offers robust performance with relatively low computational cost, its reliance on handcrafted features may limit scalability compared to deep learning methods. The model showed particularly strong performance in identifying key brain regions associated with AD pathology, with hippocampal and entorhinal cortex features ranking among the most discriminative variables.

The region-based ResNet18 approach achieved 73–74% accuracy, showing that CNNs can capture subtle morphological changes in key brain regions. Although performance was lower than some prior reports, this analysis highlights the discriminative role of hippocampal structures and points to class imbalance as an important challenge. These results indicate that CNNs remain valuable tools when paired with sufficient data and balanced sampling.

The sequence-based approach achieved about 76% accuracy, showing that transformer models can capture temporal patterns in neuroimaging data. While performance was lower than larger transformer architectures reported in the literature, our results confirm the feasibility of applying data-efficient transformers to moderate-sized datasets. This highlights their

promise as a foundation for future studies with larger and more diverse data. Despite these limitations, our approach validated that transformer architectures can effectively model sequential brain imaging patterns, establishing a foundation for future work. We also note that achieving state-of-the-art performance typically requires larger datasets and stronger optimization, which were beyond the scope of this study but remain important directions moving forward. In addition, differences in preprocessing across paradigms may have influenced performance, and future work should include ablation studies to disentangle preprocessing effects from model architecture.

Several promising directions emerge from this comparative analysis. First, incorporating Mild Cognitive Impairment as an intermediate diagnostic category would enable three-class classification, providing more clinically relevant staging information. Second, an ensemble approach combining strengths from all three paradigms—statistical robustness, regional specificity, and sequential patterns—could potentially overcome individual method limitations. Third, expanding the dataset size and addressing class imbalance through advanced sampling techniques could substantially improve model performance. Additionally, validation on Vietnamese demographic data would enhance model generalizability across different populations. Finally, future work should also include multi-center validation, statistical significance testing, error analysis, and direct comparisons with recent state-of-the-art Transformer models to further strengthen the robustness and impact of the findings.

V. CONCLUSIONS

This comparative study evaluated three computational paradigms for Alzheimer's disease classification on the ADNI dataset. The statistically-based approach achieved 91% accuracy, the region-based ResNet18 method reached 73-74% accuracy, and the sequence-based DeiT approach attained approximately 76% accuracy. Performance variations highlight the critical impact of dataset characteristics, preprocessing strategies, and architectural choices on classification outcomes. While traditional statistical methods demonstrated robust performance with limited data, deep learning approaches showed potential but required careful optimization. The sequence-based transformer approach, though promising, needs further development for medical imaging applications. Future work should focus on ensemble methods combining all three paradigms, multi-class classification including MCI, and validation on diverse populations to advance toward clinically deployable AD diagnostic systems.

REFERENCES

[1] S. El-Sappagh, J. M. Alonso, S. M. R. Islam, A. M. Sultan, and K. S. Kwak, “A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer’s disease,” *Sci Rep*, vol. 11, no. 1, Jan. 2021, doi: 10.1038/s41598-021-82098-3.

[2] Q. Li and M. Q. Yang, “Comparison of machine learning approaches for enhancing Alzheimer’s disease classification,” *PeerJ*, vol. 9, p. e10549, Feb. 2021, doi: 10.7717/peerj.10549.

[3] V. P. Nithya, N. Mohanasundaram, and R. Santhosh, “An Early Detection and Classification of Alzheimer’s Disease Framework Basedon ResNet-50,” *CMIR*, vol. 20, Oct. 2023, doi: 10.2174/1573405620666230825113344.

[4] A. Abubakar, Y. Jibrin, M. B. Maina, and A. B. Maina, “Classification of Alzheimer’s Disease Using Cnn-Based Features and Vit-Global Contextual Patterns from MRI Images,” 2024, *Elsevier BV*. doi: 10.2139/ssrn.4811438.

[5] S. Alp *et al.*, “Joint transformer architecture in brain 3D MRI classification: its application in Alzheimer’s disease classification,” *Sci Rep*, vol. 14, no. 1, Apr. 2024, doi: 10.1038/s41598-024-59578-3.

[6] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” Dec. 10, 2015, *arXiv*: arXiv:1512.03385. doi: 10.48550/arXiv.1512.03385.

[7] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, “Training data-efficient image transformers & distillation through attention,” Jan. 15, 2021, *arXiv*: arXiv:2012.12877. doi: 10.48550/arXiv.2012.12877.

[8] “recon-all - Free Surfer Wiki.” Accessed: Aug. 27, 2025. [Online]. Available: https://surfer.nmr.mgh.harvard.edu/fswiki/recon-all

[9] “fsl_anat.” Accessed: Aug. 27, 2025. [Online]. Available: https://fsl.fmrib.ox.ac.uk/fsl/docs/#/structural/fsl_anat