

# Ensemble Vision Transformer for Dementia Diagnosis

Fei Huang, Anqi Qiu, and the Alzheimer's Disease Neuroimaging Initiative

**Abstract**—In recent years, deep learning has gained momentum in computer-aided Alzheimer's Disease (AD) diagnosis. This study introduces a novel approach, Monte Carlo Ensemble Vision Transformer (MC-ViT), which develops an ensemble approach with Vision transformer (ViT). Instead of using traditional ensemble methods that deploy multiple learners, our approach employs a single vision transformer learner. By harnessing Monte Carlo sampling, this method produces a broad spectrum of classification decisions, enhancing the MC-ViT performance. This novel technique adeptly overcomes the limitation of 3D patch convolutional neural networks that only characterize partial of the whole brain anatomy, paving the way for a neural network adept at discerning 3D inter-feature correlations. Evaluations using the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset with 7199 scans and Open Access Series of Imaging Studies-3 (OASIS-3) with 1992 scans showcased its performance. With minimal preprocessing, our approach achieved an impressive 90% accuracy in AD classification, surpassing both 2D-slice CNNs and 3D CNNs.

**Index Terms**—Monte Carlo sampling, 3D Patch, Vision Transformer, structural magnetic resonance imaging, Alzheimer's disease

## I. INTRODUCTION

ALZHEIMER'S disease (AD) is a neurodegenerative condition associated with aging, marked by a gradual decline in memory, cognitive function, and physical capabilities, as documented in previous studies [1], [2]. Despite sustained research efforts, a definitive cure for AD has yet to be found. Present treatments can merely slow down its advancement, with the most substantial benefits observed when AD is detected in its early stages. Nevertheless, the accurate diagnosis of AD or Mild Cognitive Impairment (MCI) continues to present a significant clinical hurdle.

This research/project is supported by STI 2030—Major Projects (No.2022D0209000), the National Research Foundation, Singapore, and the Agency for Science Technology and Research (A\*STAR), Singapore, under its Prenatal/Early Childhood Grant (Grant No. H22POM0007), and by the Hong Kong global STEM scholar scheme. (Corresponding author: Anqi Qiu.)

Fei Huang is with the School of Computer Engineering and Science, Shanghai University, Shanghai200240, China (e-mail: huangfei8866@shu.edu.cn)

Anqi Qiu is with Department of Health Technology and Informatics, the Hong Kong Polytechnic University, Department of Biomedical Engineering, National University of Singapore, Singapore, and Department of Biomedical Engineering at Johns Hopkins University. (e-mail: anqi.qiu@polyu.edu.hk).

In recent years, structural Magnetic Resonance Imaging (sMRI) has emerged as a potent tool for early brain degeneration detection in routine clinical practice, providing invaluable insights into the evolving morphological characteristics associated with AD [3], [4]. The field of deep learning has garnered considerable attention for its application in AD diagnosis and prediction based on sMRI [5]–[7]. Deep learning's exceptional adaptability and capacity to extract crucial, distinctive features from brain MRI images have catalyzed substantial advancements in this area [8]. This technology holds the potential to significantly enhance the precision and accuracy of AD diagnosis, addressing the persistent clinical challenge highlighted earlier.

Historically, within the realm of sMRI-based deep learning methodologies, the segmentation of MRI datasets into slices or regions at various scales has been a common approach. This strategy aims to enhance feature extraction and improve the detection of localized abnormal brain structural changes. These sMRI-based studies can primarily be categorized into four groups: slice-level, patch-level, region-of-interest (ROI)-level, and subject-level analyses [8]. In slice-level investigations, well-established Convolutional Neural Network (CNN) architectures such as VGGNet, AlexNet, and ResNet are often employed [5], [9], [10]. These models process one or a few 2D brain structural image slices with 2D CNNs, yielding efficiency but potentially overlooking the brain's intricate three-dimensional nature, resulting in suboptimal accuracy in AD diagnosis.

In contrast, 3D CNNs, when applied to entire brain images, capture the brain's three-dimensional spatial intricacies, offering a more holistic understanding of its structure [11]–[13]. However, these models often grapple with overfitting issues due to their high-dimensionality. Patch-level and ROI-level methodologies are designed to alleviate computational strain [3], [14]. When employing a prior knowledge on brain regions affected early in AD, such as the hippocampus, 3D patch CNNs often require intensive image processing including brain segmentation and registration [8], [15]. Patch-level methods, on the other hand, focus on local structural changes in MRIs. Nevertheless, a significant challenge lies in effectively amalgamating these local patches into a coherent global representation of brain structure.

Ensemble learning, a technique that leverages a collective of multiple learners to make predictions, has garnered significant attention in recent years. This approach involves the amalgamation of results from multiple learners to enhance prediction

accuracy. Extensive research in this domain has investigated various input strategies, including slices or patches, tested alongside multiple deep learning architectures [6], [10], [16]. This research consistently demonstrates superior performance in terms of AD diagnosis accuracy when compared to individual learners. However, it's worth noting that traditional ensemble methods, which incorporate multiple learners, typically entail increased training demands compared to a single-learner approach.

Recent developments have witnessed the integration of CNNs and transformers, leveraging the strengths of both methodologies to attain more robust performance in AD diagnosis [17]–[19]. This fusion, which combines CNNs' capacity to capture local information with Transformers' global information processing capabilities, consistently yields improved diagnostic accuracy for AD [17], [18]. Nevertheless, it's worth noting that when dealing with limited training data, as is often the case with small MRI datasets, the model's generalization ability may be compromised [20].

This study takes advantage of 3D CNN and the idea of ensemble and introduces a novel Monte Carlo Vision Transformer Ensemble (MC-ViT) method for the diagnosis of AD. The MC-ViT leverages a 3D Patch Network-based CNN to extract spatial features from sMRI images. Then, it employs Monte Carlo sampling to generate various samples of the whole brain features and utilizes ViT to capture their relationships and excel AD diagnosis. Unlike conventional sampling approaches that rely on non-probability sampling, this study opted for Monte Carlo sampling from a probability distribution. This method streamlines the selection of patches by prioritizing the importance of their features in classification outcomes. Our experiments utilizing the ADNI and OASIS datasets highlight the effectiveness of our approach, surpassing several benchmark methods slice-level [21], patch-level [22], ROI-level [23], subject-level [24] in terms of accuracy. Furthermore, our experiments also outperform two benchmark methods utilizing ViT models, including ViT with slices [25], [26], and whole brain MRI [27], [28].

The key contributions of this study are as follows:

- A novel integration of the 3D Patch Network and Vision Transformer, combined with Monte Carlo sampling, offering improved AD diagnosis by effectively extracting both local spatial and overarching structural features from sMRI scans.
- An ensemble technique that embraces diversity, consolidating predictions from a wide range of features through Monte Carlo sampling.
- A single base learner approach, optimizing computational training while creating an efficient deep learning ensemble model.
- A model that outperforms established 2D-slice and 3D CNNs in AD diagnostic accuracy.

The following sections of this manuscript explore the relevant literature in machine learning and deep learning for AD diagnosis (Section II). We then provide a comprehensive overview of our proposed framework (Section III), followed by a discussion of preprocessing techniques and an in-depth exploration of experiments and parameters (Section IV). The

paper culminates with a robust discussion and our final conclusions in Section V, respectively.

## II. RELATED WORK

This section concisely overviews prior research on computer-aided AD diagnosis methods utilizing sMRI data. We review pertinent literature on i) Convolutional Neural Networks, ii) Transformers, and iii) Ensemble Learning.

### A. Convolutional Neural Networks

In the realm of sMRI-based deep learning methods for AD diagnosis, different scales of feature representations have been explored, including slice-level, ROI-level, subject-level, and patch-level approaches.

In the context of slice-level methods, researchers typically select 2D brain images, such as coronal, sagittal, or axial slices, from each subject and collectively use them for classification tasks. For instance, Neffati et al. [29] employed 2D discrete wavelet transform texture features extracted from coronal slices for AD classification. Jain et al. [30] and Tanveer et al. [31] used 2D CNNs to enhance AD diagnosis with informative 2D slices. Nevertheless, focusing solely on individual slices may overlook the holistic understanding of the entire brain structure. In related research, previous studies [32], [33] have attempted to leverage the combined capabilities of CNNs and Long Short-Term Memory (LSTM) networks to capture intricate relationships within the data.

In contrast, employing 3D CNN approaches on the entire brain allows for the incorporation of the three-dimensional spatial relevance of the brain [24]. However, utilizing 3D CNN on the whole brain entails a higher number of parameters and increased computational complexity. Consequently, alternative approaches utilize ROI-level [34], [35] and patch-level [3], [22] inputs for the 3D CNN. Liu et al. [34], [35] conducted a study wherein they extracted 3D texture features by partitioning the brain's ROIs into various configurations, resulting in the development of multiple hierarchical networks. Nonetheless, region levels are typically predetermined based on biological prior knowledge or anatomical brain atlases, such as hippocampal and medial temporal ROIs [36]. The utilization of 3D patches, akin to the slice-level approach, mitigates computational complexity while retaining more anatomical information than slice-level methods. A study introduced a dual-attention multi-instance deep learning network known as DA-MIDL, which leverages spatial attention blocks to identify discriminative features from patches within sMRI data, thereby enhancing early AD diagnosis [3]. Nevertheless, an approach solely focused on individual patches may encounter difficulties in selecting patches containing significant and informative content, potentially neglecting the comprehensive integrity of the entire brain structure.

### B. Transformers

In recent years, the transformer model [37] has emerged as a promising approach in the field of medical imaging, building on its remarkable success in natural language processing

and continuous advancements in computer vision. The Vision Transformer (ViT) [27], in particular, represents a departure from traditional CNN architectures, embracing a pure transformer design. Serving as an innovative feature extractor, ViT places a strong emphasis on patch-level attention rather than pixel-level attention. Notably, both in computer vision and medical imaging [17], [38], [39], ViT has demonstrated its superiority over CNN, consistently delivering outstanding performance.

A novel technique for brain age estimation has been proposed, utilizing a 2D CNN on complete sMRI data and localized 2D slices, which are then combined through self-attention to capture both global context and fine-grained features [38]. Altay et al. [39] have devised an attention transformer model aimed at proficiently amalgamating features across slices to forecast the preclinical stage of AD. Nevertheless, there remains potential for enhancing the accuracy of AD diagnosis, particularly by more effectively incorporating local spatial features into a unified global feature representation. In light of this, Hu et al. [17] propose the VGG-TSwinformer model featuring a sliding-window attention mechanism designed for meticulous fusion of spatial features to monitor brain atrophy progression using longitudinal sMRI images in patients with MCI.

### C. Ensemble Learning

Recent research has unequivocally demonstrated the efficacy of leveraging ensemble methodologies in the diagnosis of AD [40]. Instead of relying solely on the decisions of individual models, an ensemble approach has emerged as a promising strategy for enhancing diagnostic accuracy.

Numerous studies [22], [41] utilize landmark detection to pinpoint the most discriminative patches, followed by employing multiple identical CNNs to extract distinct features from these patches. Liu et al. [41] adopt a straightforward approach by consolidating the predicted patch labels from 50 landmark locations obtained by CNNs, employing a majority voting strategy. However, relying solely on a simple majority voting strategy may result in the loss of comprehensive brain information. In another investigation Liu et al. [22], features from diverse patches are extracted and combined using multiple CNNs, followed by employing fully connected (FC) layers to make diagnostic decisions for Alzheimer's disease. Training each patch exclusively with its corresponding CNN enables it to specialize in extracting features relevant to its specific content or context. Nonetheless, the utilization of multiple CNNs entails higher computational demands and time consumption. Moreover, exclusive training of each patch with its corresponding CNN may result in the loss of comprehensive brain information and a lack of diversity.

The ensemble approach, outlined in various studies [15], [25], [33], involves merging the outputs of disparate models utilizing a weighted mean methodology. This technique harnesses the strengths of model diversity and exploits their complementary features, resulting in enhanced accuracy for diagnosing AD. Notwithstanding its impressive performance, it is important to recognize the computational demands inherent in this approach. Integrating multiple models, especially

those with diverse architectures, can be time-consuming and resource-intensive.

To mitigate these computational challenges, the utilization of the snapshot ensemble technique [31] has gained prominence. This approach leverages the strength of deep neural networks by harnessing multiple locally optimal solutions. Furthermore, authors have bolstered their ensemble by incorporating distinctive feature perspectives and complementary attributes through hyper-parameter randomization. Despite the promising advancements achieved through techniques like snapshot ensembles and transfer learning, the issue of computational resource consumption in multi-framework algorithms remains a point of concern.

In summary, the ensemble approach in AD diagnosis has exhibited promising results, but it is essential to acknowledge the computational demands associated with this strategy. Efforts to mitigate these challenges through techniques like the snapshot ensemble hold potential but require further exploration.

## III. METHODS

### A. Architecture

This section introduces the MC-ViT architecture, comprised of three key components: the 3D Patch network, Monte Carlo Sampling, and ViT, as illustrated in Figure 1. To begin, the entire brain's image patches (as discussed in Section III-A.1) serve as inputs to the 3D Patch Network, generating feature representations. Subsequently, the 3D Patch Network's output is synthesized using the Monte Carlo method, which is detailed in Section III-A.2. This process yields multiple features based on the network's accuracy. These features are then subjected to processing through the ViT, as elaborated in Section III-A.3. The ViT emphasizes spatial relationships to extract global information. Finally, the Classifier and Ensemble combine the results derived from multiple features for each subject. Below is a comprehensive breakdown of the proposed approach's architectural components.

**1) 3D Patch Network:** In our methodology, we utilize a collection of localized image patches as inputs for the singular 3D patch network. Equipped with a moderate parameter size, this singular 3D patch network learner adeptly discerns anatomical features from a range of patches, much like CNNs do for feature extraction in diverse natural images. Concurrently, to optimize computational efficiency, we initially partition MR images into contiguous cubic patches of consistent dimensions, typically represented as  $W \times W \times W$ , with overlapping regions. The primary aim of the 3D patch network is to enhance the distinctive features of discriminative segments within these standardized patches. It has been demonstrated that training a solitary 3D patch network at the volumetric patch level is computationally efficient.

The 3D patch network encompasses two main tasks:

- 1) **learning local spatial patch representations:** It focuses on acquiring spatial representations from individual patches;
- 2) **outputting patch accuracy:** It gauges the accuracy of each patch processed by the 3D patch network.

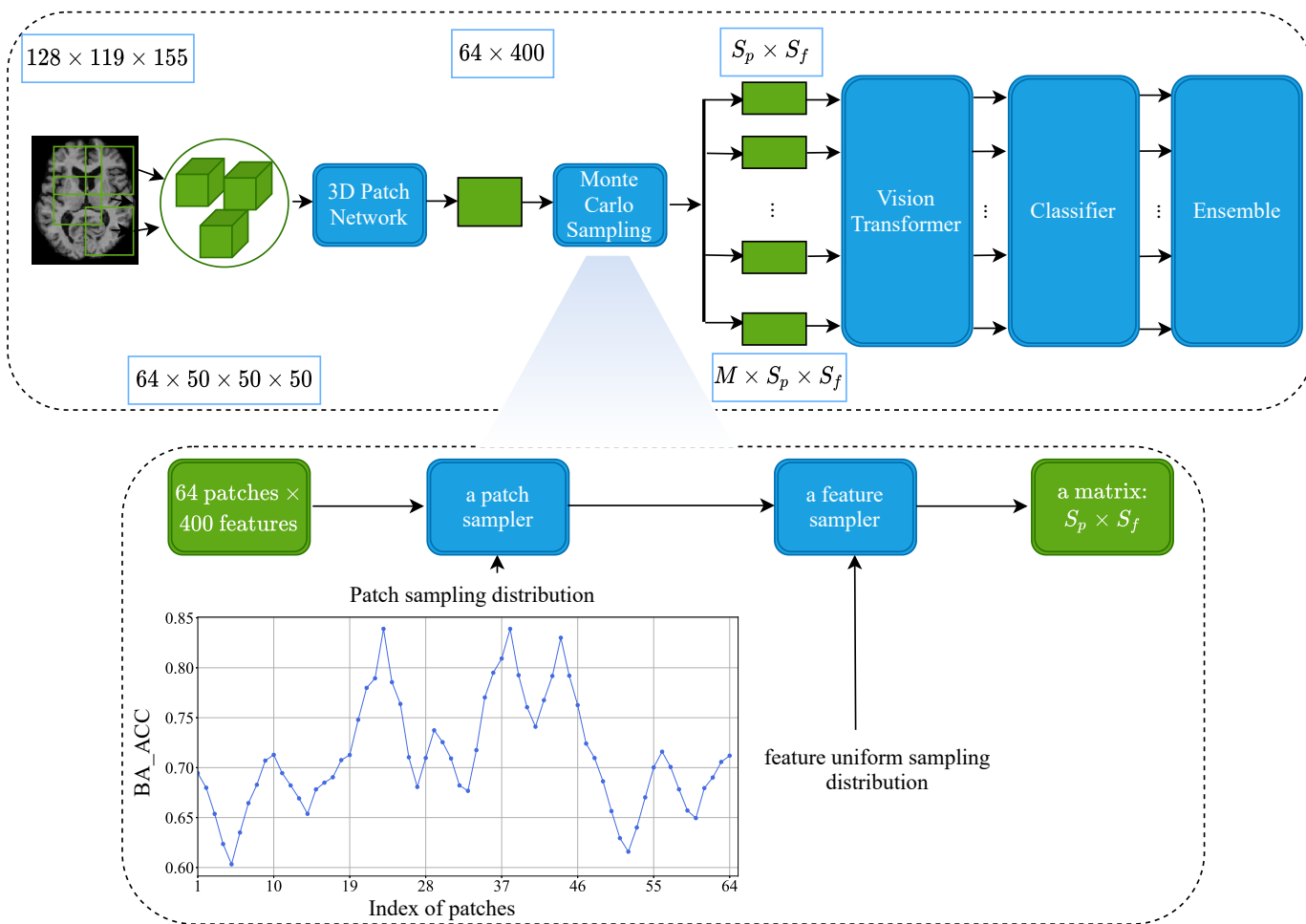


Fig. 1. The architecture of Monte Carlo Vision Transformer Ensemble (MC-ViT). The 3D Patch Network extracts 400 features from each image patch. Monte Carlo Sampling includes a patch sampler and a feature sampler. A patch sampler selects the number of patches ( $S_p$ ) based on the prior distribution. A feature sampler chooses the number of features ( $S_f$ ) in a uniform sampling distribution.

It's worth noting that we opted for a 7-layer 3D Patch Network over a deeper variant due to its superior performance, as demonstrated in previous work [8].

The 3D patch network consists of four convolutional (Conv) blocks, denoted as Conv1 to Conv4, and three fully connected (FC) layers. Each convolutional block comprises a  $3 \times 3 \times 3$  convolutional layer, a batch normalization (BN) layer, a rectified linear unit (ReLU) activation function, and a max-pooling layer. Zero padding is applied when the feature size before the max-pooling layer is not divisible by 2. The number of channels in Conv1 to Conv4 is 15, 25, 50, and 50, respectively. After the Conv blocks, the 3D patch network adds three FC layers to the output. Using the 3D patch network to process patches enables us to effectively derive information representations from a diverse set of patches. As a result, our study employs the 3D patch network for all the MRI patches acquired from each brain MRI scan.

We define a set of patches as  $\mathbf{P} = [\mathbf{P}_0, \mathbf{P}_1, \dots, \mathbf{P}_{63}] \in \mathbb{R}^{64 \times 50 \times 50 \times 50}$ , where  $\mathbf{P}_i \in \mathbb{R}^{50 \times 50 \times 50}$ . Subsequently, we employ the 3D patch network to extract features from each  $\mathbf{P}_i$  patch from  $N$  scans, denoting the resultant feature vector as  $\mathbf{f} = [f_0, f_1, \dots, f_{399}] \in \mathbb{R}^{1 \times 400}$ . Furthermore, we aggregate all the accuracy scores obtained from processing each patch,

resulting in  $p_f = [p_0, p_1, \dots, p_{63}]$ . The feature extraction from each patch occurs just before the first FC layer and is flattened to create a 400-dimensional vector. This process generates 400 features for each of the 64 patches, which are then combined to form a matrix.

**2) Monte Carlo Sampling:** In this section, we introduce the pivotal components of our network. Our network's design seamlessly integrates features derived from patch representations with a distribution of potential decisions. While prior studies [34], [35] have trained individual CNNs for each patch and employed ensemble learning to achieve AD diagnosis results, these approaches often overlook the interrelationships among these patches. To address this limitation, we employ Monte Carlo Sampling to extract additional features from each patch, capturing global information from each scan.

Let  $E = X^n(\text{random}(P'), \text{random}(\mathbf{f}))_{n=1}^N \in \mathbb{R}^{N \times S_p \times S_f}$  represent a collection of image characteristics extracted using the 3D patch network across all  $N$  scans, where  $X$  denotes the scans. Here,  $X^n(\text{random}(P'), \text{random}(\mathbf{f}))$  signifies the random selection of patch and feature from  $n_{th}$  scan. These features are extracted by the 3D Patch Network, where  $\mathbf{P}' = \{\mathbf{P}'_0, \mathbf{P}'_1, \mathbf{P}'_2, \dots, \mathbf{P}'_{63}\}$ , and  $\mathbf{f} = \{f_0, f_1, \dots, f_{399}\}$  forms a subset. Our Monte Carlo sampling process involves the patch

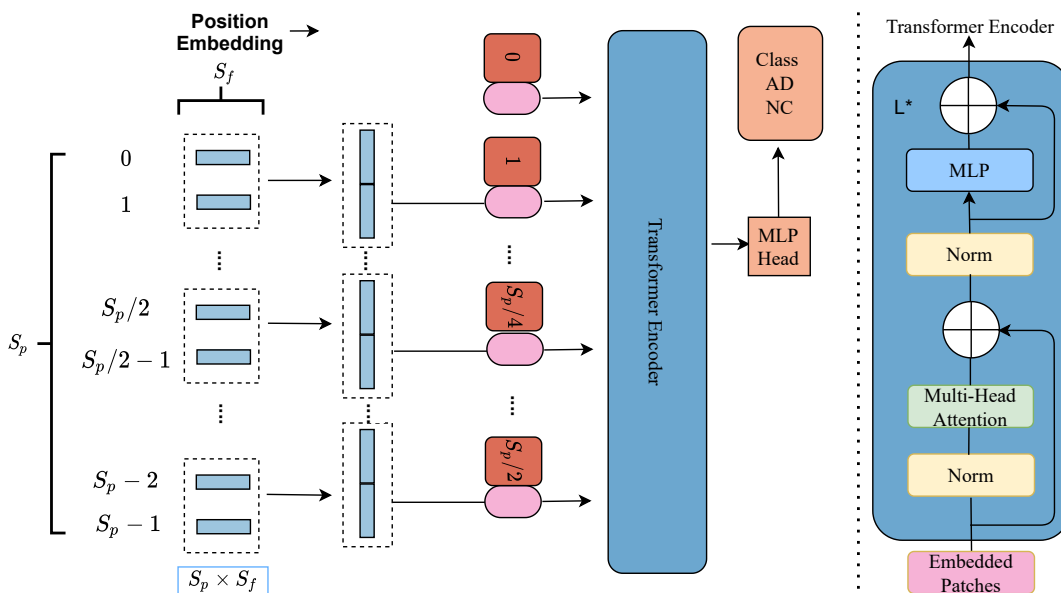


Fig. 2. The architecture of Vision Transformer. In this stage, we stitch the features of the two patches together and remove the first layer containing the CNN to reduce the inductive bias.

sampler, which prioritizes selecting and ranking patch features using  $p_f$  with varying probabilities. Furthermore, evaluating the discriminative process entails applying a 3D patch network to the training data of each patch.

Additionally, we introduce feature sampling to randomly select features, employing a uniform distribution due to the unknown contribution of features from each patch to the outcome. After the Monte Carlo Sampling, the resulting feature, as depicted in Figure 1, comprises  $M \times S_p \times S_f$  components. Here,  $M$  denotes the number employed during Monte Carlo Sampling,  $S_p$  signifies the number of patches sampled, and  $S_f$  denotes the number of features sampled. Additional implementation specifics will be expounded upon in the forthcoming sections. The Monte Carlo sampling selects patches and then repeated  $M$  times for each scan. The patch-level feature representations are then condensed and integrated into the overall global feature representation.

**3) Vision Transformer:** Following the application of Monte Carlo Sampling to aggregate all potential decisions, our study introduces a neural network tailored to capture intricate interactions among features extracted from the sampled patches. We employ a ViT, a pivotal element in our architecture, to learn the relationships among patches and ultimately facilitate AD diagnosis. Furthermore, we harness ensemble learning to consolidate these insights into a final diagnostic decision.

In contrast to traditional approaches that often utilize FC layers or convolutional neural networks (CNNs) to explore correlations among patch-level features, the ViT offers distinct advantages in high-level feature extraction, particularly concerning inter-patch relationships. Leveraging the distinctive weighted feature maps generated by the ViT, our design fosters comprehensive learning of essential feature representations encompassing the entire structural information within brain MRI scans. Consequently, the ViT plays a pivotal role in enhancing the precision of AD classification results.

Our approach centers around a modified architecture rooted in the ViT network architecture. As illustrated in Figure 2, we remove the initial layers containing CNN components from the ViT, thereby reducing the inductive bias inherent to CNNs. We adopt an innovative approach by amalgamating the features of the two patches into a unified embedding. Furthermore, the sequence of embedded patches undergoes a learnable embedding process, complemented by the inclusion of position embeddings to preserve crucial positional data within the patch embeddings.

The ViT network comprises multiple transformer blocks, housing multi-head self-attention mechanisms and feed-forward neural networks. These components empower the model to discern connections between patches and comprehend the overarching relationships embedded within the features. In cases where the resulting vector cannot be evenly divided, we pad it with zeroes to maintain consistency in the processing pipeline.

### B. Evaluation Metrics

To evaluate the classification performance, we employ a set of five standard metrics: Classification Balance Accuracy (BA\_ACC), F1 score, Sensitivity (SEN), Specificity (SPE), and the Area Under the Receiver Operating Characteristic Curve (AUC). Additionally, we introduce the geometric mean [42] as an evaluation metric during the network training process. Specifically, the BA\_ACC is defined as follows:

$$BA\_ACC = \frac{\frac{TP}{TP+FN} + \frac{TN}{FP+TN}}{2} \quad (1)$$

In the context of our study, we define the terms: true positive (TP) as correctly identified samples in the positive class ("NC"), true negative (TN) as correctly identified samples in the negative class ("AD"), false negative (FN) as misclassified

positive samples, and false positive (FP) as misclassified negative samples. Sensitivity and specificity serve as metrics to assess the proportion of accurately classified samples within the positive and negative categories, respectively.

Nevertheless, these metrics, in conjunction with classification accuracy, can be susceptible to the issue of class imbalance, which may result in inaccurate and deceptive evaluations of the classifier's effectiveness when dealing with imbalanced datasets. In order to address this concern and consider the distribution of subjects in both positive and negative classes, we compute the BA\_ACC by adding the accuracies of positive and negative samples and subsequently dividing the total by two.

The Geometric mean provides a holistic evaluation of the classifier's performance, considering the accuracy of both classes rather than solely focusing on one class's accuracy. To address the challenge of imbalanced datasets and account for the subject ratio in both positive and negative classes by integrating the geometric mean into its methodology, as defined below:

$$Geometric\ Mean = \sqrt{SEN \times SPE} \quad (2)$$

The geometric mean is especially useful when subjects are unevenly distributed between positive and negative classes [42]. Furthermore, we employ the AUC to evaluate the overall predictive performance of the model, regardless of the chosen classification threshold. The AUC value ranges from 0 to 1, where a model with 100% incorrect predictions yields an AUC of 0, while a model with 100% correct predictions achieves an AUC of 1.0.

### C. Implementation

The framework, depicted in Figure 1, is implemented using Python 3.7 along with the PyTorch 1.10.0 library. All experimental procedures are conducted on hardware that incorporates an NVIDIA Tesla V100-SXM2 GPU boasting 32GB of RAM, alongside an Intel Xeon Gold 5118 CPU running at 2.30 GHz. The ADNI images were harmonized across different sites to minimize site effects before releasing the data. Hence, mixing the data from all sites is more robust to train the network. The ADNI cohort is divided into two distinct datasets: the first and second ADNI datasets. The first ADNI dataset is employed as the training set for the 3D patch network, while the second dataset is also allocated for both training and evaluation of the MC-ViT model. To safeguard against any inadvertent data contamination, we ensure that all scans from a given subject are exclusively allocated to one of these datasets. The training process is conducted in two phases, commencing with the training of the 3D patch network and followed by MC-ViT training.

1) *3D Patch Network Training*: This study first applies the 3D patch network for solving the two-class classification problem, such as the NC/AD, MCI/AD, and NC/MCI. In this step, the input comprises a set of 3D patches extracted from an image. First, we uniformly divide the MRI images into multiple cubic patches with a fixed size (e.g.,  $50 \times 50 \times 50$ ). Then, the patches' value is normalized between -1 and 1.

The Adam optimizer with recommended parameters was used for training, and the batch size was 256. We trained the networks for 80 epochs, with an initial learning of 0.00005. In the course of network training, we implemented a learning rate scheduler that dynamically fine-tuned the learning rate according to the model's training performance, leading to improved convergence and enhanced generalization of the model. The learning rate was updated using a strategy where, if the loss did not decrease for five consecutive epochs, it would be reduced to 80% of its previous value. The minimum learning rate was set to one-fifth of the initial learning rate. In this stage, the CrossEntropyLoss is used.

Alternatively, the sampling distribution of patches is determined based on the classification accuracies of the patches using a 3D patch network applied to the first ADNI dataset.

2) *MC-ViT Training*: The second ADNI dataset is harnessed for both training and evaluating the performance of the MC-ViT model. Specifically, 37.5% of subjects are allocated to the training set, 12.5% to the validation set, and the remaining 50% to the evaluation set. For each experimental run, it is imperative to first determine the Monte Carlo Sampling parameters, including patch ( $S_p$ ) and feature ( $S_f$ ) specifications, as well as the repetition rate ( $M$ ). The GM metric is employed to strike a balance between the SEN and SPE of the neural network. The MC-ViT model is trained using the Adam optimizer and defined training parameters: a batch size of 32; an initial learning rate set at 0.0005; weight decay is 0.001; and the number of epochs is 50; the learning rate scheduler is the same as 3D Patch Network training. The MC-ViT is trained with LabelSmoothSoftMaxCE, which is a regularization technique used in classification tasks to mitigate overconfidence in model predictions, as shown in Eq. (3). It incorporates label smoothing by assigning a smoothed label distribution instead of binary labels, encouraging more robust and generalized learning. Combining the softmax function with cross-entropy loss strikes a balance between exploring other classes and maintaining confidence in the target class.

$$H(y, \hat{y}) = - \sum_{i=1}^K y_i \times \log(\hat{y}_i) \quad (3)$$

$$y_i = \begin{cases} 0, & \text{otherwise} \\ 1, & i = j \end{cases}$$

As mentioned above, the key hyperparameters, such as the number of patches ( $S_p$ ), features ( $S_f$ ), and Monte Carlo Sampling ( $M$ ), determine the input of the ViT.

## IV. EXPERIMENTS

### A. MRI Data

Two datasets, namely ADNI and OASIS-3, utilized in our study were obtained from the publicly available Alzheimer's Disease Neuroimaging Initiative (ADNI) database (<http://adni.loni.usc.edu>) and the Open Access Series of Imaging Studies (OASIS) database (<http://www.oasis-brains.org>). Table I provides demographic and clinical information for subjects in the ADNI and OASIS-3 cohorts, including age,

TABLE I  
DEMOGRAPHIC AND CLINICAL INFORMATION OF THE ADNI, OASIS-3. ABBREVIATIONS: MMSE, MINI-MENTAL STATE EXAM; CDR, CLINICAL DEMENTIA RATING.

Group Type	Number (Subjects/Scans)	Gender (Male/Female)	Age (Mean $\pm$ SD)	MMSE (Mean $\pm$ SD)	CDR Sum of Box (Mean $\pm$ SD)
ADNI					
NC	546/2190	1095/1095	76.07 $\pm$ 5.80	29.0 $\pm$ 1.2	0.1 $\pm$ 0.3
MCI	910/3393	1837/1273	74.27 $\pm$ 7.61	27.4 $\pm$ 2.3	1.6 $\pm$ 1.1
AD	592/1616	893/694	75.99 $\pm$ 7.41	21.9 $\pm$ 4.3	5.4 $\pm$ 2.6
OASIS-3					
NC	712/1531	923/608	69.0 $\pm$ 9.3	29.0 $\pm$ 1.4	0.1 $\pm$ 0.5
MCI	102/126	58/68	75.0 $\pm$ 8.4	27.9 $\pm$ 2.7	1.1 $\pm$ 1.4
AD	274/335	147/188	76.9 $\pm$ 8.3	24.0 $\pm$ 5.1	4.2 $\pm$ 3.4

gender, mini-mental state examination (MMSE) results, and clinical dementia rating (CDR).

The ADNI project, initiated in 2003 under the leadership of Principal Investigator Michael W. Weiner, MD, is a public-private partnership. The ADNI study encompasses over 1,000 participants, including individuals classified as normal controls (NC), those with MCI, and subjects with AD. The study consists of four subsequent phases: ADNI 1, ADNI "Grand Opportunities" (ADNI GO), ADNI 2, and ADNI 3. For our study, we specifically utilize data from three phases: ADNI 1 (n=811), ADNI GO (n=118), and ADNI 2 (n=1019). We combine data from the same subject across ADNI 1, ADNI GO, and ADNI 2. The frequency of visits per subject ranges from 1 to 12. In accordance with the diagnostic criteria outlined in the ADNI protocol at each visit, subjects were categorized as either NC, individuals with MCI, or those diagnosed with AD. Our study's dataset consists of 7,199 brain scans, including 2,190 scans from 546 NC subjects, 3,393 scans from 910 MCI subjects, and 1,616 scans from 592 AD subjects. It is important to note that we excluded 648 scans from 115 subjects who experienced a transition from MCI to NC during the study.

The OASIS-3 dataset is publicly available and consists of two collections: OASIS-Cross-sectional, which contains sagittal MRI images, and OASIS-Longitudinal, which includes longitudinal slices. The number of visits per subject in the dataset varies between 1 and 7. Both collections feature images acquired with a resolution of  $256 \times 256$  pixels from different patients. This study focused on the OASIS-Cross-sectional collection, which included 436 sagittal images. The images were grouped into three categories: NC, MCI, and AD. The majority of the images (712 subjects or 1531 scans) were classified as NC, while 102 subjects (126 scans) were categorized as MCI, and 247 subjects (335 scans) were classified as AD. It is worth noting that there was a significant imbalance in favor of cognitively normal cases. The dataset included images from 168 male patients and 268 female patients aged 18 to 98 years.

### B. Image Preprocessing

In our study, we followed a four-step process to preprocess structural images. First, we corrected for bias field and used rigid transformations (rotation and translation) to adjust the images. Then, we normalized the image intensities and removed the skull from the images. To correct for intensity

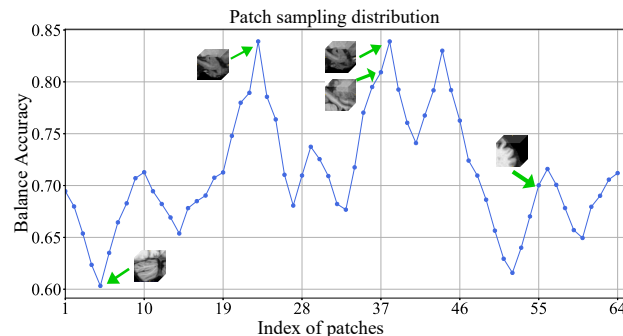


Fig. 3. The NC and AD classification accuracy from 3D Patch Network.

non-uniformity in T1-weighted MRI images, we used the non-parametric non-uniform intensity normalization (N3) method [43]. We registered each image to the MNI space using a linear transformation derived from FLIRT [44]. We also rescaled the image intensities to establish the mean intensity of the white matter at 110. Finally, we removed the skull from the brain images using the watershed algorithm described in a previous study [45].

Since not all brain parts contain valid information, processing the brain, particularly the peripheral background, was necessary. We removed the uninformative peripheral background outside the effective brain area and cropped the images to achieve a uniform size of  $128 \times 119 \times 115$ .

### C. Influence of Patch Sampling ( $S_p$ )

In our work, the first experiment investigates the influence of the patch sampling  $S_p$ . We select all 400 features from the chosen patches and utilize Monte Carlo sampling to ensemble NC/AD classification results. The entire dataset is randomly split into training and testing sets, repeated five times for different numbers of patches (i.e., 8, 16, 32, and 64, totaling 64 patches). For each random sampling with patches and features, different patches and all features are selected for all scans, with this process repeated 50 times, as  $M = 50$ . The definitive classification determination for each MRI scan is established through a majority voting mechanism that factors in the results of these 50 classifications.

The sampling distribution is determined by assessing the AD probability generated by the 3D patch network when a patch discriminates between NC and AD. The patches that

TABLE II  
INFLUENCE OF PATCH SAMPLING ( $S_p$ ) ON THE NC AND AD CLASSIFICATION.

$S_p$	BA_ACC	F1-score	SEN	SPE	AUC
4	0.8746 ±0.0058	0.8984 ±0.0053	0.8696 ±0.0132	0.8667 ±0.0179	0.9386 ±0.0103
8	0.8848 ±0.0072	0.8927 ±0.0087	0.8959 ±0.0418	0.8737 ±0.0422	0.9457 ±0.0071
16	0.8894 ±0.0077	0.8992 ±0.007	0.9081 ±0.0222	0.8707 ±0.0223	0.9479 ±0.0074
32	0.9014 ±0.0043	0.9069 ±0.0026	0.9007 ±0.0085	0.9022 ±0.0154	0.9557 ±0.0065
No Sampling (64)	0.8846 ±0.005	0.8900 ±0.0103	0.8783 ±0.0362	0.8909 ±0.0296	0.9408 ±0.0106

TABLE III  
THE EFFECTS OF THE REPETITION SAMPLING ( $M$ ).

$M$	BA_ACC	F1-score	SEN	SPE	AUC
40	0.8922 ±0.0066	0.9004 ±0.0070	0.8993 ±0.0220	0.8851 ±0.0215	0.9503 ±0.0065
50	0.9014 ±0.0043	0.9069 ±0.0026	0.9007 ±0.0085	0.9022 ±0.0154	0.9557 ±0.0065
60	0.8902 ±0.0077	0.8997 ±0.0068	0.9115 ±0.0139	0.8690 ±0.0231	0.9514 ±0.0069
70	0.8837 ±0.0099	0.8930 ±0.0080	0.8948 ±0.0185	0.8726 ±0.0106	0.9481 ±0.0126
80	0.8935 ±0.0042	0.8969 ±0.0059	0.8765 ±0.0232	0.9105 ±0.0178	0.9537 ±0.0052

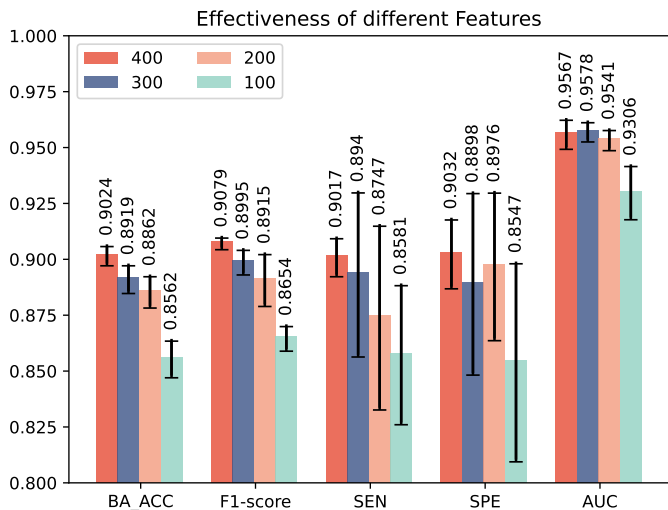


Fig. 4. The effects of the feature sampling ( $S_f$ ) on the NC and AD classification.

cover the hippocampus exhibit the highest probability, as illustrated in Fig. 3.

Table II displays the mean and standard deviation of the NC/AD classification outcomes, which were determined by majority voting across the 50 classification results. In this experiment, we utilized all  $S_f = 400$  features while adjusting  $S_p = 4, 8, 16, 32, 64$ , where the case of 64 patches signifies no patch sampling. Among these configurations, the results obtained with 32 patches without features and  $M = 50$  repetitions demonstrate a higher accuracy than the others. Student's  $t$ -test indicates that the performance using 32 patches is superior to that using 64 patches ( $t = 3.98, p < 0.005$ ) based on BA\_ACC, suggesting that MC sampling aids in improving classification performance. Furthermore, the performance using 32 patches also outperforms that using 4 patches ( $t = 8.28, p < 0.001$ ), 8 patches ( $t = 4.72, p = 0.0015$ ), and is comparable to that using 16 patches ( $t = -1, p = 0.347$ ) based on BA\_ACC. In the following, 32 patches were utilized for the remainder of the experiments detailed in the paper.

#### D. Effects of Feature Sampling ( $S_f$ )

In this section, we delve into the impact of random feature sampling, utilizing a set of 32 patches. As previously discussed in Section IV-C, these patches are selected with

different probabilities derived from the 3D patch network without feature sampling. However, in this section, we adopt a feature sampling approach based on uniform selection. The 3D patch network extracts a latent hierarchical representation comprising 400 features from each patch. Notably, the classification scores acquired for each patch can be regarded as high-level, task-specific features, akin to the auto-context strategy employed in image segmentation. These subsequent classification scores have the potential to offer more direct and semantically rich information pertaining to the diagnostic task, serving as a valuable complement to the patch-level feature representations.

Building upon the observations made in Section IV-C, where it was noted that selecting 32 patches led to higher accuracy, we maintain this patch count and  $M = 50$  in this section. We vary the feature sampling across [100, 200, 300, 400], representing different subsets of the total 400 features. The results, as illustrated in Fig.4, demonstrate a noteworthy trend: accuracy is notably higher when all 400 features are selected compared to other feature subsets.

#### E. Effects of Monte Carlo Sampling Repetition ( $M$ )

Building upon the effects of the number of patches and features discussed above, it is also essential to consider the influence of the repetition sampling ( $M$ ) on the classification between NC and AD. In this experiment, we set  $S_p = 32$  and  $S_f = 400$  and vary  $M = [40, 50, 60, 70, 80]$ . The highest accuracy and AUC value are achieved when  $M = 50$ .

Table.III presents the mean and standard deviation of various evaluation metrics, including BA\_ACC, F1-score, SEN, SPE, and AUC, using the ADNI dataset for the two-class classifiers between NC and AD. The highest accuracy and AUC value are achieved when  $M = 50$ .

#### F. Effects of Patch Size

In our previous results, we consistently used the image patches of  $50 \times 50 \times 50$ . However, in order to explore the influence of the patch size, we conducted a series of experiments where networks were individually trained using local patches of varying dimensions, including  $40 \times 40 \times 40$ ,  $45 \times 45 \times 45$ ,  $50 \times 50 \times 50$ ,  $55 \times 55 \times 55$ , and  $60 \times 60 \times 60$ . The evaluation of classification results was based on metrics such as BA ACC and AUC, and the results are presented in Table.IV.

TABLE IV

COMPARISONS OF OUR PROPOSED WORK WITH DIFFERENT SIZES OF PATCH.

Patch size	BA_ACC	F1-score	SEN	SPE	AUC
$40 \times 40 \times 40$	0.8408	0.8589	0.8487	0.8471	0.9121
	$\pm 0.0088$	$\pm 0.0084$	$\pm 0.0211$	$\pm 0.0317$	$\pm 0.0119$
$45 \times 45 \times 45$	0.8174	0.8307	0.8296	0.8051	0.8801
	$\pm 0.0043$	$\pm 0.0144$	$\pm 0.0168$	$\pm 0.0283$	$\pm 0.0189$
$50 \times 50 \times 50$	0.9014	0.9069	0.9007	0.9022	0.9557
	$\pm 0.0043$	$\pm 0.0026$	$\pm 0.0085$	$\pm 0.0154$	$\pm 0.0065$
$55 \times 55 \times 55$	0.8700	0.8741	0.8472	0.8927	0.9346
	$\pm 0.0121$	$\pm 0.0085$	$\pm 0.0229$	$\pm 0.0460$	$\pm 0.0126$
$60 \times 60 \times 60$	0.8810	0.8873	0.8732	0.8888	0.9382
	$\pm 0.0122$	$\pm 0.0128$	$\pm 0.0267$	$\pm 0.0250$	$\pm 0.0154$

Table IV provides an insightful analysis of our proposed method, revealing its relatively robust performance across a spectrum of input patch sizes, ranging from  $40 \times 40 \times 40$  to  $60 \times 60 \times 60$ . Particularly promising results were observed when utilizing patch sizes within the range of  $50 \times 50 \times 50$  to  $60 \times 60 \times 60$ . However, adopting smaller image patches, such as  $40 \times 40 \times 40$  and  $45 \times 45 \times 45$ , yielded suboptimal outcomes, highlighting a trade-off between performance and patch size.

Conversely, patches sized  $50 \times 50 \times 50$  (BA\_Acc: range: 0.8968 ~ 0.9066) demonstrate statistically superior performance compared to patches sized  $55 \times 55 \times 55$  (BA\_Acc: range: 0.8588 ~ 0.8861; Student's  $t$ -test:  $t = 5.46, p = 0.0006$ ) and patches sized  $60 \times 60 \times 60$  (BA\_Acc range: 0.8700 ~ 0.8995; Student's  $t$ -test:  $t = 3.53, p = 0.008$ ). The performance of  $55 \times 55 \times 55$  and  $60 \times 60 \times 60$  patches is comparable (Student's  $t$ -test:  $t = 1.44, p = 0.188$ ). This experiment underscores that classification performance may not exhibit a linear increase with patch size, suggesting that the performance of our model with Monte Carlo sampling of patches relies on patch size due to its connection with the randomness of feature selection in our model.

### G. Comparisons

This section compares our proposed method with other existing methods based on T1-weighted structural MRI data from the ADNI dataset, as reported in the literature [8]. Specifically, our method is compared to four conventional feature-based methods, including 1) 2D ResNet50 with slices, 2) 3D CNN with patch, 3) 3D CNN with ROI, and 4) 3D CNN with whole brain MRI. We have opted for these established approaches, as they have undergone thorough scrutiny and comparative assessment in a recent review [8]. Furthermore, for a fair comparison, we compared two additional ViT feature-based methods based on our ADNI dataset, including 5) 2D ViT with slices [25], [26], and 6) 3D ViT with whole brain MRI [27], [28].

We used the pre-trained ResNet50 model to fine-tune the last five convolutional layers and the final and additional FC layers for 2D slice CNNs. To mitigate the risk of overfitting, we maintained the weights and biases of the remaining CNN layers in a frozen state during fine-tuning, as recommended [37]. A statistical analysis using a  $t$ -test shows that

TABLE V

OUR PROPOSED WORK IS COMPARED WITH 2D-SLICE AND 3D CNNs FOR CLASSIFYING NC AND AD OF THE ADNI DATASET. THE RESULTS OF 2D CNN AND 3D CNN MODELS ARE FROM TABLE. 6 IN [8].

Model	BA_ACC (Mean $\pm$ SD)	Accuracy of 5 repetitions (Of 5 repetitions)
2D-slice CNNs		
ResNet50	$0.79 \pm 0.04$	0.83,0.83,0.72,0.82,0.83
3D CNNs		
3D-patch single-CNN	$0.74 \pm 0.08$	0.75,0.84,0.78,0.75,0.59
3D-patch multi-CNN	$0.81 \pm 0.03$	0.82,0.84,0.83,0.77,0.79
3D-ROI CNN	$0.88 \pm 0.03$	0.84,0.89,0.90,0.89,0.85
3D whole-brain CNN	$0.82 \pm 0.05$	0.74,0.90,0.83,0.77,0.83
Vision Transformer		
2D-slice single-ViT	$0.74 \pm 0.02$	0.73,0.71,0.74,0.76,0.76
3D whole-brain ViT	$0.76 \pm 0.02$	0.74,0.79,0.77,0.74,0.74
Our proposed	$0.90 \pm 0.00$	0.90,0.90,0.91,0.90,0.90

our proposed model outperforms the best classification results achieved with trained ResNet50 (Student's  $t$ -test:  $t = 4.43, p < 0.002$ ).

To ensure equitable comparisons, we leveraged the outcomes from Wen's 2020 study for the evaluation of 3D CNNs [8]. Our 3D patch CNN architecture comprises four convolutional blocks and three fully connected (FC) layers. The training of this model employed three distinct approaches. Initially, in the first approach, we accommodated 36 patches, each measuring  $50 \times 50 \times 50$  mm<sup>3</sup>, within a single 3D patch CNN, denoted as a 3D patch single-CNN. The second approach employed the use of separate 3D patch convolutional neural networks (CNNs) for each patch. This yielded a total of 36 distinct CNN models, collectively referred to as the 3D patch multi-CNN. As for the 3D ROI CNN model, it was tailored to focus exclusively on the region of interest (ROI) encompassing the Hippocampus, which had dimensions of  $50 \times 50 \times 50$  mm<sup>3</sup>. This ROI served as input for the 3D patch CNN. Lastly, we devised the 3D whole-brain CNN, comprising five convolutional blocks and three FC layers, designed to process the entire brain volume as input. In accordance with standard scientific conventions and requisites, this methodology was pursued. An analysis utilizing Student's  $t$ -tests demonstrated that the proposed model outperforms the 3D patch single-CNN (Student's  $t$ -test:  $t = 3.86, p < 0.005$ ), the 3D patch multi-CNN (Student's  $t$ -test:  $t = 6.97, p < 0.001$ ), the 3D ROI CNN (Student's  $t$ -test:  $t = 2.29, p < 0.052$ ), and the 3D whole-brain CNN (Student's  $t$ -test:  $t = 3.55, p < 0.008$ ).

We trained the 2D ViT-B/16 [25], [26] and 3D ViT [27], [28] models from scratch to assess the ADNI dataset, utilizing both 2D slices and whole-brain MRI data, for comparative analysis with our methodologies. For the implementation of 2D ViT and 3D ViT [27], [28], featuring pure-transformer networks, the transformer sequence is utilized to extract 2D patch embeddings and 3D patches, respectively. Analysis conducted using Student's  $t$ -tests demonstrated that our proposed model outperforms both the 2D ViT-B/16 (Student's  $t$ -test:  $t = 16.71, p < 0.001$ ) and the 3D ViT (Student's  $t$ -test:  $t = 13.92, p < 0.001$ ).

TABLE VI  
ROBUSTNESS OF PROPOSED WORK IN THE ADNI AND OASIS-3 DATASETS.

	BA_ACC	F1-score	SEN	SPE	AUC
ADNI					
NC/AD	0.9014 ±0.0043	0.9069 ±0.0026	0.9007 ±0.0085	0.9022 ±0.0154	0.9557 ±0.0065
NC/MCI	0.6426 ±0.0122	0.6648 ±0.0452	0.6447 ±0.1026	0.6405 ±0.1020	0.6837 ±0.0164
MCI/AD	0.7202 ±0.0095	0.6492 ±0.0187	0.7332 ±0.0551	0.7073 ±0.0404	0.7953 ±0.0083
OASIS-3					
NC/AD	0.8028 ±0.0073	0.5807 ±0.0253	0.7780 ±0.0384	0.8276 ±0.0402	0.8632 ±0.0199
NC/MCI	0.6147 ±0.0013	0.2704 ±0.0110	0.3990 ±0.0879	0.8305 ±0.0673	0.6740 ±0.0409
MCI/AD	0.6448 ±0.0255	0.6837 ±0.0587	0.6271 ±0.1044	0.6626 ±0.0859	0.6643 ±0.0241

### H. Robustness analysis

We are testing the suggested model’s ability to remain robust across various datasets, which include ADNI and OASIS-3. To assess its performance, we first train 3D Patch Network on the first ADNI dataset, which consists of 2,583 subjects. The scans have three classifications: NC, MCI, and AD. We use the second ADNI dataset, which has 3,968 scans, to train and evaluate the Monte Carlo Sampling and ViT. We focus on two-class classification problems, specifically comparing NC vs. AD, NC vs. MCI and MCI vs. AD. We set the model hyperparameters without feature sampling to  $M = 50$ ,  $S_p = 32$ . To ensure the reliability of our results, we randomly split the second ADNI dataset into two halves for training and evaluation purposes. We repeat this process five times to account for any potential variations. Table VI presents the mean and standard deviation of various performance metrics, such as BA\_ACC, SEN, SPE, and AUC, for the two-class classifiers in the ADNI dataset.

Subsequently, we applied transfer learning to finetune the MC-ViT model using half of the OASIS-3 dataset and evaluated its performance on the remaining half. This experiment was repeated five times. Table VI presents the BA\_ACC, SEN, SPE, and AUC for NC/AD, NC/MCI, and MCI/AD on the OASIS-3 dataset. It is noteworthy that the classification accuracy rates between NC and AD on the OASIS-3 dataset are comparatively lower than those on the ADNI dataset. This difference can be attributed to the relatively smaller sample sizes of MCI and AD patients in the OASIS-3 dataset, as well as the relatively better performance on the MMSE and CDR sum of box scores by AD patients in the OASIS-3 dataset when compared to the ADNI dataset (refer to Table I). The classification accuracy of OASIS-3 is comparatively lower than ADNI’s due to its smaller dataset size and the presence of milder AD cases. Therefore, by employing our proposed model on the OASIS-3 dataset, we observe differences in the classification performance between the two datasets. These may be due to the sample size, demographic characteristics, and clinical scores of individuals in the respective datasets.

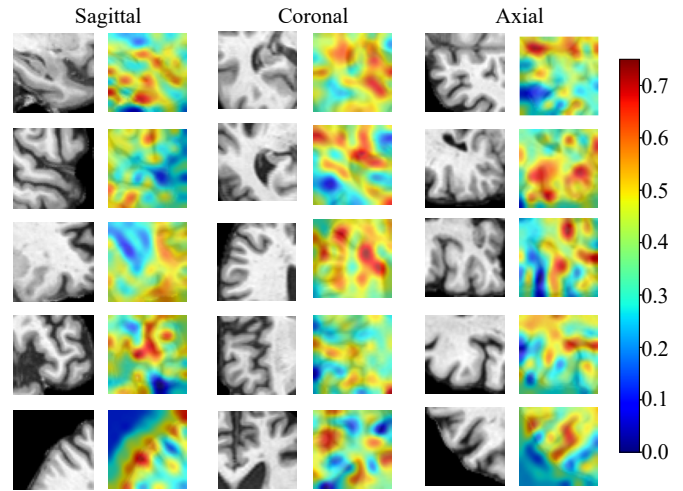


Fig. 5. Saliency maps for AD diagnosis. The saliency maps are displayed in three MR planes (i.e., sagittal, coronal, and axial), where red and blue colors denote high and low discriminative features in sMRI, respectively.

### I. Saliency Map

We employed the 3D Grad-CAM method as a heatmap visualization tool to illustrate the importance of each patch for interpretation. This technique effectively highlights the critical brain regions that contributed to the AD diagnosis. Fig.5 also demonstrates that the Hippocampus is an important region. Due to the overlap of the relative locations in the whole image, the patches with the most discriminative locations. Fig.5 shows that the most discriminative locations with red colors are located at the edges of the sulcus gyrus and gray matter, which may effectively reflect the local structural changes by brain atrophy. This is consistent with the brain atrophy patterns highlighted in literature [3], [7], [46], [47].

## V. DISCUSSION

This study introduces an innovative MC-ViT approach utilizing 3D image patches for the computer-aided diagnosis of AD. We highlight the pivotal role of patch and feature sampling in improving classification performance by enabling the generation of informed decisions. Moreover, integrating the ViT within our model effectively captures global structural features across diverse patches. Furthermore, our results indicate that employing 32 patches with all features in MC-ViT (24.04 million parameters) reduces the model size by approximately 74% compared to using 64 patches with all features (93.21 million parameters). Additionally, our experiments demonstrate that setting MC = 50 and utilizing 32 patches with 400 features per patch resulted in a runtime of 3.52 hours, which was 20.5% lower than the runtime of employing 64 patches with 400 features per patch across 50 ViT models, taking 4.43 hours. Our methodology only requires minimal preprocessing of structural brain images, which includes rigid transformation, intensity normalization, and skull removal. Our experimental findings demonstrate that the proposed model can achieve accurate AD diagnosis even with a limited number of brain patches.

To assess the robustness of our model, we conducted experiments on the OASIS-3 dataset, training the model on the ADNI dataset. The results reveal that transfer learning empowers the model to perform exceptionally well on the OASIS-3 data for AD classification. Notably, the NC vs. AD classification accuracy achieved by our proposed model on the OASIS-3 dataset surpasses that of other models, such as the 3D-ROI CNN, 3D-patch multi-CNN model, and 3D whole-brain CNN (with accuracies ranging from 64% to 67%, as shown in Table.6 in [8]). This observation reinforces the notion presented in [8] that classification performance is intricately linked to clinical characteristics. When clinical diagnoses differ across datasets due to variations in practitioner experiences and diagnostic tools, transfer learning emerges as a pragmatic approach for bridging diagnostic gaps between the two datasets.

However, it's crucial to acknowledge the limitations of our study. Notably, There are comparatively few numbers of MCI and AD patients in the OASIS-3 dataset warrants further investigations with larger sample sizes to establish the robustness of the MC-ViT model. Additionally, the integration of prior knowledge concerning AD-related brain regions has the potential to enhance the model's performance. Nevertheless, this enhancement comes at the cost of more intensive image processing, such as segmentation and image registration, which may be less feasible in clinical settings. Therefore, it is imperative for future research endeavors to seek a harmonious blend between the utilization of prior knowledge and the attainment of swift computational capabilities tailored to practical clinical use.

#### ACKNOWLEDGEMENTS

This research/project is supported by STI 2030—Major Projects (No.2022ZD0209000), the National Research Foundation, Singapore, and the Agency for Science Technology and Research (A\*STAR), Singapore, under its Prenatal/Early Childhood Grant (Grant No. H22P0M0007), and by the Hong Kong global STEM scholar scheme.

#### REFERENCES

[1] Mahsa Dadar, Tharick A Pascoal, Sarinporn Manitsirikul, Karen Misquitta, Vladimir S Fonov, M Carmela Tartaglia, John Breitner, Pedro Rosa-Neto, Owen T Carmichael, Charles Decarli, et al. Validation of a regression technique for segmentation of white matter hyperintensities in alzheimer's disease. *IEEE transactions on medical imaging*, 36(8):1758–1768, 2017.

[2] Colin J Holmes, Rick Hoge, Louis Collins, Roger Woods, Arthur W Toga, and Alan C Evans. Enhancement of mr images using registration for signal averaging. *Journal of computer assisted tomography*, 22(2):324–333, 1998.

[3] W. Y. Zhu, L. Sun, J. S. Huang, L. X. Han, and D. Q. Zhang. Dual attention multi-instance deep learning for alzheimer's disease diagnosis with structural mri. *Ieee Transactions on Medical Imaging*, 40(9):2354–2366, 2021.

[4] Mengjin Dong, Long Xie, Sandhitsu R. Das, Jiancong Wang, Laura E.M. Wisse, Robin deFlores, David A. Wolk, and Paul A. Yushkevich. Deepatrophy: Teaching a neural network to detect progressive changes in longitudinal mri of the hippocampal region in alzheimer's disease. *NeuroImage*, 243:118514, 2021.

[5] Yanteng Zhang, Qizhi Teng, Yuyang Liu, Yan Liu, and Xiaohai He. Diagnosis of alzheimer's disease based on regional attention with smri gray matter slices. *Journal of Neuroscience Methods*, 365:109376, 2022.

[6] Ruoxuan Cui and Manhua Liu. Hippocampus analysis by combination of 3-d densenet and shapes for alzheimer's disease diagnosis. *IEEE journal of biomedical and health informatics*, 23(5):2099–2107, 2018.

[7] Chunfeng Lian, Mingxia Liu, Jun Zhang, and Dinggang Shen. Hierarchical fully convolutional network for joint atrophy localization and alzheimer's disease diagnosis using structural mri. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):880–893, 2018.

[8] J. Wen, E. Thibeau-Sutre, M. Diaz-Melo, J. Samper-Gonzalez, A. Routier, S. Bottani, D. Dormont, S. Durrleman, N. Burgos, O. Colliot, Initiative Alzheimer's Disease Neuroimaging, Biomarkers Australian Imaging, and ageing Lifestyle flagship study of. Convolutional neural networks for classification of alzheimer's disease: Overview and reproducible evaluation. *Med Image Anal*, 63:101694, 2020.

[9] Saman Sarraf and Ghassem Tofghi. Classification of alzheimer's disease structural mri data by deep learning convolutional neural networks. *arXiv preprint arXiv:1607.06583*, 2016.

[10] Muhammad Mujahid, Amjad Rehman, Teg Alam, Faten S Alamri, Suliman Mohamed Fati, and Tanzila Saba. An efficient ensemble approach for alzheimer's disease detection using an adaptive synthetic technique and deep learning. *Diagnostics*, 13(15):2489, 2023.

[11] Huan Lao and Xuejun Zhang. Regression and classification of alzheimer's disease diagnosis using nmf-tdnet features from 3d brain mr image. *IEEE Journal of Biomedical and Health Informatics*, 26(3):1103–1115, 2021.

[12] Hossein Shahamat and Mohammad Saniee Abadeh. Brain mri analysis using a deep learning based evolutionary approach. *Neural Networks*, 126:218–234, 2020.

[13] Xin Zhang, Liangxiu Han, Wenyong Zhu, Liang Sun, and Daoqiang Zhang. An explainable 3d residual self-attention deep neural network for joint atrophy localization and alzheimer's disease diagnosis using structural mri. *IEEE journal of biomedical and health informatics*, 26(11):5289–5297, 2021.

[14] Xiaoqi Shen, Lan Lin, Xinze Xu, and Shuicai Wu. Effects of patchwise sampling strategy to three-dimensional convolutional neural network-based alzheimer's disease classification. *Brain Sciences*, 13(2):254, 2023.

[15] Manhua Liu, Fan Li, Hao Yan, Kundong Wang, Yixin Ma, Li Shen, Mingqing Xu, Alzheimer's Disease Neuroimaging Initiative, et al. A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in alzheimer's disease. *Neuroimage*, 208:116459, 2020.

[16] M. Tanveer, A. H. Rashid, M. A. Ganaie, M. Reza, I. Razzak, and K. L. Hua. Classification of alzheimer's disease using ensemble of deep neural networks trained through transfer learning. *Ieee Journal of Biomedical and Health Informatics*, 26(4):1453–1463, 2022.

[17] Zhentao Hu, Zheng Wang, Yong Jin, and Wei Hou. Vgg-tswinformer: Transformer-based deep learning model for early alzheimer's disease prediction. *Computer Methods and Programs in Biomedicine*, 229:107291, 2023.

[18] Xingyu Gao, Hongjie Cai, and Manhua Liu. A hybrid multi-scale attention convolution and aging transformer network for alzheimer's disease diagnosis. *IEEE Journal of Biomedical and Health Informatics*, 2023.

[19] Zhentao Hu, Yanyang Li, Zheng Wang, Shuo Zhang, Wei Hou, Alzheimer's Disease Neuroimaging Initiative, et al. Conv-swformer: Integration of cnn and shift window attention for alzheimer's disease classification. *Computers in Biology and Medicine*, 164:107304, 2023.

[20] Yang Liu, Yao Zhang, Yixin Wang, Feng Hou, Jin Yuan, Jiang Tian, Yang Zhang, Zhongchao Shi, Jianping Fan, and Zhiqiang He. A survey of visual transformers. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[21] Jyoti Islam and Yanqing Zhang. Brain mri analysis for alzheimer's disease diagnosis using an ensemble system of deep convolutional neural networks. *Brain informatics*, 5:1–14, 2018.

[22] Mingxia Liu, Jun Zhang, Ehsan Adeli, and Dinggang Shen. Landmark-based deep multi-instance learning for brain disease diagnosis. *Medical image analysis*, 43:157–168, 2018.

[23] Weiming Lin, Tong Tong, Qinquan Gao, Di Guo, Xiaofeng Du, Yonggui Yang, Gang Guo, Min Xiao, Min Du, Xiaobo Qu, et al. Convolutional neural networks-based mri image analysis for the alzheimer's disease prediction from mild cognitive impairment. *Frontiers in neuroscience*, 12:777, 2018.

[24] H. F. Wang, Y. Y. Shen, S. Q. Wang, T. F. Xiao, L. M. Deng, X. Y. Wang, and X. Y. Zhao. Ensemble of 3d densely connected convolutional

- network for diagnosis of mild cognitive impairment and alzheimer's disease. *Neurocomputing*, 333:145–156, 2019.
- [25] Wei Chen, Fan Sun, Yi Luo, and Xiaomin Wang. Ensemble model for predicting alzheimer's disease and disease stages with cnn and transformer models. In *Proceedings of the 2023 2nd International Conference on Algorithms, Data Mining, and Information Technology*, pages 46–51, 2023.
- [26] Gia Minh Hoang, Ue-Hwan Kim, and Jae Gwan Kim. Vision transformers for the prediction of mild cognitive impairment to alzheimer's disease progression using mid-sagittal smri. *Frontiers in Aging Neuroscience*, 15:1102869, 2023.
- [27] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [28] Jinseong Jang and Dosik Hwang. M3t: three-dimensional medical image classifier using multi-plane and multi-slice transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20718–20729, 2022.
- [29] Syrine Neffati, Khaoula Ben Abdellafou, Ines Jaffel, Okba Taouali, and Kais Bouzrara. An improved machine learning technique based on downsized kpca for alzheimer's disease classification. *International Journal of Imaging Systems and Technology*, 29(2):121–131, 2019.
- [30] Rachna Jain, Nikita Jain, Akshay Aggarwal, and D Jude Hemanth. Convolutional neural network based alzheimer's disease classification from magnetic resonance brain images. *Cognitive Systems Research*, 57:147–159, 2019.
- [31] Muhammad Tanveer, Ashraf Haroon Rashid, MA Ganaie, Motahar Reza, Imran Razzak, and Kai-Lung Hua. Classification of alzheimer's disease using ensemble of deep neural networks trained through transfer learning. *IEEE Journal of Biomedical and Health Informatics*, 26(4):1453–1463, 2021.
- [32] Amir Ebrahimi-Ghahnavieh, Suhuai Luo, and Raymond Chiong. Transfer learning for alzheimer's disease detection on mri images. pages 133–138, 2019.
- [33] Mohit Dua, Drishti Makhija, PYL Manasa, and Prashant Mishra. A cnn-rnn-lstm based amalgamation for alzheimer's disease detection. *Journal of Medical and Biological Engineering*, 40:688–706, 2020.
- [34] Jin Liu, Jianxin Wang, Bin Hu, Fang-Xiang Wu, and Yi Pan. Alzheimer's disease classification based on individual hierarchical networks constructed with 3-d texture features. *IEEE transactions on nanobioscience*, 16(6):428–437, 2017.
- [35] Jin Liu, Min Li, Wei Lan, Fang-Xiang Wu, Yi Pan, and Jianxin Wang. Classification of alzheimer's disease using whole brain hierarchical network. *IEEE/ACM transactions on computational biology and bioinformatics*, 15(2):624–632, 2016.
- [36] X. F. Zhu, H. I. Suk, L. Wang, S. W. Lee, D. Shen, and Alzheimers Dis Neuroimaging Initia. A novel relational regularization feature selection method for joint regression and classification in ad diagnosis. *Medical Image Analysis*, 38:205–214, 2017.
- [37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. 2017.
- [38] Sheng He, P. Ellen Grant, and Yangming Ou. Global-local transformer for brain age estimation. *IEEE Transactions on Medical Imaging*, 41(1):213–224, 2022.
- [39] Fatih Altay, Guillermo Ramón Sánchez, Yanli James, Stephen V Faraone, Senem Velipasalar, and Asif Salekin. Preclinical stage alzheimer's disease detection using magnetic resonance image scans. 35(17):15088–15097, 2021.
- [40] Ye Ren, Le Zhang, and Ponnuthurai N Suganthan. Ensemble classification and regression-recent developments, applications and future directions. *IEEE Computational intelligence magazine*, 11(1):41–53, 2016.
- [41] Mingxia Liu, Jun Zhang, Dong Nie, Pew-Thian Yap, and Dinggang Shen. Anatomical landmark based deep feature representation for mr images in brain disease diagnosis. *IEEE Journal of Biomedical and Health Informatics*, 22(5):1476–1485, 2018.
- [42] Victoria López, Alberto Fernández, Salvador García, Vasile Palade, and Francisco Herrera. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information sciences*, 250:113–141, 2013.
- [43] John G Sled, Alex P Zijdenbos, and Alan C Evans. A nonparametric method for automatic correction of intensity nonuniformity in mri data. *IEEE transactions on medical imaging*, 17(1):87–97, 1998.
- [44] Mark Jenkinson and Stephen Smith. A global optimisation method for robust affine registration of brain images. *Medical image analysis*, 5(2):143–156, 2001.
- [45] Florent Ségonne, Anders M Dale, Evelina Busa, Maureen Glessner, David Salat, Horst Karl Hahn, and Bruce Fischl. A hybrid approach to the skull stripping problem in mri. *Neuroimage*, 22(3):1060–1075, 2004.
- [46] Daoqiang Zhang, Dinggang Shen, Alzheimer's Disease Neuroimaging Initiative, et al. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in alzheimer's disease. *NeuroImage*, 59(2):895–907, 2012.
- [47] Yuanchen Wu, Yuan Zhou, Weiming Zeng, Qian Qian, and Miao Song. An attention-based 3d cnn with multi-scale integration block for alzheimer's disease classification. *IEEE Journal of Biomedical and Health Informatics*, 26(11):5665–5673, 2022.