IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 10 March 2024, accepted 22 March 2024, date of publication 27 March 2024, date of current version 2 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3381862

RESEARCH ARTICLE

Multimodal 3D Deep Learning for Early Diagnosis of Alzheimer's Disease

SEUNG KYU KIM[1], QUAN ANH DUONG[1], AND JIN KYU GAHM[2,3], for the Alzheimer's Disease Neuroimaging Initiative

[1]Department of Information Convergence Engineering, Pusan National University, Busan 46241, South Korea
[2]School of Computer Science and Engineering, Pusan National University, Busan 46241, South Korea
[3]Center for Artificial Intelligence Research, Pusan National University, Busan 46241, South Korea

Corresponding author: Jin Kyu Gahm (gahmj@pusan.ac.kr)

This work was supported in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development grant funded by Korean Government [Ministry of Science and ICT (MSIT)] under Grant IITP-2023-RS-2023-00254177 and in part by the National Research Foundation of Korea (NRF) grant funded by Korean Government (MSIT) under Grant NRF-2020R1C1C1008362 and Grant 2022R1A4A1030189.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Alzheimer's Disease Neuroimaging Initiative (ADNI) Database.

ABSTRACT Alzheimer's disease (AD) is a neurodegenerative disease that affects the elderly and leads to cognitive decline and memory loss. Treatments for stopping or slowing the progression of AD have not been discovered yet; therefore, delaying the progression of AD is the only option, which makes early diagnosis of AD crucial. Additionally, although Aβ plaques and tau proteins are considered the causes of early AD, few studies have used this information to diagnose early AD. In this study, a middle-fusion multimodal model is proposed for the diagnosis of early AD. The proposed multimodal model extracts features without loss using a depthwise separable convolution block without an activation function. Subsequently, middle fusion is applied using mix skip connection and sharing weight convolution blocks, both designed to learn the complex relationships between modalities. In contrast to other studies, the proposed approach has three main novelties. 1) A middle-fusion multimodal model is proposed for the early diagnosis of AD. 2) The proposed model is evaluated using the entire ADNI series, including T1-weighted magnetic resonance imaging (T1w MRI) and 18F-FluoroDeoxyGlucose positron emission tomography (FDG PET) from the ADNI1 dataset, as well as Aβ PET and tau protein PET from ADNI2 and ADNI3 datasets. 3) A novel region-of-interest (ROI) extraction method is proposed for the hippocampus, middle temporal, and inferior temporal regions, which are known to be affected in the early stages of AD. In the experimental results, the proposed multimodal model achieved a balanced accuracy of 1.00, for the task of Alzheimer's disease vs cognitive normal (CN) and 0.76 for the task of mild cognitive impairment vs cognitive normal.

INDEX TERMS Computer aided diagnosis, convolutional neural networks, deep learning, dementia, image classification, magnetic resonance imaging, positron emission tomography.

I. INTRODUCTION

Alzheimer's disease (AD) is a degenerative brain disease that causes loss of nerve cells and tissue in the brain, leading to cognitive impairment and memory loss, especially in the elderly [1], [2]. By 2050, one in 85 people worldwide will suffer from AD or other types of dementia [3]. Although nursing and treatment costs are expected to increase significantly as the number of patients increases, no other treatment has been found to stop or treat disease progression except for treatments that slow disease progression [4]. Since AD is a degenerative brain disease, loss of nerve cells or tissue occurs when the disease progresses. Therefore, the diagnosis of mild cognitive impairment (MCI), which is considered a precursor to AD, is an important part of early AD diagnosis.

The associate editor coordinating the review of this manuscript and approving it for publication was Byung-Gyu Kim.

Computer-aided diagnostics based on deep-learning approaches have been widely studied in the medical imaging field. Research on the early diagnosis of AD has been conducted by many research groups focusing on diagnostic performance improvement and self-supervised learning methods [5], [9] using deep learning, machine learning, and other algorithms. Most research has been conducted using magnetic resonance imaging (MRI), which provides structural information on brain tissue, and 18F-FluoroDeoxyGlucose positron emission tomography (FDG-PET), which provides metabolic information on the brain. Structural changes in the brain MRI indicate that nervous tissue loss has already occurred; therefore, it is too late for an early diagnosis of AD, and the metabolic information on the brain does not provide information on the cause of AD. Current research on the causes of Alzheimer's disease attempts to explain the causes of AD through amyloid β plaques (Aβ plaques) and tau protein hypotheses. The Aβ hypothesis states that the precipitation of Aβ peptides causes AD, whereas the tau protein hypothesis states that hyperphosphorylation of tau protein causes neurofibrillary tangles which cause AD [10], [11]. An A/T/N biomarker classification scheme was proposed for the clinical diagnosis of AD based on the tau protein hypothesis [14], where A refers to the Aβ biomarker; T is a tau protein biomarker; and N indicates neurodegeneration or neuronal injury, with the category of the biomarker classified as positive or negative. Similar to the A/T/N biomarker classification scheme for the clinical diagnosis of AD, developing a multimodal deep learning model using MRI, Aβ PET and Tau PET is essential.

In this study, multimodal models based on 3D subjects and ROIs are proposed and applied to MRI and PET images collected from the Alzheimer's Disease Neuroimaging Initiative (ADNI) [15]. In ADNI1, subjects who were scanned using both MRI and FDG-PET were included, whereas in ADNI2 and ADNI3, subjects scanned by MRI, Tau PET, and Aβ PET were included. The performance of the proposed model was validated through the following three tasks: AD diagnosis task of Alzheimer's disease (AD) vs. cognitively normal (CN); early AD diagnosis task of mild cognitive impairment (MCI) vs. CN; and AD predictive diagnosis task of stable MCI (sMCI) vs. progressive MCI (pMCI)). The proposed multimodal model extracts features while preventing losses by applying a depthwise separable convolution (DS-Conv) block [16] without an activation function. Subsequently, each extracted modality preserves features through a DS-Conv block, and middle fusion is applied through a mix skip connection convolution (MSC-Conv) block to learn the complex relationship between modalities. Subsequently, a multimodal model is learned during training to extract common features between the modalities related to labels while sharing the weights through a sharing weight convolution (SW-Conv) block. In addition, a new region of interest (ROI) extraction method optimized for the tau protein and Aβ plaque is proposed. Research related to early AD biomarkers suggests that the middle temporal and inferior temporal regions show changes in Aβ plaque and tau protein [17] in the early stages of AD. The hippocampus was selected as the ROI in another study [18]. Therefore, our new ROI extraction method focused on the hippocampus, middle temporal, and inferior temporal regions. The contributions of this study are as follows:

- A new multimodal model is proposed for early diagnosis of Alzheimer's disease.
- Unlike most other studies, the proposed research uses the entire ADNI dataset consisting of ADNI1, ADNI2, and ADNI3. In particular, ADNI2 and ADNI3 cover MRI, Tau PET, and Aβ PET.
- A new ROI extraction method is proposed for identifying precipitating Aβ plaque and tau protein in the early stages of AD.

The remainder of this paper is organized as follows. Section II presents related work classified by model input type. In Section III, the preprocessing method and the proposed model are explained in detail. Section IV presents the experimental setting and evaluation methods. In Section V, the results of the proposed model are presented.

II. RELATED WORK

In this study, deep-learning research on the early diagnosis of Alzheimer's disease is classified into three categories: 3D subject-based methods using the entire image of the subject, ROI/Patch-based methods utilizing specific areas or patches of the image, and 2D slice-based methods employing 2D slices of the subject's 3D image.

A. 3D SUBJECT-BASED METHODS

This classification involves the use of the entire 3D image volume of the subject in early diagnostic models of AD. However, these methods require substantial computational resources. Zhang et al. [19] introduced self-attention into residual connection blocks and proposed a 3D residual self-attention deep learning network for MRI images. Visualization of important areas in the classification results was achieved using gradient-weighted class activation mapping (GRAD-CAM) [20]. Yee et al. [21] presented a convolution model with residual connections and performed classification using a 1×1 convolution layer with global average pooling and softmax. Punjabi et al. [22] used a simple 3D CNN model to combine MRI and PET images. They extracted the features of each modality using convolutional neural networks (CNN) and performed late fusion through feature concatenation. Zou et al. [23] applied Tau PET to the Inception-V3 3D model and conducted a CN vs. AD/MCI task. Spasov et al. [24] proposed an early fusion method for combining multimodal images and a late fusion method for combining clinical information with image features.

B. ROI/PATCH-BASED METHODS

This category utilizes 3D patches and a region of interest (ROI) for learning, instead of the entire image of the subject.

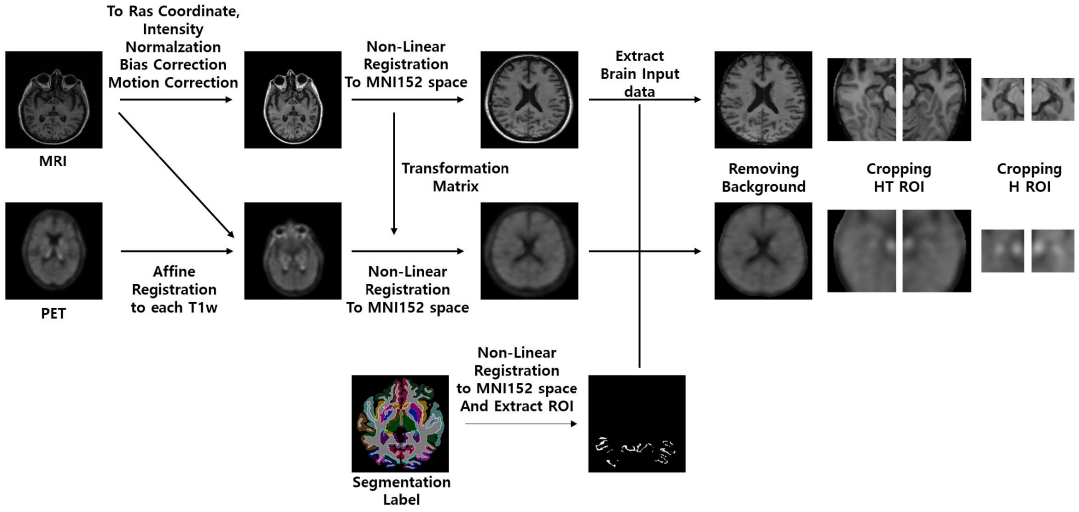

FIGURE 1. Preprocessing flowchart for MRI and PET images. After applying different preprocessing steps to each modality, the same cropping methods were applied to extract the ROI from the entire image of each subject.

Models using only specific areas, fall under the ROI-based method, and those using entire patches constitute the patch-based method. ROI/patch-based methods typically require fewer computing resources than 3D subject-based methods, as image size is smaller. Wen et al. [25] proposed both patch- and ROI-based methods. The patch-based method divides the entire image into non-overlapping patches and performs classification through pretraining using an autoencoder and fine-tuning. The ROI-based method extracts two patches, each containing the left and right hippocampal ROI, thereby learning them in a manner similar to the patch-based method. Huang et al. [26] employed a VGG-like framework to learn MRI and FDG-PET images, using regions containing the hippocampus as the ROI. Zhang et al. [27] extracted ROIs using score-CAM [28] and learned FDG-PET images using a 3D subject-based method. The extracted ROIs undergo separate network processing and their features are fused through late fusion and average voting. Cui et al. [29] used FSL [30] to extract the hippocampus ROIs, considering not only the hippocampus area but also the degree of atrophy through late fusion.

C. 2D SLICE-BASED METHODS

In slice-based methods, 2D slices of modalities are used for training. Zhang et al. [31] proposed a 2D multimodal model, combining features extracted through channel-wise attention and convolution layers. Liang et al. [32] evaluated the early and late fusion performance using AlexNet [33] and ResNet [34], transitioning from 3D to a 2D approach to improve the training speed. Qiu et al. [35] used VGG-11 [36] to extract MRI features and the multilayer perceptron (MLP) for Mini-Mental Status Examination (MMSE) and logical memory capabilities. These features were combined using late fusion and majority voting. Valliani et al. [37] employed a pretrained ResNet from ImageNet to augment limited medical image data, to improve model performance. Pan et al. [38] introduced a multiview separable pyramid network (MiSePyNet) to train axial, sagittal, and coronal images separately using a slice-wise CNN. A separable convolution is used to train the spatial information using fewer parameters.

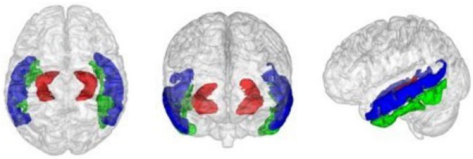

FIGURE 2. Display of ROI regions. Red stands for hippocampus; blue stands for middle temporal; green stands for inferior temporal regions.

III. METHOD

A. PREPROCESSING AND EXTRACTION METHOD OF ROIs

The distinct characteristics of MRI and PET images necessitate separate preprocessing methods. The preprocessing flowchart of the entire dataset is shown in Figure 1. In contrast to other studies, we opted to automatically extract ROIs using the segmentation label of Freesurfer [40], eliminating the need for manual specification of the center point of the ROI. Previous research by Frisoni et al. [18] highlighted the association between MRI and tau protein, as well as hippocampal atrophy, as markers of MCI. Similarly, Insel et al. [17] investigated the brain regions affected by amyloid β plaques and tau protein in early AD by selecting ROIs in the inferior temporal, middle temporal, and entorhinal regions. Based on these insights, we chose the hippocampus, middle temporal, and inferior temporal ROIs for training, as illustrated in Figure 2.

1) MRI IMAGES

The preprocessing for the MRI images was conducted as follows. First, all T1w MRI scans were transformed into the RAS coordinate system. FreeSurfer was used to apply motion correction, bias field correction, and intensity normalization.

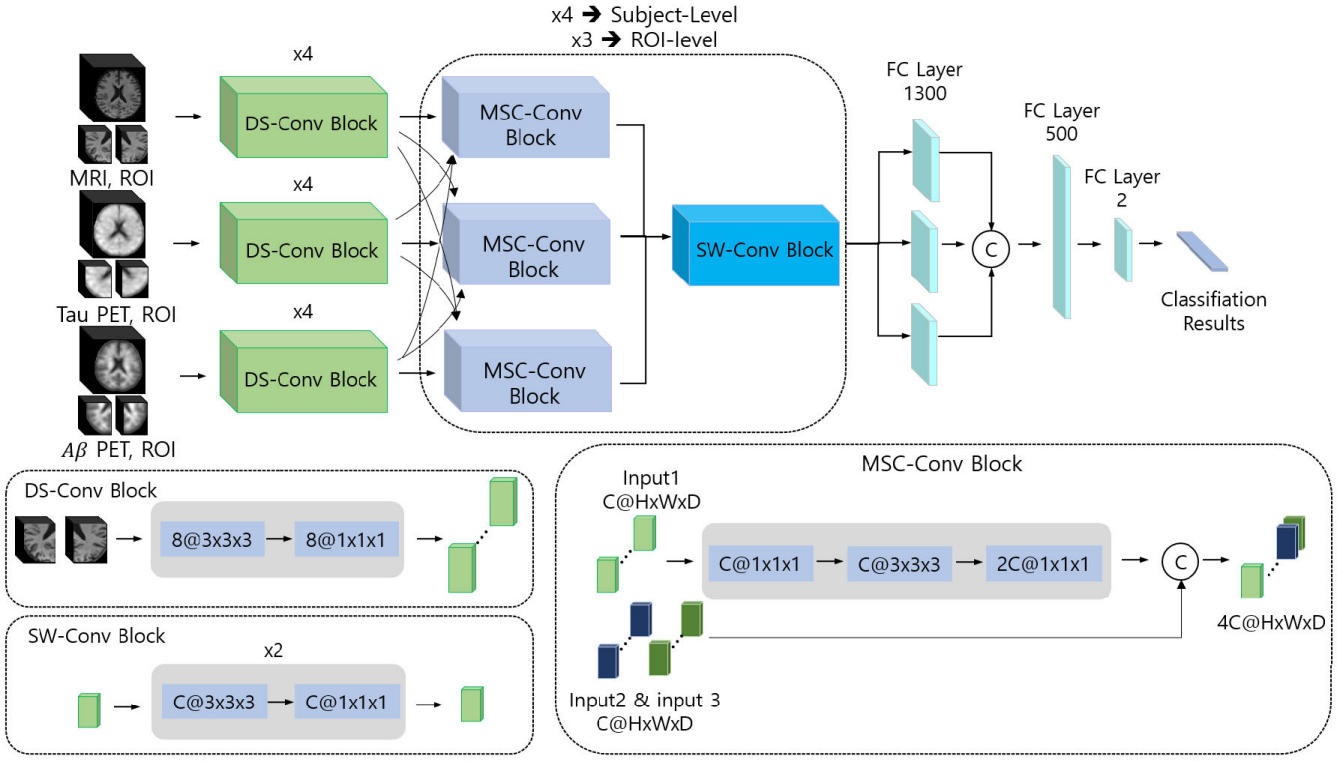

FIGURE 3. Multimodal architecture of the proposed model in three modalities. DS-conv, MSC-conv, and SW-conv blocks denote depthwise separable convolution, mixed skip connection convolution, and sharing weight convolution blocks, respectively.

Subsequently, nonlinear registration was performed from the T1w space to the MNI152 space using ANTs [41]. The registered MRI images were skull-stripped using Freesurfer. The entire image dataset in the 3D subject-based method was cropped to remove the background, resulting in images of consistent size. In the ROI-based method, two ROIs were automatically extracted using segmentation labels from Freesurfer. The hippocampal region ROI consisted of left and right ROIs of size 50 × 50×50. The hippocampus, middle temporal, and inferior temporal ROIs consisted of left and right ROIs of size 80 × 96×80. The processing time is 30 minutes per subject. All images were then normalized using min-max normalization.

2) PET IMAGES

The PET images were preprocessed as follows: The preprocessed data were collected from ADNI, where preprocessing included co-registration, averaging of six five-minute frames, image and voxel standardization, and uniform resolution. Each collected PET image of the subject was registered to the subject's T1w space via Freesurfer, and the PET images were registered based on the transformation matrix that matched the T1w MRI of each subject to the MNI 152 space. The remaining preprocessing steps for PET images followed the same as the MRI preprocessing procedure.

B. PROPOSED MODEL

The proposed model was inspired by MobileNet [16], [19], [42]. We designed an end-to-end deep-learning multimodal model for three binary classification tasks: AD vs CN, MCI vs CN, and sMCI vs pMCI. The overall architecture of the multimodal model using these three modalities is illustrated in Figure 3, two modalities in Figure 4, and single modality in Figure 5. Table 1 presents a detailed view of the model architecture based on the training method.

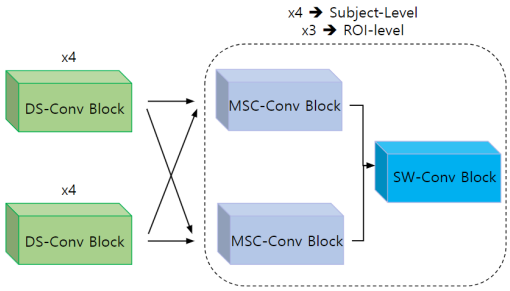

FIGURE 4. Multimodal with two modalities. Each modality has an MSC-Conv block which is concatenated to the output of the convolutional layers.

1) DEPTHWISE SEPARABLE CONVOLUTION BLOCK (DS-CONV BLOCK)

Convolutional neural networks, consisting of multiple convolutional layers, are among the most effective methods for feature extraction. Traditional CNN models typically include convolutional layers, normalization, and activation functions, such as rectified linear units (ReLUs) [43] or leaky ReLUs [44]. However, these activation functions may cause a loss of features during the feature extraction

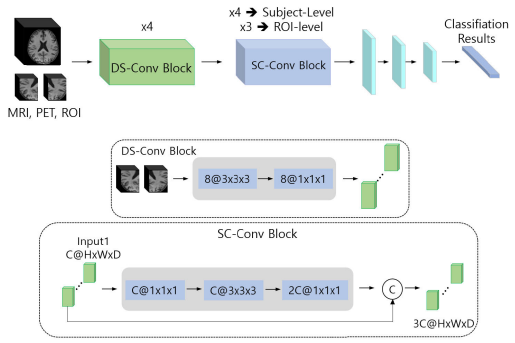

FIGURE 5. The proposed unimodal model architecture.

TABLE 1. Detailed architecture of 3D subject-based and ROI-based methods. Channels can differ in the case of the multimodal model with three modalities regarding the number of modalities to concatenate.

Layer Name	3D subject-based output size (CxHxWxD)	ROI-based output size (CxHxWxD)	ROI-based output size (CxHxWxD)	Model
Region	Whole	H	HT	
Input size	1x160x192x160	1x50x50x50	80x96x80	
DS-Conv Block (× 4)	8x80x96x80	8x25x25x25	8x40x48x40	$\begin{pmatrix}3\times3\times3\\1\times1\times1\end{pmatrix}\times4$ *Maxpool*
Bottleneck (sharing) Block 1,2,3 & Sharing block	24x40x48x40 72x20x24x20 216x10x12x10	24x12x12x12 72x6x6x6 216x3x3x3	24x20x24x20 72x10x12x10 216x5x6x5	$\left(\begin{pmatrix}1\times1\times1\\3\times3\times3\\1\times1\times1\\Maxpool\end{pmatrix}\begin{pmatrix}3\times3\times3\\1\times1\times1\end{pmatrix}\times2\right)\times3$
Bottleneck (sharing) Block 4 & Sharing block	648x5x6x5			$\begin{pmatrix}1\times1\times1\\3\times3\times3\\1\times1\times1\\Maxpool\end{pmatrix}\begin{pmatrix}3\times3\times3\\1\times1\times1\end{pmatrix}\times2$
FC Layer	1300 500 2	1300 500 2	1300 500 2	Flatten Dropout 0.5 FC Layer x3

TABLE 2. Experimental results for the multimodal model with or without activation function in DS-Conv blocks using the MRI and FDG-PET images of ADNI1 dataset.

Task	Activation Function	Acc	Bacc	F1
AD vs CN	Leaky ReLU	0.90 (±0.04)	0.90 (±0.04)	0.89 (±0.05)
	X	0.94 (±0.02)	0.95 (±0.01)	0.93 (±0.03)
sMCI vs pMCI	Leaky ReLU	0.79 (±0.04)	0.77 (±0.02)	0.71 (±0.03)
	X	0.84 (±0.05)	0.87 (±0.05)	0.81 (±0.05)
MCI vs CN	Leaky ReLU	0.76 (±0.02)	0.77 (±0.04)	0.81 (±0.01)
	X	0.77 (±0.04)	0.78 (±0.04)	0.81 (±0.04)

process. To address this, we designed a block for extracting the preserved features using four DS-Conv blocks without activation functions.

The DS-Conv block comprises two convolutional layers with kernel sizes of $3\times3\times3$ and $1\times1\times1$. After the four DS-Conv blocks, a max pooling layer with a stride of two was applied to reduce the resolution of the features. This step was applied consistently in both the unimodal and multimodal models. The experimental results comparing the presence and absence of an activation function are presented in Table 2.

2) MULTIMODAL MODEL

Most multimodal models of early AD diagnosis have traditionally employed independent feature extraction networks for each modality and later used a late fusion approach involving concatenation before passing through fully connected (FC) layers. However, late fusion has limitations in learning complex relationships between modalities because features are concatenated before reaching the FC layers. To address this, we adopted a middle-fusion approach to facilitate the learning of more intricate relationships between modalities.

Middle fusion was implemented using the mix skip connection convolution (MSC-Conv) block after features were extracted by the DS-Conv block. The MSC-Conv block includes a depthwise separable convolution, instance normalization [45], and a leaky ReLU, applied to each modality. After the features of one modality are extracted, the features of the other modality are concatenated using skip connections, similar to those in U-Net [46]. Subsequently, the features of each modality, obtained through the MSC-Conv block, are passed through the sharing weight convolution (SW-Conv) block to extract common features related to the labels by sharing the weights of the convolution layers. The SW-Conv block comprises two depthwise separable convolutional layers: batch normalization [47] and leaky ReLU.

The features extracted using the SW-Conv block are then fed into an FC layer in each modality and concatenated before passing through the last two FC layers for diagnosis. Each FC layer includes batch normalization and leaky ReLU. The architecture of the multimodal model with two modalities is shown in Figure 3. The 3D subject-based multimodal model employs four MSC-Conv and SW-Conv blocks. All convolutional and FC layers, except for the last FC layer, are initialized using Xavier Normalization [48].

3) ROI-BASED MODEL

To improve model performance and reduce the large computational costs associated with 3D subject-based methods, we designed an ROI-based model. The hippocampus ROI (H ROI) had dimensions of $50\times50\times50$, and the combined ROI of the hippocampus, middle temporal, and inferior temporal regions (HT ROI) had dimensions of $80\times96\times80$. These ROI sizes are smaller than the entire image size of $160\times192\times160$.

In the ROI-based model, the number of MSC-Conv and SW-Conv blocks was reduced from four to three, to prevent excessive compression and feature extraction from small ROI images and ensure optimal performance without information loss. By leveraging smaller ROIs, the ROI-based model offers computational advantages while maintaining a high level of accuracy in the early diagnosis of AD.

IV. EXPERIMENT

A. DATASET

Data used in the preparation of 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

TABLE 3. Demographics of ADNI1, ADNI2&3 Dataset.

Dataset	Group	Gender (Female/Male)	Age (Mean±STD)	MMSE (Mean±STD)	CDR (Mean±STD)
ADNI1	CN	101 (44/57)	75.33±7.38	29.07±1.11	0.03±0.13
	sMCI	129 (39/90)	74.91±7.38	27.33±1.65	1.50±0.74
	pMCI	79 (29/50)	75.05±6.78	26.85±1.65	1.63±0.86
	AD	84 (34/50)	75.10±7.35	23.64±2.10	4.40±1.66
ADNI2 & ADNI3	CN	258 (145/113)	70.17±5.76	29.14±1.07	0.06±0.21
	MCI	159 (99/60)	71.34±7.11	27.89±1.99	1.44±1.01
	AD	55 (32/22)	74.47±7.64	22.80±3.87	5.09±2.37

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). All ADNI studies are conducted according to the Good Clinical Practice guidelines, the Declaration of Helsinki, and U.S. 21 CFR Part 50 (Protection of Human Subjects), and Part 56 (Institutional Review Boards). Written informed consent was obtained from all participants before protocol-specific procedures were performed. The ADNI protocol was approved by the Institutional Review Boards of all of the participating institutions. Unlike most previous studies, we validated the proposed model using the entire ADNI series (ADNI1, ADNI2, and ADNI3). For the ADNI1 dataset, data from subjects who underwent both MRI and FDG-PET were collected. For training, only MRI and FDG-PET scans of subjects obtained within 60 days of training were used. Similarly, for the ADNI2 and ADNI3 datasets, images from subjects with MRI, Aβ PET, and Tau PET scans were considered. Subjects with scans taken within 60 days between modalities were included in the training, following the same criteria as in ADNI1. Moreover, only baseline scans were utilized for both training and testing purposes, while scans from subsequent visits were disregarded. This approach was adopted because identifying Alzheimer's disease at its initial stages is critical for effective intervention and care, highlighting the importance of baseline diagnosis for early detection.

Wen [25] pointed out data leakage issues in the process of splitting training and test datasets in many early diagnosis deep learning research studies on AD. Fung et al. [39] demonstrated a decrease in performance due to data leakage resulting from inappropriate data splitting methods. To avoid such issues and ensure more accurate model performance validation, we split our dataset using a stratified five-fold cross-validation based on the subject's diagnosis.

Additionally, we only included subjects with reliable diagnosis based on the following restrictions:

- CN (cognitively normal): Diagnosed as CN at baseline and remaining stable during the follow-up.
- sMCI (stable mild cognitive impairment): Diagnosed as MCI at baseline, with diagnosis not converting to CN or AD within 36 months.
- pMCI (progressive mild cognitive impairment): Diagnosed as MCI at baseline, and converting to AD within 36 months.
- MCI (mild cognitive impairment): Diagnosed as MCI at baseline, with diagnosis not reverting to CN within 36 months. MCI comprises sMCI and pMCI.
- AD (Alzheimer's Disease): Diagnosed as AD at baseline, the baseline diagnosis remains stable during the follow-up.

The ADNI1 dataset consists of 101 CN, 129 sMCI, 79 pMCI, and 84 AD subjects. The ADNI2 and ADNI3 datasets include 258 CN, 159 MCI, and 55 AD subjects. MCI subjects were not further divided into sMCI and pMCI groups because of the limited number of pMCI subjects in the ADNI2 and ADNI3 datasets. Clinical information, including clinical dementia rating (CDR) and Mini-Mental State Examination (MMSE), as well as the overall demographics of the collected dataset, are presented in Table 3.

B. EXPERIMENTAL SETTING

To ensure a more accurate evaluation of the proposed model, we conducted a stratified five-fold cross-validation split by subjects. The hyperparameters used to train the model are as follows:

- The model was trained for 500 epochs and early stopping was applied to prevent overfitting and reduce learning time.
- The Adam optimization algorithm was used for training [49], with an initial learning rate of 0.0003. The learning rate was designed to decrease for each batch step through the cosine annealing learning rate scheduler [50].
- No data augmentation techniques were employed during training.
- To prevent overfitting in the FC layer, a dropout rate of 0.5 was applied.
- All convolutional and FC layers were initialized using Xavier normalization.
- In the ROI-based model, each subject had two ROIs (left and right), resulting in one prediction per subject obtained through soft voting.

The AD vs CN, sMCI vs pMCI, and MCI vs CN experiments, were conducted on the ADNI1 dataset. Because of the small number of pMCI samples in the ADNI2 and ADNI3 datasets, the AD vs CN and MCI vs CN experiments

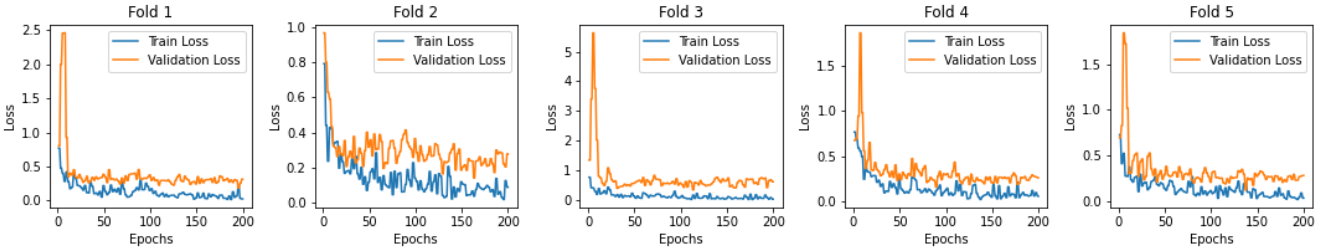

FIGURE 6. Learning curves for AD vs. CN task on ADNI1 dataset. The training process appears to be stable even on limited data.

were conducted separately on the ADNI2 and ADNI3 datasets.

All experiments were performed on a single NVIDIA A100 40GB. Each epoch took approximately 18 seconds to complete. Training 500 epochs for five folds took a total of 13 hours. Learning curves for AD vs. CN task on ADNI1 dataset are shown in Figure 6, indicating that no overfitting occurred during training.

C. EVALUATION METRICS

Three binary classification tasks were performed: AD diagnosis (AD vs CN), AD predictive diagnosis (sMCI vs pMCI), and early AD diagnosis (MCI vs CN). The results of the proposed model for all tasks were evaluated using the following metrics: accuracy (Acc), balanced accuracy (Bacc), sensitivity (Sen), specificity (Spe), and F1-Score (F1). These metrics are commonly used to evaluate the model performance of binary classification tasks and are calculated as follows:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$
$$Sen = \frac{TP}{TP + FN}$$
$$Spe = \frac{TN}{TN + FP}$$
$$Bacc = \frac{SEN + SPE}{2}$$
$$F1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

where TP, TN, FP, and FN, as depicted in Figure 7, refer to true positive, true negative, false positive, and false negative, respectively.

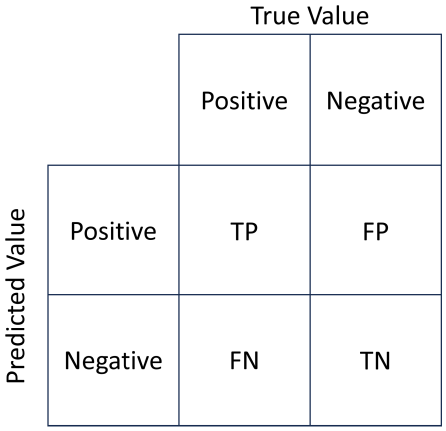

FIGURE 7. Definition of True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN).

V. RESULTS

A. RESULTS OF ADNI1 DATASET

The ADNI1 dataset was used to verify the performance of the proposed model. Three classification tasks were conducted to assess overall model performance: AD vs CN, sMCI vs pMCI, and MCI vs CN.

To evaluate the effects of the activation functions on DS-Conv Blocks, we experimented with the presence and absence of LeakyReLU. The results showed a performance improvement of 5% on BACC for the AD vs CN task, 10% on BACC for the sMCI vs pMCI task, and 1% on BACC for the MCI vs CN task in the absence of LeakyReLU for DS-Conv blocks. The experimental results for the presence or absence of an activation function in the DS-Conv blocks are listed in Table 2.

Next, the proposed model was evaluated using both the ROI and the entire image of each subject. In the AD vs CN task, the 3D subject-based model showed a similar performance to that of HT ROI, achieving a BACC of 0.95. In the sMCI vs pMCI task, the 3D subject-based model outperformed the ROI-based model, with a BACC of 0.87. Although the ROI-based model achieved a BACC of 0.81, indicating a lower performance than that of the entire image, the ROI-based model had a relatively smaller drop in performance with lower computational cost. This performance drop in the ROI-based model could be attributed to the ROI extraction method being based on tau proteins and amyloid beta plaques rather than using metabolic information from FDG-PET.

The results using the hippocampus, middle temporal, and inferior temporal ROIs (HT ROI) and the entire image (All) as input are presented in Table 5. The overall experimental results of the hippocampus ROI are shown in Table 8.

B. COMPARISON OF RESULTS WITH OTHER RESEARCH

The experimental results obtained using the ADNI1 dataset were compared with those of other studies to evaluate model performance, as shown in Table 4. Instead of using the

TABLE 4. Comparison with other research. All, H, and HT refer to the 3D subject-based method, RoI-based method (hippocampus region ROI), and RoI-based method (hippocampus, middle temporal, inferior temporal ROI regions), respectively.

Task	Method	Data type	Subjects number	Region	Acc	Bacc	Sen	Spe	F1
AD vs CN	Ruoxuan et al. [29]	MRI	192AD/223CN	H	0.92	0.92	0.91	0.94	-
	Junhao Wen et al. [25]	MRI	336AD/330CN	H	-	0.88	-	-	-
	Xin Zhang et al. [19]	MRI	200AD/231CN	All	0.91	0.91	0.91	0.92	-
	Xiaoxi Pan et al. [38]	FDG-PET	237AD/242CN	RoI	0.93	0.93	0.91	0.97	-
	Jin Zhang et al. [27]	FDG-PET	146AD/184CN	All	**0.98**	**0.97**	**0.96**	**0.99**	**0.96**
	Ours(ADNI1)	MRI, FDG-PET	84AD/101CN	All	0.94	0.95	0.99	0.91	0.93
sMCI vs pMCI	Ruoxuan et al. [29]	MRI	231sMCI/165pMCI	H	0.75	0.75	0.73	0.76	-
	Junhao Wen et al. [25]	MRI	298sMCI/295pMCI	H	-	0.74	-	-	-
	Xin Zhang et al. [19]	MRI	232sMCI/172pMCI	All	0.82	0.81	0.81	0.81	-
	Xiaoxi Pan et al. [38]	FDG-PET	360sMCI/166pMCI	All	0.83	0.80	0.72	**0.88**	-
	Ours(ADNI1)	MRI, FDG-PET	129sMCI/79pMCI	All	**0.84**	**0.87**	**0.97**	0.76	**0.81**
MCI vs CN	Ruoxuan et al. [29]	MRI	396MCI/223CN	H	0.75	0.74	0.77	0.70	-
	Jin Zhang et al. [27]	FDG-PET	347MCI/184CN	RoI	0.69	0.68	**0.78**	0.58	0.74
	Ours(ADNI1)	MRI, FDG-PET	208MCI/101CN	All	**0.77**	**0.78**	0.76	**0.80**	**0.81**

TABLE 5. ADNI1 results for the tasks of AD vs CN, MCI vs CN, and sMCI vs pMCI. HT denotes the enclosing region of hippocampus, middle temporal, and inferior temporal ROIs. All refers to the 3D subject-based method using the entire image of each subject. All the metrics are expressed as average(±) standard deviation. The overall experimental results are shown in Table 8.

Task	Region	Acc	Bacc	F1
AD vs CN	HT	0.96 (±0.03)	0.95 (±0.02)	0.93 (±0.03)
	All	0.94 (±0.02)	0.95 (±0.01)	0.93 (±0.03)
sMCI vs pMCI	HT	0.81 (±0.02)	0.81 (±0.03)	0.76 (±0.04)
	All	0.84 (±0.05)	0.87 (±0.05)	0.81 (±0.05)
MCI vs CN	HT	0.75 (±0.03)	0.73 (±0.03)	0.81 (±0.04)
	All	0.77 (±0.04)	0.78 (±0.04)	0.81 (±0.04)

ADNI3 dataset, which has been used in a limited number of research studies, we focused on comparisons with other studies that used the ADNI1 dataset, evaluating their models under similar conditions.

Many studies have highlighted the importance of avoiding data leakage during the process of splitting a dataset as this is crucial for accurate model performance evaluation. Therefore, the proposed model was compared with research that splits the dataset by subject to ensure a fair evaluation. Additionally, many studies report sensitivity and specificity but do not include the balanced accuracy (BACC) metric. To provide a comprehensive comparison, we calculated the BACC values according to the formula in the "Evaluation Metrics" section.

In the AD vs CN task, the proposed model achieved a BACC of 0.95, which is lower than the highest performance reported in other studies. However, in the sMCI vs pMCI task, the proposed model outperformed the other models with a BACC of 0.87. Furthermore, in the MCI vs CN task, the proposed model exhibited the best performance with a BACC of 0.77. These results indicate that the proposed model performs better in early AD diagnosis tasks despite having a smaller number of training subjects compared to other studies.

The superiority of the proposed model in the sMCI vs pMCI and MCI vs CN tasks highlights its effectiveness in early AD diagnosis, even with a limited number of training subjects compared to other studies.

C. RESULTS OF ADNI2 AND ADNI3 DATASETS

In the ADNI2 and ADNI3 datasets, experiments were conducted by adding a small number of ADNI2 subjects to the collected ADNI3 dataset to maximize the use of Tau PET and PET data for training purposes. However, the sMCI vs pMCI task could not be performed because of the small number of pMCI subjects. The results are shown in Table 6.

In the AD vs CN task, the proposed model exhibited the highest performance, with a BACC of 1.00 with all of HT ROI of MRI, Tau PET, and Aβ PET utilized. Similarly, in the MCI vs CN task, the highest performance of 0.76 for BACC was achieved when the HT ROI of MRI, Tau PET, and Aβ PET were all utilized. These results indicate that utilizing the HT ROI, which represents early changes in regions with tau protein and Aβ plaques, led to a performance improvement compared to using all images, despite the low computational cost.

In the MCI vs CN task, no significant performance improvement was observed under any of the experimental conditions. This can be attributed to two reasons. First, the MCI diagnosis is less stable on the ADNI1 dataset because of the short history of MCI subjects. Some subjects often transition back to CN during follow-up visits, resulting in an unstable MCI diagnosis. Second, there was no significant difference between the Tau PET and Aβ PET images of sMCI and CN; however, sMCI accounted for the majority of MCI subjects (143 out of 159), with few pMCI subjects (16). Tau PET and Aβ PET images of pMCI subjects differed from those of CN and sMCI subjects, making it difficult to accurately distinguish them.

TABLE 6. ADNI2 and ADNI3 results for the tasks of AD vs CN and MCI vs CN. HT denotes the enclosing region of hippocampus, middle Temporal, and inferior temporal ROIs. All refers to subject-based methods which use the entire image for each subject. All the metrics are expressed as average(±) standard deviation. The overall experimental results are shown in Table 9.

Task	Modality	Region	Acc	Bacc	Sen	Spe	F1
AD vs CN	MRI, Tau PET	HT	0.99(±0.01)	0.98(±0.02)	0.98(±0.04)	0.99(±0.01)	0.96(±0.03)
	MRI, Aβ PET	HT	0.94(±0.02)	0.94(±0.04)	0.95(±0.07)	0.94(±0.03)	0.83(±0.05)
	MRI, Tau PET, Aβ PET	HT	**1.00(±0.00)**	**1.00(±0.00)**	**1.00(±0.00)**	**1.00(±0.00)**	**1.00(±0.00)**
	MRI, Tau PET	All	0.95(±0.02)	0.92(±0.04)	0.88(±0.08)	0.97(±0.02)	0.86(±0.05)
	MRI, Aβ PET	All	0.94(±0.02)	0.97(±0.01)	**1.00(±0.00)**	0.93(±0.02)	0.82(±0.05)
	MRI, Tau PET, Aβ PET	All	0.99(±0.03)	0.99(±0.01)	**1.00(±0.00)**	0.99(±0.03)	0.96(±0.03)
MCI vs CN	MRI, Tau PET	HT	0.76(±0.03)	0.74(±0.03)	0.65(±0.06)	0.83(±0.06)	0.68(±0.04)
	MRI, Aβ PET	HT	0.71(±0.05)	0.71(±0.03)	0.69(±0.09)	0.72(±0.11)	0.64(±0.03)
	MRI, Tau PET, Aβ PET	HT	**0.79(±0.01)**	**0.76(±0.02)**	0.66(±0.07)	0.86(±0.05)	**0.69(±0.03)**
	MRI, Tau PET	All	0.74(±0.02)	0.74(±0.03)	**0.74(±0.11)**	0.73(±0.07)	0.67(±0.04)
	MRI, Aβ PET	All	0.73(±0.03)	0.73(±0.02)	0.72(±0.08)	0.73(±0.08)	0.66(±0.01)
	MRI, Tau PET, Aβ PET	All	0.76(±0.04)	0.74(±0.04)	0.68(±0.05)	0.80(±0.05)	0.68(±0.05)

TABLE 7. Comparison with early, late, and middle fusion models on hippocampus, middle temporal, and inferior temporal regions. Hippocampal-temporal region of MRI, Aβ PET, and Tau PET of ADNI2 and ADNI3 datasets were used.

Task	Fusion Stage	Acc	Bacc	Sen	Spe	F1
AD vs CN	Early Fusion	0.97(±0.04)	0.97(±0.03)	0.97(±0.06)	0.97(±0.05)	0.90(±0.10)
	Late Fusion	0.96(±0.02)	0.93(±0.04)	0.89(±0.06)	0.98(±0.02)	0.88(±0.07)
	Middle Fusion	**1.00(±0.00)**	**1.00(±0.00)**	**1.00(±0.00)**	**1.00(±0.00)**	**1.00(±0.00)**
MCI vs CN	Early Fusion	0.70(±0.03)	0.68(±0.04)	0.56(±0.11)	0.80(±0.06)	0.58(±0.07)
	Late Fusion	0.73(±0.03)	0.69(±0.04)	0.53(±0.12)	0.84(±0.06)	0.58(±0.07)
	Middle Fusion	**0.79(±0.01)**	**0.76(±0.02)**	**0.66(±0.07)**	**0.86(±0.05)**	**0.69(±0.03)**

TABLE 8. ADNI1 results on the tasks of AD vs CN, MCI vs CN, and sMCI vs pMCI. HT denotes the enclosing region of hippocampus, middle temporal, and inferior temporal regions. All refers to subject-based methods using the entire image for each subject. All the metrics are expressed as average(±) standard deviation.

Task	Modality	Region	Acc	Bacc	Sen	Spe	F1
AD vs CN	MRI	H	0.88(±0.05)	0.88(±0.04)	0.89(±0.05)	0.88(±0.09)	0.87(±0.05)
	FDG-PET	H	0.95(±0.02)	0.96(±0.01)	**1.00(±0.00)**	0.94(±0.03)	0.95(±0.02)
	MRI, FDG-PET	H	0.93(±0.03)	0.93(±0.03)	0.97(±0.06)	0.89(±0.00)	0.92(±0.03)
	MRI	HT	0.81(±0.02)	0.81 (±0.02)	0.81(±0.06)	0.81(±0.09)	0.79(±0.02)
	FDG-PET	HT	**0.97(±0.02)**	**0.97(±0.02)**	**1.00(±0.00)**	0.94(±0.03)	**0.97(±0.03)**
	MRI, FDG-PET	HT	0.96(±0.03)	0.95(±0.02)	0.95(±0.03)	**0.96(±0.03)**	0.93(±0.03)
	MRI	All	0.81(±0.04)	0.82(±0.03)	0.90(±0.10)	0.75(±0.14)	0.81(±0.03)
	FDG-PET	All	0.95(±0.03)	0.95(±0.03)	0.95(±0.06)	0.94(±0.00)	0.94(±0.03)
	MRI, FDG-PET	All	0.94(±0.02)	0.95(±0.01)	0.99(±0.03)	0.91(±0.03)	0.93(±0.03)
sMCI vs pMCI	MRI	H	0.78(±0.06)	0.79(±0.05)	0.81(±0.01)	0.77(±0.09)	0.71(±0.07)
	FDG-PET	H	0.76(±0.07)	0.74(±0.08)	0.67(±0.15)	**0.81(±0.10)**	0.67(±0.11)
	MRI, FDG-PET	H	0.79(±0.05)	0.79(±0.03)	0.80(±0.15)	0.78(±0.15)	0.73(±0.04)
	MRI	HT	0.76(±0.03)	0.76(±0.02)	0.74(±0.02)	0.78(±0.06)	0.70(±0.02)
	FDG-PET	HT	0.78(±0.03)	0.78(±0.03)	0.76(±0.10)	0.79(±0.08)	0.72(±0.04)
	MRI, FDG-PET	HT	0.81(±0.02)	0.81(±0.03)	0.83(±0.06)	0.79(±0.04)	0.76(±0.04)
	MRI	All	0.73(±0.06)	0.73(±0.03)	0.74(±0.12)	0.73(±0.14)	0.65(±0.04)
	FDG-PET	All	0.78(±0.04)	0.78(±0.04)	0.79(±0.13)	0.78(±0.08)	0.70(±0.05)
	MRI, FDG-PET	All	**0.84(±0.05)**	**0.87(±0.05)**	**0.97(±0.04)**	0.76(±0.06)	**0.81(±0.05)**
MCI vs CN	MRI	H	0.73(±0.07)	0.72(±0.01)	0.77(±0.14)	0.67(±0.12)	0.80(±0.07)
	FDG-PET	H	0.80(±0.05)	0.78(±0.03)	0.85(±0.10)	0.71(±0.05)	0.84(±0.05)
	MRI, FDG-PET	H	0.80(±0.02)	0.77(±0.02)	0.86(±0.05)	0.68(±0.05)	0.85(±0.02)
	MRI	HT	0.73(±0.02)	0.69(±0.03)	0.81(±0.06)	0.57(±0.10)	0.80(±0.02)
	FDG-PET	HT	**0.82(±0.02)**	**0.78(±0.02)**	**0.88(±0.04)**	0.69(±0.06)	**0.87(±0.02)**
	MRI, FDG-PET	HT	0.75(±0.03)	0.73(±0.03)	0.79(±0.09)	0.67(±0.11)	0.81(±0.04)
	MRI	All	0.72(±0.04)	0.68(±0.04)	0.78(±0.07)	0.58(±0.09)	0.79(±0.04)
	FDG-PET	All	0.77(±0.04)	0.75(±0.05)	0.80(±0.04)	0.70(±0.09)	0.82(±0.03)
	MRI, FDG-PET	All	0.77(±0.04)	0.78(±0.04)	0.76(±0.07)	**0.80(±0.08)**	0.81(±0.04)

D. THE EFFECT OF MODALITIES SETTINGS AND ROIs

We conducted experiments to evaluate the effect of all possible modality settings and Regions of Interest (ROIs), as shown in Table 8 and Table 9. Experimental results on the ADNI1 datasets demonstrate that combining FDG-PET with the hippocampal (HT) region yielded the highest

TABLE 9. ADNI2 and ADNI3 results on the tasks of AD vs CN and MCI vs CN. HT denotes the enclosing region of hippocampus, middle temporal, and inferior temporal regions. All refers to subject-based methods using the entire image for each subject. All the metrics are expressed as average(±) standard deviation.

Task	Modality	Region	Acc	Bacc	Sen	Spe	F1
AD vs CN	MRI	H	0.93(±0.02)	0.91(±0.05)	0.88(±0.11)	0.95(±0.03)	0.80(±0.05)
	Tau PET	H	0.99(±0.03)	0.99(±0.01)	**1.00(±0.00)**	0.99(±0.03)	0.96(±0.06)
	Aβ PET	H	0.96(±0.03)	0.95(±0.04)	0.95(±0.07)	0.96(±0.03)	0.88(±0.08)
	MRI, Tau PET	H	0.99(±0.01)	0.99(±0.01)	**1.00(±0.00)**	0.99(±0.01)	0.96(±0.04)
	MRI, Aβ PET	H	0.96(±0.02)	0.97(±0.03)	0.98(±0.04)	0.96(±0.02)	0.90(±0.05)
	MRI, Tau PET, Aβ PET	H	0.98(±0.02)	0.98(±0.04)	0.97(±0.06)	0.98(±0.02)	0.93(±0.07)
	MRI	HT	0.90(±0.05)	0.79(±0.10)	0.63(±0.20)	0.95(±0.04)	0.67(±0.11)
	Tau PET	HT	0.99(±0.02)	1.00(±0.01)	**1.00(±0.00)**	0.99(±0.02)	0.97(±0.06)
	Aβ PET	HT	0.95(±0.03)	0.95(±0.04)	0.97(±0.07)	0.94(±0.03)	0.84(±0.09)
	MRI, Tau PET	HT	0.99(±0.01)	0.98(±0.02)	0.98(±0.04)	0.99(±0.01)	0.96(±0.03)
	MRI, Aβ PET	HT	0.94(±0.02)	0.94(±0.04)	0.95(±0.07)	0.94(±0.03)	0.83(±0.05)
	MRI, Tau PET, Aβ PET	HT	**1.00(±0.00)**	**1.00(±0.00)**	**1.00(±0.00)**	**1.00(±0.00)**	**1.00(±0.00)**
	MRI	All	0.86(±0.06)	0.80(±0.04)	0.70(±0.17)	0.89(±0.09)	0.61(±0.06)
	Tau PET	All	0.93(±0.10)	0.90(±0.11)	0.84(±0.17)	0.96(±0.06)	0.85(±0.14)
	Aβ PET	All	0.94(±0.03)	0.90(±0.04)	0.83(±0.07)	0.96(±0.03)	0.82(±0.07)
	MRI, Tau PET	All	0.95(±0.02)	0.92(±0.04)	0.88(±0.08)	0.97(±0.02)	0.86(±0.05)
	MRI, Aβ PET	All	0.94(±0.02)	0.97(±0.01)	**1.00(±0.00)**	0.93(±0.02)	0.82(±0.05)
	MRI, Tau PET, Aβ PET	All	0.99(±0.03)	0.99(±0.01)	**1.00(±0.00)**	0.99(±0.03)	0.96(±0.03)
MCI vs CN	MRI	H	0.71(±0.05)	0.70(±0.06)	0.63(±0.11)	0.77(±0.07)	0.62(±0.09)
	Tau PET	H	0.77(±0.02)	0.74(±0.02)	0.61(±0.06)	0.87(±0.04)	0.67(±0.03)
	Aβ PET	H	0.74(±0.02)	0.70(±0.02)	0.53(±0.07)	0.87(±0.06)	0.60(±0.03)
	MRI, Tau PET	H	0.77(±0.02)	0.75(±0.03)	0.67(±0.10)	0.82(±0.06)	0.69(±0.04)
	MRI, Aβ PET	H	0.73(±0.03)	0.71(±0.03)	0.62(±0.09)	0.80(±0.08)	0.63(±0.04)
	MRI, Tau PET, Aβ PET	H	0.76(±0.04)	0.72(±0.03)	0.56(±0.10)	**0.88(±0.10)**	0.63(±0.05)
	MRI	HT	0.69(±0.04)	0.68(±0.02)	0.66(±0.10)	0.71(±0.10)	0.62(±0.03)
	Tau PET	HT	0.74(±0.02)	0.71(±0.03)	0.54(±0.08)	0.87(±0.06)	0.61(±0.05)
	Aβ PET	HT	0.70(±0.02)	0.67(±0.02)	0.55(±0.06)	0.79(±0.03)	0.58(±0.02)
	MRI, Tau PET	HT	0.76(±0.03)	0.74(±0.03)	0.65(±0.06)	0.83(±0.06)	0.68(±0.04)
	MRI, Aβ PET	HT	0.71(±0.05)	0.71(±0.03)	0.69(±0.09)	0.72(±0.11)	0.64(±0.03)
	MRI, Tau PET, Aβ PET	HT	**0.79(±0.01)**	**0.76(±0.02)**	0.66(±0.07)	0.86(±0.05)	**0.69(±0.03)**
	MRI	All	0.73(±0.03)	0.69(±0.02)	0.56(±0.08)	0.82(±0.08)	0.60(±0.04)
	Tau PET	All	0.74(±0.03)	0.71(±0.02)	0.59(±0.09)	0.82(±0.09)	0.62(±0.04)
	Aβ PET	All	0.73(±0.04)	0.70(±0.03)	0.58(±0.07)	0.81(±0.07)	0.61(±0.04)
	MRI, Tau PET	All	0.74(±0.02)	0.74(±0.03)	**0.74(±0.11)**	0.73(±0.07)	0.67(±0.04)
	MRI, Aβ PET	All	0.73(±0.03)	0.73(±0.02)	0.72(±0.08)	0.73(±0.08)	0.66(±0.01)
	MRI, Tau PET, Aβ PET	All	0.76(±0.04)	0.74(±0.04)	0.68(±0.05)	0.80(±0.05)	0.68(±0.05)

performances in AD vs. CN and MCI vs. CN tasks, while utilizing all modalities with the whole scans achieved better results in sMCI vs. pMCI task. Similarly, the results on the ADNI2 & ADNI3 datasets show that using all modalities and the HT region consistently improved AD vs. CN and MCI vs. CN tasks. Overall, in most cases, employing multiple modalities alongside focusing on the HT region improves diagnostic accuracy.

E. FUSION STRATEGIES

Experiments were conducted to verify the performance of the proposed model in comparison with the early and late fusion methods. The comparison results for the early, middle, and late fusion models are presented in Table 7. In the early fusion model, MRI, Tau PET, and Amyloid PET images were concatenated during the data input stage and then fed into the unimodal model. The late fusion model concatenated the three modalities at the FC layer after passing through three different backbones. The experimental results demonstrated that the middle fusion model proposed in this study showed the best performance, outperforming both early and late fusion methods.

VI. DISCUSSION

In this study, a multimodal model was proposed for early Alzheimer's disease diagnosis using MRI, FDG-PET, Tau PET, and PET data. The model utilized DS-Conv blocks to extract the preserved features by removing activation functions and MSC-Conv blocks with skip connections to learn the complex relationships between modalities. The SW-Conv block was employed to share weights and extract common features related to labels between modalities, thereby enabling efficient feature fusion. Additionally, novel ROI extraction methods were proposed to reduce the computational costs while maintaining or improving model performance. The ROI extraction method focused on the hippocampus, middle temporal, and inferior temporal regions known for early changes in Aβ plaques and tau protein in the brain during the early stages of AD.

The proposed model evaluated on the ADNI1 dataset outperformed other research models in the sMCI vs pMCI and MCI vs CN tasks, despite having a smaller number of subjects compared to other studies. We further experimented with the AD vs CN and MCI vs CN tasks using Tau PET and Aβ PET, which provide information on the tau protein and Aβ plaques. In the ADNI2 and ADNI3 datasets, the proposed

model demonstrated good performance in the AD vs CN task, particularly when utilizing HT ROI of MRI, Tau PET, and Aβ PET. This indicates the significance of focusing on early change regions related to the tau protein and Aβ plaques for early AD diagnosis.

However, there are two main drawbacks in this research, particularly in experiments conducted on ADNI2 and ADNI3 datasets. Although the proposed method achieved perfect AD vs CN classification, this was limited by the small number of AD subjects in the test set (11 subjects for each fold). Consequently, further experiments on datasets with more AD subjects are necessary to validate its performance. Furthermore, the proposed model did not demonstrate a significant performance improvement in the MCI vs CN task, potentially due to the unstable MCI diagnosis resulting from short histories and the small proportion of pMCI subjects. To address this limitation, additional research involving a larger number of pMCI patients and improved models capable of discerning the differences between sMCI and CN subjects is required.

VII. CONCLUSION

In conclusion, the proposed multimodal model demonstrated promising results for early AD diagnosis, especially in distinguishing sMCI from pMCI and MCI from CN subjects. Despite some limitations, such as the availability of PET images and data labeling costs, we believe that future research focusing on generating PET images or employing robust self-supervised and few-shot learning methods for medical images can help overcome these challenges and further enhance the performance of early AD diagnostic models. Overall, this study contributes to the field of early AD diagnosis by introducing a multimodal approach and exploring the significance of ROI-based methods in improving model efficiency and performance. This opens avenues for further research and potential applications in the early detection and understanding of Alzheimer's disease.

VIII. DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

APPENDIX A
RESULTS OF ADNI1 DATASET

The overall experimental results of ADNI1 dataset are shown in Table 8.

APPENDIX B
RESULTS OF ADNI2 AND ADNI3 DATASETS

The overall experimental results of ADNI2 and ADNI3 datasets are shown in Table 9.

ACKNOWLEDGMENT

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

REFERENCES

[1] M. Crous-Bou, C. Minguilln, N. Gramunt, and J. L. Molinuevo, "Alzheimer's disease prevention: From risk factors to early intervention," *Alzheimer's Res. Therapy*, vol. 9, no. 1, pp. 1–9, 2017.

[2] G. McKhann, D. Drachman, M. Folstein, R. Katzman, D. Price, and E. M. Stadlan, "Clinical diagnosis of Alzheimer's disease: Report of the NINCDS ADRDA Work Group under the auspices of department of health and human services task force on Alzheimer's disease," *Neurology*, vol. 34, no. 7, p. 939, 1984.

[3] R. Brookmeyer, E. Johnson, K. Ziegler-Graham, and H. M. Arrighi, "Forecasting the global burden of Alzheimer's disease," *Alzheimer's Dementia*, vol. 3, no. 3, pp. 186–191, Jul. 2007.

[4] L.-K. Huang, S.-P. Chao, and C.-J. Hu, "Clinical trials of new drugs for Alzheimer disease," *J. Biomed. Sci.*, vol. 27, no. 1, pp. 1–13, Dec. 2020.

[5] M. A. Ebrahimighahnavieh, S. Luo, and R. Chiong, "Deep learning to detect Alzheimer's disease from neuroimaging: A systematic literature review," *Comput. Methods Programs Biomed.*, vol. 187, Apr. 2020, Art. no. 105242.

[6] S. Fathi, M. Ahmadi, and A. Dehnad, "Early diagnosis of Alzheimer's disease based on deep learning: A systematic review," *Comput. Biol. Med.*, vol. 146, Jul. 2022, Art. no. 105634.

[7] J. K. Gahm, Y. Tang, and Y. Shi, "Patch-based mapping of transentorhinal cortex with a distributed atlas," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2018, pp. 689–697.

[8] J. Ouyang, Q. Zhao, E. Adeli, E. V. Sullivan, and K. M. Pohl, "Self-supervised longitudinal neighbourhood embedding," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervent.*, 2021, pp. 80–89.

[9] M. Hon and N. M. Khan, "Towards Alzheimer's disease classification through transfer learning," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2017, pp. 1166–1169.

[10] R. M. Nisbet and J. Götz, "Amyloid-and tau in Alzheimer's disease: Novel pathomechanisms and non-pharmacological treatment strategies," *J. Alzheimer's Disease*, vol. 64, no. s1, pp. S517–S527, Jun. 2018.

[11] F. Kametani and M. Hasegawa, "Reconsideration of amyloid hypothesis and tau hypothesis in Alzheimer's disease," *Frontiers Neurosci.*, vol. 12, p. 25, Jan. 2018.

[12] E. Mohandas, V. Rajmohan, and B. Raghunath, "Neurobiology of Alzheimer's disease," *Indian J. Psychiatry*, vol. 51, no. 1, pp. 55–61, 2009.

[13] J. W. Vogel, A. L. Young, N. P. Oxtoby, R. Smith, R. Ossenkoppele, O. T. Strandberg, R. La Joie, L. M. Aksman, M. J. Grothe, Y. Iturria-Medina, A. D. N. Initiative, M. J. Pontecorvo, M. D. Devous, G. D. Rabinovici, D. C. Alexander, C. H. Lyoo, A. C. Evans, and O. Hansson, "Four distinct trajectories of tau deposition identified in Alzheimer's disease," *Nature Med.*, vol. 27, no. 5, pp. 871–881, 2021.

[14] C. R. Jack, D. A. Bennett, K. Blennow, M. C. Carrillo, H. H. Feldman, G. B. Frisoni, H. Hampel, W. J. Jagust, K. A. Johnson, D. S. Knopman, R. C. Petersen, P. Scheltens, R. A. Sperling, and B. Dubois, "A/T/N: An unbiased descriptive classification scheme for Alzheimer disease biomarkers," *Neurology*, vol. 87, no. 5, pp. 539–547, Aug. 2016.

[15] S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett, "The Alzheimer's disease neuroimaging initiative," *Neuroimag. Clin.*, vol. 15, no. 4, pp. 869–877, Nov. 2005.

[16] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[17] P. S. Insel, C. B. Young, P. S. Aisen, K. A. Johnson, R. A. Sperling, E. C. Mormino, and M. C. Donohue, "Tau positron emission tomography in preclinical Alzheimer's disease," *Brain*, vol. 146, no. 2, pp. 700–711, Feb. 2022.

[18] G. B. Frisoni, N. C. Fox, C. R. Jack, P. Scheltens, and P. M. Thompson, "The clinical use of structural MRI in Alzheimer disease," *Nature Rev. Neurol.*, vol. 6, no. 2, pp. 67–77, Feb. 2010.

[19] X. Zhang, L. Han, W. Zhu, L. Sun, and D. Zhang, "An explainable 3D residual self-attention deep neural network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 11, pp. 5289–5297, Nov. 2022.

[20] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[21] E. Yee, K. Popuri, and M. Faisal, "Beg and Alzheimer's disease neuroimaging initiative, quantifying brain metabolism from FDG-PET images into a probability of Alzheimer's dementia score," *Hum. Brain Mapping*, vol. 41, no. 1, pp. 5–16, 2020.

[22] A. Punjabi, A. Martersteck, Y. Wang, T. B. Parrish, and A. K. Katsaggelos, "Neuroimaging modality fusion in Alzheimer's classification using convolutional neural networks," *PLoS ONE*, vol. 14, no. 12, Dec. 2019, Art. no. e0225759.

[23] J. Zou, D. Park, A. Johnson, X. Feng, M. Pardo, J. France, Z. Tomljanovic, A. M. Brickman, D. P. Devanand, J. A. Luchsinger, W. C. Kreisl, and F. A. Provenzano, "Deep learning improves utility of tau PET in the study of Alzheimer's disease," *Alzheimer's Dementia, Diagnosis, Assessment Disease Monitor.*, vol. 13, no. 1, Jan. 2021, Art. no. e12264.

[24] S. Spasov, L. Passamonti, A. Duggento, P. Liò, and N. Toschi, "A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer's disease," *NeuroImage*, vol. 189, pp. 276–287, Apr. 2019.

[25] J. Wen, E. Thibeau-Sutre, M. Diaz-Melo, J. Samper-González, A. Routier, S. Bottani, D. Dormont, S. Durrleman, N. Burgos, and O. Colliot, "Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation," *Med. Image Anal.*, vol. 63, Jul. 2020, Art. no. 101694.

[26] Y. Huang, J. Xu, Y. Zhou, T. Tong, and X. Zhuang, "Diagnosis of Alzheimer's disease via multi-modality 3D convolutional neural network," *Frontiers Neurosci.*, vol. 13, p. 509, May 2019.

[27] J. Zhang, X. He, L. Qing, Y. Xu, Y. Liu, and H. Chen, "Multi-scale discriminative regions analysis in FDG-PET imaging for early diagnosis of Alzheimer's disease," *J. Neural Eng.*, vol. 19, no. 4, Aug. 2022, Art. no. 046030.

[28] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, "Score-CAM: Score-weighted visual explanations for convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 111–119.

[29] R. Cui and M. Liu, "Hippocampus analysis by combination of 3-D DenseNet and shapes for Alzheimer's disease diagnosis," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 5, pp. 2099–2107, Sep. 2019.

[30] M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, and S. M. Smith, "FSL," *NeuroImage*, vol. 62, no. 2, pp. 782–790, 2012.

[31] T. Zhang and M. Shi, "Multi-modal neuroimaging feature fusion for diagnosis of Alzheimer's disease," *J. Neurosci. Methods*, vol. 341, Jul. 2020, Art. no. 108795.

[32] G. Liang, X. Xing, L. Liu, Y. Zhang, Q. Ying, A.-L. Lin, and N. Jacobs, "Alzheimer's disease classification using 2D convolutional neural networks," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 3008–3012.

[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[35] S. Qiu, G. H. Chang, M. Panagia, D. M. Gopal, R. Au, and V. B. Kolachalama, "Fusion of deep learning models of MRI scans, mini–mental state examination, and logical memory test enhances diagnosis of mild cognitive impairment," *Alzheimer's Dementia, Diagnosis, Assessment Disease Monitor.*, vol. 10, no. 1, pp. 737–749, Jan. 2018.

[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[37] A. Valliani and A. Soni, "Deep residual nets for improved Alzheimer's diagnosis," in *Proc. 8th ACM Int. Conf. Bioinf., Comput. Biol. Health Informat.*, Aug. 2017, p. 615.

[38] X. Pan, T.-L. Phan, M. Adel, C. Fossati, T. Gaidon, J. Wojak, and E. Guedj, "Multi-view separable pyramid network for AD prediction at MCI stage by 18F-FDG brain PET imaging," *IEEE Trans. Med. Imag.*, vol. 40, no. 1, pp. 81–92, Jan. 2021.

[39] Y. Ren Fung, Z. Guan, R. Kumar, J. Yeahuay Wu, and M. Fiterau, "Alzheimer's disease brain MRI classification: Challenges and insights," 2019, *arXiv:1906.04231*.

[40] B. Fischl, "FreeSurfer," *NeuroImage*, vol. 62, no. 2, pp. 774–781, Aug. 2012.

[41] B. Avants, N. J. Tustison, and G. Song, "Advanced normalization tools: V1.0," *Insight J.*, vol. 2, no. 365, pp. 1–35, Jul. 2009.

[42] J. Dolz, K. Gopinath, J. Yuan, H. Lombaert, C. Desrosiers, and I. B. Ayed, "HyperDense-Net: A hyper-densely connected CNN for multi-modal image segmentation," *IEEE Trans. Med. Imag.*, vol. 38, no. 5, pp. 1116–1126, May 2019.

[43] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 807–814.

[44] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. Int. Conf. Mach. Learn.*, 2013, vol. 30, no. 1, p. 3.

[45] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2016, *arXiv:1607.08022*.

[46] O. Ronneberger and P. Fischer, "Thomas Brox, U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.

[47] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[48] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.

[49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[50] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," 2016, *arXiv:1608.03983*.

SEUNG KYU KIM received the B.S. degree in mechanical engineering from Pusan National University, Republic of Korea, in 2019, where he is currently pursuing the M.S. degree in artificial intelligence. His research interests include computer vision, medical image analysis, and early diagnosis of Alzheimer's disease.

QUAN ANH DUONG received the B.S. degree in computer science from the VNU University of Science, Vietnam, in 2021. He is currently pursuing the M.S. degree in computer engineering from Pusan National University, Republic of Korea. His research interests include computer vision, medical image analysis, and cortical surface analysis.

JIN KYU GAHM received the B.S. degree in computer engineering from Seoul National University, Republic of Korea, in 2000, and the Ph.D. degree in computer science from the University of California at Los Angeles, Los Angeles, CA, USA, in 2014. He was a Postdoctoral Associate with the Laboratory of Neuro Imaging, University of Southern California, USA, from 2014 to 2019. He is currently an Associate Professor with the School of Computer Science and Engineering, Pusan National University. His research interests include medical image processing, machine learning, and brain shape analysis.