IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Early Diagnosis of Alzheimer's Disease using 18F-FDG PET with Soften Latent Representation

ABDUL REHMAN[1], (Member,EEE), MYUNG-KYU YI[2], ABDUL MAJEED[2], AND SEONG OUN HWANG.[2], (Member, IEEE)
[1]Department of IT Convergence Engineering, Gachon University, South Korea
[2]Department of Computer Engineering, Gachon University, South Korea

Corresponding author: Seong Oun Hwang (e-mail: seongoun.hwang@gmail.com).

ABSTRACT Mild cognitive impairment (MCI) is an early stage of Alzheimer's disease (AD), which is currently incurable. Early diagnosis of AD is essential for effective intervention since the World Alzheimer's Report 2015 predicted the number of cases will triple by 2050. The 18F-FDG PET imaging technique, although effective in detecting metabolic activities in the brain, faces challenges such as low signal-to-noise ratios and limited data availability, which complicates the existing methods to extract necessary lesion information effectively for diagnosing early-stage MCI. To overcome these challenges, we introduce a novel deep learning-based model, ResGLPyramid, that combines convolution operations, MobileViTv3, and a global-local attention module (GLAM) block, to capture local and global representations. By utilizing a softened cross-entropy (SCE) objective function, the model reduces overfitting, improves generalization, and enhances the detection of subtle metabolic changes. The proposed model enhances the sensitivity and specificity of Alzheimer's detection by leveraging local- and long-range interactions among critical diagnostic features that lead to more precise and efficient analyses. Experimental results show the ResGLPyramid model achieved an accuracy of 92.75% in classifying MCI and AD individuals, which is 3.44% higher than state-of-the-art methods.

INDEX TERMS 18F-FDG PET, Alzheimer's disease, deep learning, global feature representation, local feature representation, MobileViT

I. INTRODUCTION

ALzheimer's disease (AD) is a chronic and progressive brain disorder, characterized by a form of dementia, that gradually impairs cognitive abilities and eventually disrupts the inherent skills to perform even the simplest tasks [1]. As reported by the World Alzheimer's Report 2015, approximately 46.85 million people globally are affected by AD and other forms of dementia, with projections indicating that figure will double by 2030 and triple by 2050 [2]. Mild cognitive impairment (MCI) represents a transitional stage between normal cognition (NC) function and AD, characterized by slightly reduced cognitive abilities that do not significantly affect daily activities. Despite this, MCI substantially increases the risk of developing dementia, highlighting the critical need for precise early detection methods and effective interventions to slow symptom progression and manage cognitive decline [3].

Diagnostic techniques such as magnetic resonance imaging (MRI) [4], [5] and positron emission tomography (PET) [6] are vital tools in detecting AD and MCI. MRI captures structural and functional changes in the brain, while PET, particularly 18F-fluorodeoxyglucose PET (18F-FDG PET), is more effective in the early stages of AD because it measures metabolic activities at the tissue level [7], detecting decreases in metabolic activities before structural changes occur [8]. However, PET faces several challenges: images often have a low signal-to-noise ratio, contain repetitive information among slices within the same class, and suffer from a lack of sufficient data volume. These issues complicate the use of deep learning (DL) models for accurate AD prediction, underscoring the necessity for improved imaging techniques and data handling to enhance the diagnosis and management of AD and MCI [9] [10].

To address these challenges, deep learning approaches have become essential for classifying and diagnosing neurological conditions such as NC, MCI, and AD. Various

studies have employed DL models to enhance diagnostic accuracy. For instance, Liu et al. [11] leveraged convolutional neural network (CNN) layers, to extract features from brain slices, processed by a gated recurrent unit (GRU), in order to integrate inter-slice features, achieving an area under the curve (AUC) of 95.3% for AD vs. NC and 83.9% for MCI vs. NC. Ding et al. [12] applied the InceptionV3 model [13] with ImageNet weights [14] for extraction of AD brain features, achieving a detection rate of 87.5% for AD and 61% for MCI. Zhang et al. [15] employed CNNs, pooling layers, and fully connected networks (FCNs) to extract information and correlate it with clinical scores, achieving an 84.2% detection rate, which Kim et al. [16] improved to 91.02% by substituting a global average pooling layer for FCNs to classify AD and NC from multi-slice PET images. In the preceding year, Song et al. [17] introduced a U-shaped multi-scale architecture that effectively extracts inter-slice and intra-slice features, enhancing the accuracy of AD and NC classification to 92% and achieving 73.01% and 72.6% accuracy for AD vs. MCI and MCI vs. NC, respectively.

Despite these advancements, significant limitations persist in the existing approaches. First, these methods often fail to capture comprehensive features from PET images, primarily extracting local features and overlooking crucial global contextual information. This limitation hinders early detection of AD, which is vital for effective intervention and management. Secondly, many models struggle with classification confidence, particularly when distinguishing between closely related classes due to the subtle variations in PET image features. The use of traditional cross-entropy loss impairs this issue by not effectively managing data points near the decision boundary. To address these challenges, Chen et al. [18] shifted to a contrastive loss approach within a double-attention-based CNN framework, although this method demands extensive training time and relies heavily on effective data augmentation strategies. These challenges underline the need for ongoing development in DL techniques to improve the robustness and accuracy of diagnosing AD and MCI using PET imaging.

Addressing these constraints, the Vision Transformer (ViT) model [19] presents a promising alternative with a self-attention mechanism that captures long-range dependencies, potentially outperforming traditional CNNs in feature extraction. However, ViT faces challenges such as an inductive bias problem and higher data requirements, which complicates its application in the medical field where data acquisition is severely restricted. Consequently, most approaches employ supervised fine-tuning or self-supervised learning, and although some adopt transfer learning, its effectiveness is often limited by significant domain shifts [20] [21].To overcome these challenges, researchers have explored combining a CNN with ViT to leverage both local and global feature extraction capabilities [22] [23] [24], although these approaches demand significant computational resources. Another innovative solution, MobileViT [25], targets efficient operation on mobile and edge devices by balancing local and global information processing with fewer model parameters, but this design may not fully meet the accuracy requirements for complex PET image classification of MCI and AD [26]. These efforts underscore the critical need for innovative approaches that balance efficiency and precision in medical imaging diagnostics.

In this paper, we introduce a novel DL model, ResGLPyramid, consisting of the Tri-Convolutional-Transformer (TCT) and Global Local Attention Module (GLAM) module, designed to enhance the diagnosis of MCI and AD using PET images. The proposed DL model combines the strengths of the CNN, MobileViTv3, and GLAM module to address the challenges inherent in PET imaging diagnostics. This model is precisely structured to extract local and global features, effectively solving the inductive bias problem associated with traditional transformer models and markedly increasing AD and MCI classification accuracy. Our significant contributions are as follows:

1) To improve the detection of MCI and AD by overcoming the challenges of analyzing PET images with a complex nature, we proposed ResGLPyramid which incorporates TCT and GLAM modules to process local information and leverage global contextual understanding effectively. ResGLPyramid employs patch-based self-attention to capture long-range dependencies and fuse them to expand the scope of model attributes. This strategy allows for a comprehensive analysis by capturing detailed and predominant patterns in PET images and significantly helps the model identify subtle nuances associated with early stage of AD to overcome redundant information, boosting diagnostic accuracy. The ResGLPyramid model achieved an accuracy of 92.75% in the classification of MCI and AD individuals, which is 3.44% higher than the state-of-the-art (SOTA) method [18]. The results show that predicted regions of interest are interpretable and consistent with the AD lesion region in clinical studies.
2) To the best of our knowledge, this is the first DL model that employs label smoothing with cross-entropy, referred to as softened cross-entropy (SCE), to optimize the objective function in orders to diagnose AD's early stage. This technique effectively increases the distance between feature representations of different classes while minimizing the gap within the same class, enhancing the model's robustness and accuracy in distinguishing between closely related diagnostic categories. This combination addresses the problem of overfitting caused by highly similar PET image slices. Compared to using only cross-entropy with the ResGLPyramid model, the proposed combination can achieve 6.5% higher accuracy in classifications between AD and NC, 6.4% higher accuracy in classifications between AD and MCI, and 6.23% higher accuracy in classifications between NC and MCI. This technique can be applied in various similar medical fields and is beneficial for

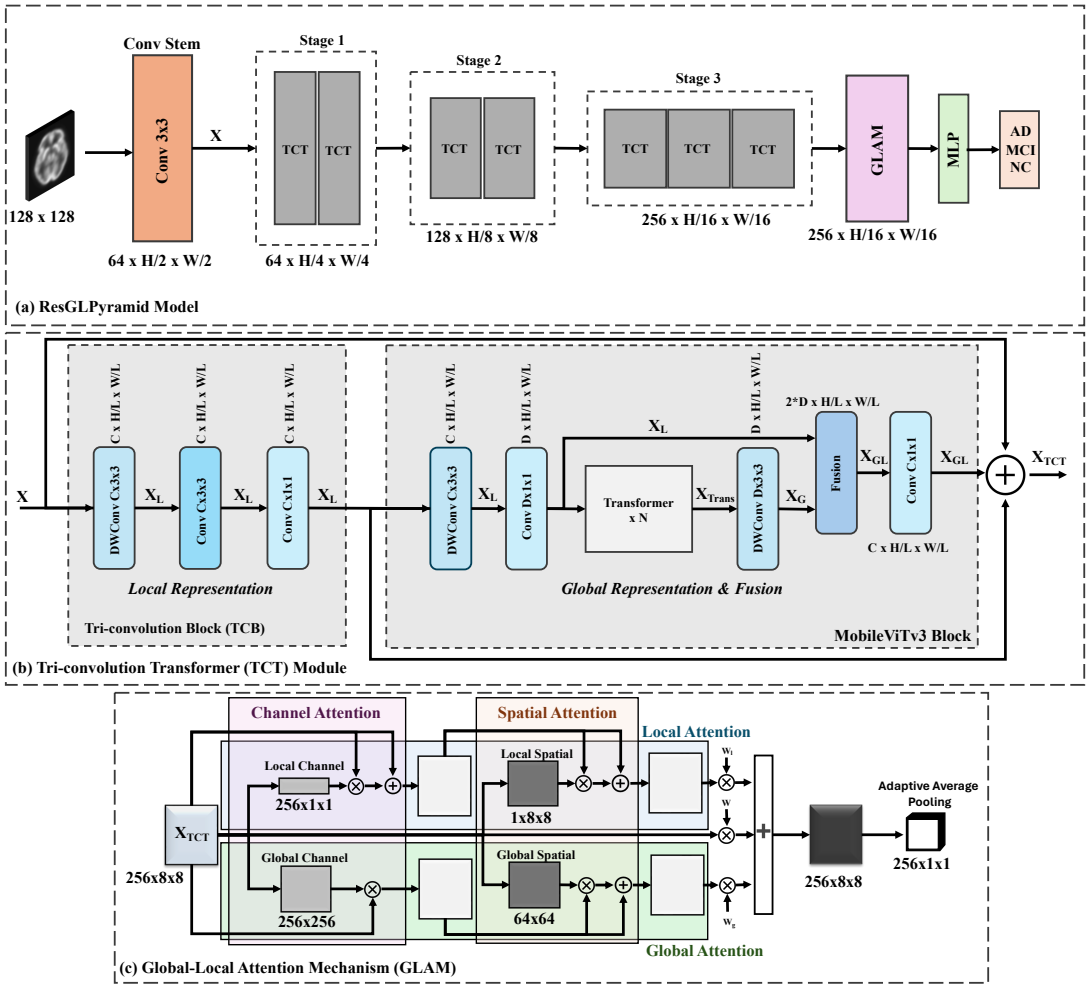

FIGURE 1. (a) Overview of the Proposed ResGLPyramid Model's three stages, each with TCT modules. At the end of the third stage, the GLAM module is integrated to apply channel and spatial attention to the deep-layer feature maps. (b) The TCT module, consists of tri-convolution block and MobileViTv3 for extracting local and global contextual information and fusing them. (c) All four attentions are applied to an input feature map at 256x8x8 to extract more features related to the metabolic activities across the brain.

diverse diagnostic and treatment scenarios.

The architectural design of the ResGLPyramid model is well-suited to handling low-resolution PET images, providing detailed and accurate diagnostic insights that are essential in clinical environments where high fidelity in image analysis is crucial. The proposed model enhances diagnostic accuracy for AD and MCI as well as offers a scalable and efficient solution adaptable to similar challenges across various medical imaging fields by holding significant potential for broad application in clinical practice, contributing to better patient outcomes through earlier and more precise diagnosis.

The rest of this paper is organized as follows: Section II presents the methodology of the proposed ResGLPyramid, Section III analyzes and discusses the experimental results, and Section IV provides concluding remarks.

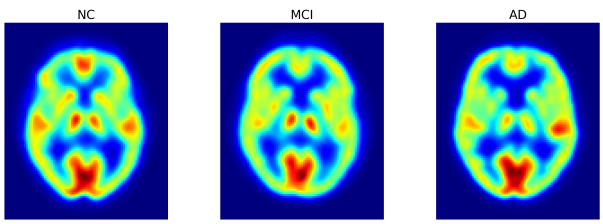

FIGURE 2. Random brain 18F-FDG PET random patient slices of NC, MCI, and AD.

II. PROPOSED METHODOLOGY

The Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset (adni.loni.usc.edu) collected data from 720 subjects using 18F-FDG PET imaging, which included 212 subjects diagnosed with AD, 290 with MCI, and 218 with NC.

TABLE 1. Demographics of the subjects in this study.

Dataset/ Modality	Class	Cases (Male/Female)	Age (Mean ± Std)	MMSE (Mean ± Std)
ADNI/ PET	NC	130/88	73.5 ± 8.1	28.9 ± 1.3
	MCI	165/125	74.3 ± 7.8	27.2 ± 2.3
	AD	122/90	72.2 ± 9.3	23.0 ± 2.0

Each subject underwent a 30-minute dynamic 3D PET scan for six five-minute frames, initiated 30-60 minutes after an intravenous injection of $185\,\text{MBq} \pm 10\%$ of 18F-FDG. Random sample images from each class displayed in Figure 2 highlights the distinct imaging characteristics observed in the stages of cognitive decline. The preprocessing of these images was precisely performed using the Statistical Parametric Mapping tool SPM12 [27], and involved spatial normalization to the Montreal Neurological Institute (MNI) template, having a dimensions of $91 \times 109 \times 91$ with a voxel size of $2 \times 2 \times 2\,\text{mm}^3$, intensity normalization based on the global mean, and skull stripping with a PET mask to isolate brain tissue, followed by smoothing with a Gaussian filter of 8 mm full width at half maximum (FWHM) [28]. This rigorous standardization facilitates a uniform analysis framework crucial for subsequent diagnostic evaluations detailed in the demographic data and Mini-Mental State Examination (MMSE) scores in Table 1.

To further improve neuroimaging accuracy, we developed the ResGLPyramid model, shown in figure 1(a), which employs the tri-convolution transformer (TCT) module with a residual connection to effectively extract and integrate local and global features throughout the analysis. The GLAM enhances this process by refining neuroimaging data using attention mechanisms, culminating in comprehensive diagnostic output through a multilayer perceptron (MLP) layer, as seen in Figure 2. This integrated approach ensures precise and reliable diagnostic predictions, advancing the field of neuroimaging in detecting cognitive decline. The details of each component are as follows.

A. CONV STEM

The purpose of Conv Stem is to reduce the computational load, highlight essential features, prevent overfitting, and detect features at multiple scales [29]. Given the properties of the image, we selected a small kernel size of 3×3 for convolution, with a stride of 2 and padding of 1. The number of output channels, denoted as C, is set to 64. This initial step is particularly effective in processing neuroimages, which often contain redundant information, thus enabling the extraction of relevant features. Subsequently, batch normalization (BN) is applied, followed by a 3×3 max pooling operation to produce $X \in \mathbb{R}^{C \times \frac{H}{2} \times \frac{W}{2}}$, where X represents the output features from this block, H is the height and W is the width of the input image. These features are then fed into a three-stage network comprising a TCT module for further processing.

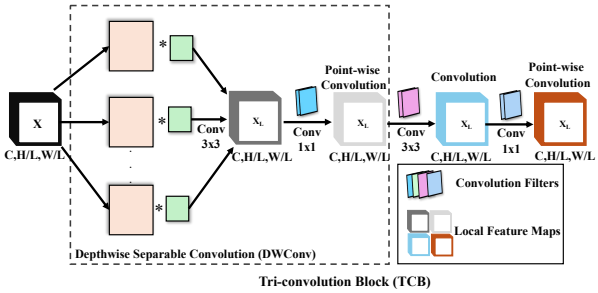

FIGURE 3. An illustration of the TCB block that consists of depth-wise separabale, simple and point-wise convolution for local spatial features extraction.

B. THE PROPOSED TCT MODULE

The tri-convolution transformer (TCT) module as shown in figure 1(b), integrates three convolution blocks to encode local information alongside patch-based self-attention mechanisms for capturing global context and fusing them. This synthesis is essential in a clinical environment for accurate diagnosis of AD [24], because it provides a nuanced view of both specific regions of interest (ROIs) and broader brain structures, including cerebrospinal fluid and the hippocampus. The TCT module enhances this process by incorporating residual connections that facilitate information flow via skip paths, thus improving integration efficiency. Structurally, the network is designed with three stages: the first two stages contain two TCT modules each, while the final stage includes three TCT modules. This configuration produces feature maps with increasing channel dimensions C of 64, 128, and 256 across the stages, allowing detailed and comprehensive extraction and integration of both local and global information critical for AD diagnosis.

1) LOCAL FEATURE EXTRACTOR

The clinical heterogeneity of AD and the different absorption rates cause PET scans to vary among individuals which complicates diagnosis. We apply the TCB, as shown in Figure 3. to capture range of features from different regions of the brain, enhancing detection of subtle pathological changes. This helps to minimize data redundancy and prioritizes critical diagnostic features, leading to precise and efficient analysis. The module processes features with dimensions $C \times \frac{H}{L} \times \frac{W}{L}$, and L values of 4, 8, and 16. The processing begins with depth-wise separable convolution (DWConv), followed by a 3×3 kernel convolution, and inference with point-wise convolution using a 1×1 kernel. The first two stages of this sequence incorporate batch normalization and the Gaussian error linear unit (GELU) activation function to enhance feature integration, as detailed in Equations 1-3. After the final convolution, batch normalization is applied in isolation, setting the stage for the subsequent attention mechanism. The output from this attention block is then seamlessly integrated with residual output, optimizing the

feature processing workflow for detailed and effective analysis.

$$X_L = \text{Gelu}\left(\text{BN}\left(\text{DWConv}_{3\times3}(X)\right)\right) \tag{1}$$

$$X_L = \text{Gelu}\left(\text{BN}\left(\text{Conv}_{3\times3}(X_L)\right)\right) \tag{2}$$

$$X_L = \text{BN}\left(\text{Conv}_{1\times1}(X_L)\right) \tag{3}$$

where X_L is the local feature map.

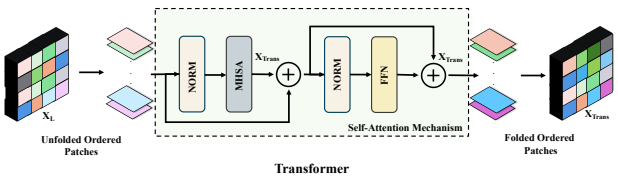

FIGURE 4. The transformer block first inputs features that are unfolded then employs normalization (NORM), multi-head self-attention (MHSA), and a feed forward network (FFN) for global representation extraction that is folded back into the same dimension as the input

2) GLOBAL FEATURE EXTRACTOR AND FUSION

We utilize the MobileViTv3 [30] module to employ spatial inductive bias to consistently capture long-range feature dependencies across spatial pixels and patch order, as depicted in Figure 2(b). That figure illustrates the process of global information extraction and fusion. Locally extracted features from the TCB undergo convolution through depth-wise 3×3 and point-wise 1×1 convolutions, resulting in output $X_{\text{L}} \in \mathbb{R}^{D\times\frac{H}{L}\times\frac{W}{L}}$ with channel dimension D and spatial dimension H. These features are then fed into a transformer block, detailed in Figure 4., where X_{L} is segmented into N-ordered, non-overlapping patches p, reshaping the dimension to $D \times p \times N$, with p representing the area of each patch (wh), and N, the total number of patches $(\frac{WH}{p})$. Each patch, with height h and width w not exceeding the kernel size k, passes through normalization (Norm), multi-head self attention (MHSA), and feed-forward network (FFN) layers. After processing, these patches are reassembled back into the original dimension to produce X_{Trans}, which is further processed through depth-wise convolution to generate X_{G}. Extracted features X_L and X_{G} are then concatenated during the fusion step and subsequently processed with point-wise 1×1 convolution to produce X_{TCT}, with dimensions $\mathbb{R}^{C\times\frac{H}{L}\times\frac{W}{L}}$, where X_{TCT}, X_{Trans}, and X_{G} denote the outputs from their respective processes. The sequence of operations from capturing long-range dependencies to final output of the feature maps is outlined in equations 4 to 11, highlighting the comprehensive fusion of local and global information within the TCT module.

$$X_L = \text{DWConv}_{3\times3}(X_F) \tag{4}$$

$$X_L = \text{Conv}_{1\times1}(X_L) \tag{5}$$

$$X_{Trans} = \text{MHSA}(\text{LN}(X_L(p)) + X_{L(p)} \tag{6}$$

$$X_{Trans} = \text{FFN}(\text{LN}(X_{Trans(p)})) + X_{Trans(p)} \tag{7}$$

$$X_G = \text{DWConv}_{3\times3}(X_{Trans}) \tag{8}$$

$$X_{GL} = \text{Fusion}(X_{\text{G}}, X_{\text{L}}) \tag{9}$$

$$X_{TCT} = \text{Gelu}\left(X_{GL} + X_L + X\right) \tag{10}$$

C. THE GLOBAL LOCAL ATTENTION MODULE (GLAM)

We incorporated the GLAM [31] into our network to enhance representation learning and enrich embedding by applying all attention i.e., local channel attention, global channel attention, local spatial attention, and global spatial attention, as shown in Figure 1(c). The module captures the subtle early signs by identifying hypometabolism in specific brain regions and contextualizing these findings within the overall brain metabolism. Initially applied in image retrieval challenges, the GLAM module shows significant improvements by employing a comprehensive approach to attention, utilizing a weighted vector to optimize the attention process. In our implementation, we adapted the traditional global average pooling to adaptive average pooling for its effectiveness in handling pooling tasks. Features from X_{TCT} with a dimension of $256 \times \frac{H}{16} \times \frac{W}{16}$ are the input for the GLAM block, where local, global, and incoming features are equally weighted by using w_l,w_g and w before fusion. This integration allows selective enhancement of the most informative features, which are subsequently refined through adaptive average pooling using a 1×1 kernel, achieving superior feature extraction and pooling outcomes.

In the end, a two-layer multilayer perceptron (MLP) is utilized to perform binary classification. The extracted X_{TCT} features are flattened and fed into the MLP, which comprises dense layers with 128 and 2 hidden units, respectively. Each layer is equipped with batch normalization and a GELU activation function.

D. LOSS FUNCTION

Label smoothing is implemented to prevent overfitting and to address the issue of redundant information across slices [32], which helps the model avoid overconfidence in its predictions. Our model employs the SCE loss function, from the PyTorch Image Models library (TIMM) [33]. SCE loss, for a batch of predictions x and the corresponding true labels y, is evaluated by obtaining the log probabilities (*LP*) from applying the log softmax function to the logits produced by the model. Following this, the negative log likelihood (NLL) loss is computed for the actual class labels.

$$\text{LP}_i = \log(p(x)_i) \tag{11}$$

$$\text{NLLLoss} = -\frac{1}{N}\sum_{i=1}^{N}\log(p(x)_{y_i}) \tag{12}$$

where $p(x)_i = \frac{e^{x_i}}{\sum_{j=1}^{K} e^{x_j}}$, $p(x)_i$ is the probability of the i-th element and K is the number of classes. N is the number

of instances in the batch, and y_i denotes the true class for the i-th instance.

To further refine the model's predictions, label smoothing is applied by averaging the log probabilities across all classes. This ensures that the probability mass is evenly distributed among all classes, contributing to a more balanced learning process.

$$\text{SmoothLoss} = \frac{1}{N}\sum_{i=1}^{N} LP_i \tag{13}$$

The final SCE loss is computed as a convex combination of the NLL loss and Smooth Loss, with smoothing parameter α adjusting the balance between these two components, enhancing the model's generalization capabilities.

$$\text{SCELoss} = (1-\alpha)\cdot \text{NLLLoss} + \alpha \cdot \text{SmoothLoss} \tag{14}$$

where $\alpha = 0.2$ is the smoothing factor that balances the trade-off between adhering to the true class distribution and promoting probability distribution uniformity across all classes.

III. EXPERIMENTS AND DISCUSSION

In this section, we discuss the experimental environment, evaluation metrics, and experimental results obtained with the proposed model, including a visualization of the results and an ablation study to underscore the importance of each component.

A. EXPERIMENTAL ENVIRONMENT

Our experiments were conducted on a Windows system equipped with 32 GB of RAM and an Nvidia RTX 3060 Ti GPU, utilizing the PyTorch framework for its proficient handling of neural networks. To optimize performance, we set a batch size of 8 and an initial learning rate of 0.001, adjusted by an exponential learning rate scheduler with a 0.95 decay rate. We also used the Adam optimizer with beta coefficients of 0.9 and 0.99 for weight adjustments. An early stopping mechanism monitored validation loss to mitigate overfitting. The dataset was distributed at a 0.8:0.1:0.1 ratio. The model's reliability was ensured through 10-fold cross-validation, with training limited to 120 epochs to prevent overfitting while allowing adequate learning depth. Data augmentation techniques such as horizontal flipping, zooming, and random rotation were applied to enhance the robustness of the training process.

B. EVALUATION METRICS

The model's performance was assessed using several metrics, such as specificity (Spec) to measure the proportion of true negatives correctly identified, sensitivity (Sen) or True Positive Rate (Tpr) to measure the proportion of true positives correctly identified, F1-score to evaluate the harmonic mean of precision and sensitivity to indicate the balance between them, Accuracy (ACC) to assesses the overall correctness of predictions, and AUC to predict the model's ability to discriminate between classes(values close to 1 indicate excellent discrimination and values close to 0.5 suggest random guessing). These metrics are vital for evaluating the precision and reliability of the model and are defined as follows:

$$\text{ACC} = \frac{TP+TN}{TP+TN+FP+FN} \tag{15}$$

$$\text{Sen} = \frac{TP}{TP+FN} \tag{16}$$

$$\text{Spec} = \frac{TN}{TN+FP} \tag{17}$$

$$\text{Pre} = \frac{TP}{TP+FP} \tag{18}$$

$$\text{False Positive Rate (FPR)} = \frac{FP}{TN+FP} \tag{19}$$

$$\text{F1-Score} = 2 \times \frac{\text{Pre}\times\text{Sen}}{\text{Pre}+\text{Sen}} \tag{20}$$

$$\text{AUC} = \sum_{i=1}^{n-1} \frac{(Fpr_{i+1}-Fpr_i)\cdot(Tpr_i+Tpr_{i+1})}{2} \tag{21}$$

These metrics are calculated based on the counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The AUC is particularly derived from the relationship between the Tpr and the False Positive Rate (Fpr), providing a comprehensive view of model performance across different thresholds and The variable n represents the total number of data points on the Receiver Operating Characteristic (ROC) curve.

C. EXPERIMENT RESULTS AND COMPARISON

The receiver operating characteristic (ROC) curves for our proposed model, illustrated in Figure 5. showcase its performance across three binary classifications: AD vs. NC, AD vs. MCI, and MCI vs. NC. The highest AUC recorded was 96.90% for the AD vs. NC classification, with the AUC for AD vs. MCI and MCI vs. NC at 92.48% and 93.08%, respectively. Notably, the ROC curve for AD vs. MCI showed a steep rise between the false positive rates of 0.1 and 0.6, highlighting the model's sensitivity in distinguishing between AD and MCI, thus yielding lower test errors. Overall, the model effectively differentiated between the most challenging cases.

In comparing our results with SOTA methods in Table 2., our model demonstrated superior performance in classifying AD vs. MCI, AD vs. NC, and MCI vs. NC using deep neural network models, encompassing both 3D and 2D approaches. While a direct comparison is limited by the variance in subject data across studies, our evaluation employed 10-fold cross-validation to ensure robustness. Previous 3D network studies by Song et al. [17] and Gao et al. [35] reported lower accuracies (92.1% and 88.8% respectively) for AD vs. NC due to redundant information in 3D brain data volumes and the high computational costs associated with training these networks. Studies employing 2D slice images, however, such

TABLE 2. Comparison of the proposed model with other DL models on the ADNI database.

Reference	Year	Subjects			Performance Metrics (%)				
		AD	MCI	NC	Sensitivity	Specificity	Accuracy	AUC	F1-Score
Pan et al. [34]	2020	237	-	242	90.32	95.49	93.13	97.11	–
Kim et al. [16]	2020	141	–	348	87.93	93.57	91.02	–	–
Zhang et al. [15]	2019	91	–	101	96.58	95.39	98.47	98.61	–
Liu et al. [11]	2018	93	–	100	91.40	91.00	91.20	95.30	–
Song et al. [17]	2021	95	–	126	89.13	94.27	92.10	–	–
Gao et al. [35]	2021	196	–	227	86.10	90.80	88.80	94.30	–
Tuan et al. [36]	2022	323	–	497	90.04	93.54	91.83	96.84	–
Chen et al. [18]	2023	124	–	212	**97.18**	96.29	**98.54**	**98.76**	–
Our proposed	–	**218**	**–**	**212**	96.10	**97.50**	96.90	96.90	**96.20**
Liu et al. [11]	2018	–	146	100	78.10	80.00	78.90	83.90	–
Zhang et al. [15]	2019	–	200	101	84.54	75.36	84.62	87.44	–
Song et al. [17]	2021	–	160	126	72.81	70.56	72.00	–	–
Chen et al. [18]	2023	–	192	84	93.13	91.66	**93.56**	**94.70**	–
Our proposed	–	**–**	**290**	**212**	**93.80**	**92.47**	93.08	93.13	**92.61**
Zhang et al. [15]	2019	91	200	–	**94.97**	79.24	84.20	87.92	–
Song et al. [17]	2021	95	160	–	57.46	89.69	78.30	–	–
Chen et al. [18]	2023	124	192	–	90.59	89.86	89.31	90.69	–
Our proposed	–	**212**	**290**	**–**	90.80	**94.14**	**92.75**	**92.48**	**91.28**

as those by Kim et al. [16], Zhang et al. [15], Liu et al. [11], Yonglin et al. [18], and Pan et al. [34] listed in Table 2., provided more promising results, with improvements in sensitivity, specificity, and accuracy in all cases, particularly for difficult classifications like MCI vs. NC and AD vs. MCI.

In this study, we propose a 2D-slice AD neuroimage prediction model that synergizes local and global features to refine early AD diagnosis. The model's architecture, combining convolutional operations' inductive bias with self-attention mechanisms on patches, markedly enhanced performance, especially in the complex cases AD vs. MCI and MCI vs. NC. Notably, our model showed accuracy and AUC improvements (3.44% and 1.79%, respectively, for AD vs. MCI) as well as sensitivity and specificity improvements (0.67% and 0.81%, respectively) for MCI vs. NC in these difficult classifications, underscoring the value of integrating local and global information to effectively capture brain metabolic activities. Although the AUC for AD vs. NC was slightly lower by 1.86% compared to other methodologies, our model required less data owing to the generalization capabilities of CNNs, suggesting the potential for further improvements with more diverse training datasets.

D. MODEL INTERPRETATION AND VISUALIZATION

1) FEATURE VISUALIZATION

t-distributed stochastic neighbor embedding (t-SNE) [37] is a non-linear dimensionality reduction technique employed to visualize high-dimensional features in low-dimensional spaces. This method, illustrated in Figure 6., arranges similar data points close to each other while positioning dissimilar ones at a distance. It facilitates visualization of the feature distribution of network predictions to assess the effectiveness of the network in learning PET lesion features. Figure 6(c) illustrates AD vs. MCI clusters in close proximity, with some points existing within the opposing cluster. This minimal variation between the two class features complicates the model's predictive accuracy. Such a difficulty is mirrored in clinical findings, emphasizing the challenge of distinguishing between these conditions. For the classification of AD vs. NC, Figure 6(a) reveals optimally aggregated features, indicating the model's high sensitivity and precision. Meanwhile, Figure 6(b) displays well-defined clusters for MCI vs. NC, demonstrating the model's effectiveness in discriminating between these closely related cases with very few data points positioned within the alternate cluster.

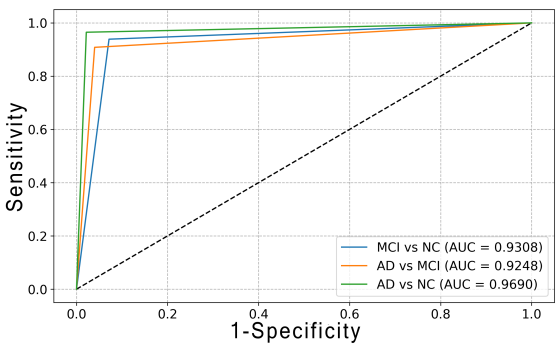

FIGURE 5. ROC curves for the three binary classifications in this paper.

2) MODEL INTERPRETATION

The class activation map is employed in our proposed model to identify the regions of focus by following the methodology outlined in [12]. Figure 7(a) presents heatmaps images of AD slices, and figures 7(b), (c), and (d) illustrate the progression from initial convolution features to deeper layer feature maps. These colormaps, generated using the JET algorithm, clearly highlight the model's concentration on critical brain areas for diagnosing AD, such as the posterior temporal lobe, posterior cingulate cortex, hippocampus, thalamus, parahippocampal

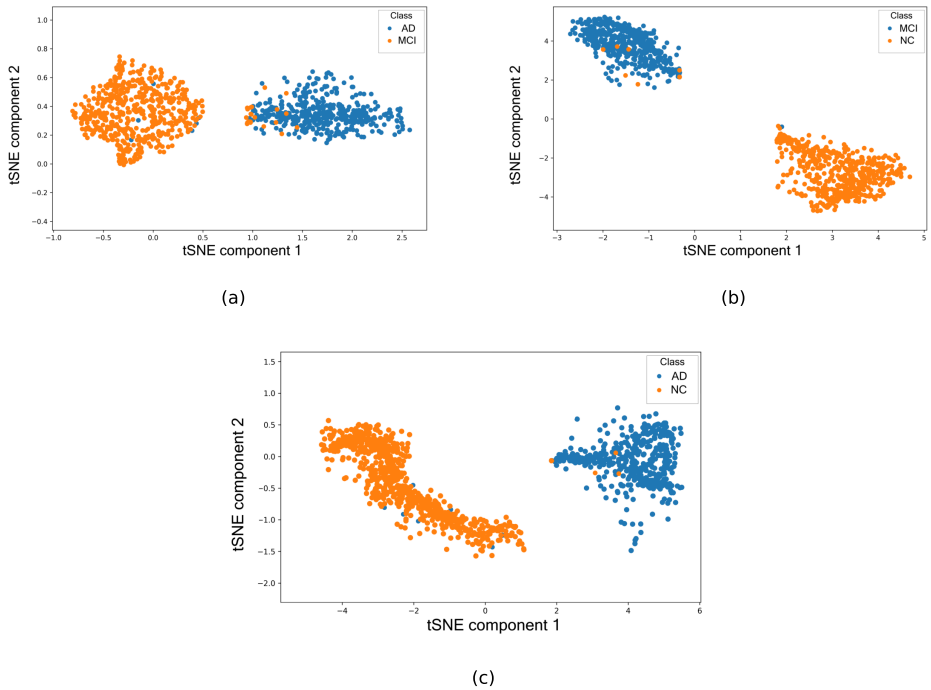

FIGURE 6. t-SNE visualizations for different classes: (a) AD vs. NC, (b) MCI vs. NC, (c) AD vs. MCI.

TABLE 3. The ablation Study of w/o SCE, w/o global features, and w/o the GLAM.

Binary Classification	Proposed Model	Sen(%)	Spec(%)	ACC(%)	AUC(%)	F1-Score(%)
AD vs. NC	TCB	70.83	77.26	74.30	74.00	70.90
	TCB+SCE	79.10	82.10	80.80	80.50	78.31
	TCT+SCE	89.81	93.22	91.70	91.32	90.20
	TCT+GLAM+SCE	**96.10**	**97.50**	**97.00**	**96.90**	**96.20**
AD vs. MCI	TCB	68.90	74.50	71.90	71.40	70.10
	TCB+SCE	75.40	80.60	78.30	78.00	76.60
	TCT+SCE	85.70	90.20	88.10	87.80	86.80
	TCT+GLAM+SCE	**90.81**	**94.14**	**92.75**	**92.48**	**91.28**
MCI vs. NC	TCB	70.12	75.433	72.98	72.67	70.59
	TCB+SCE	76.38	81.27	79.21	78.90	77.98
	TCT+SCE	87.56	89.65	88.89	88.58	87.40
	TCT+GLAM+SCE	**93.80**	**92.47**	**93.08**	**93.13**	**92.61**

gyrus, and supramarginal gyrus. The maps reveal a sharpened focus on these regions as the model analyzes deeper layers. Figure 7(e) displays heatmaps images from an MCI subject, and figures 7(f), (g), and (h) show visualizations of the deeper layers. Initially, broader regions were targeted compared to the AD subjects, but as the depth increases, the focus narrows distinctly to areas differentiating MCI from AD. These include the superior parietal regions, angular gyrus, right superior frontal gyrus, precuneus regions, marginal sulcus, and bilateral postcentral regions. This focused sensitivity aligns with clinical findings, confirming that results are both visualizable and interpretable, ensuring the model's relevance and applicability in clinical environments.

E. ABLATION STUDIES OF THE PROPOSED MODEL

In this section, we detail the ablation studies conducted to evaluate the individual contributions of different components within our proposed model. Initially, the model was trained using only triconvolutional blocks (TCB) with traditional cross-entropy (CE) loss to establish a baseline. To mitigate the issue of redundant information between slices and to curb overfitting, we incorporated SCE loss. Subsequently, global

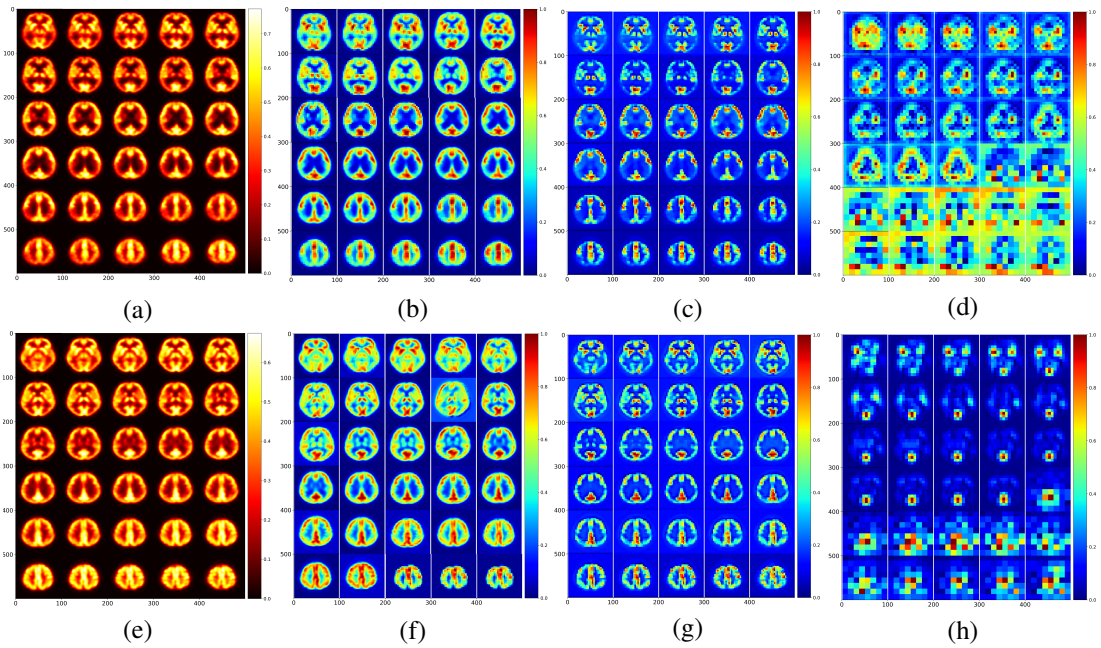

FIGURE 7. From AD slices (a) shows heatmaps images, whereas (b), (c), and (d) show colormaps of deep-layer features; from MCI slices (e) shows heatmaps images, whereas (f), (g), and (h) show colormaps from deep layers of the proposed model.

features were derived from locally encoded features using a transformer, enhancing the model's performance. The final integration involved the GLAM at the end of the backbone network to extract and refine features into $256 \times 1 \times 1$ vectors, which were then processed through the MLP layer.

1) ABLATION EXPERIMENTS

- **Effectiveness of the TCB and SCE Loss:** Table 3. shows the results for each model component. Incorporating TCB with SCE loss improved the AUC by 6.5% for AD vs. NC, by 6.7% for AD vs. MCI, and by 6.2% for MCI vs. NC. This enhancement indicates the TCB effectively generalizes features, reducing the tendency to overfit.
- **Global Feature Integration:** Further enhancements were observed when global features were extracted and fused with local features, capturing long-range dependencies. This step notably increased the AUC by 11% for AD vs. NC, 9.8% for AD vs. MCI, and 9.5% for MCI vs. NC. The integration of global features is pivotal for diagnosing early stages of AD, demonstrating the critical role of comprehensive feature analysis in capturing brain metabolic activity.
- **The GLAM Contribution:** The GLAM applied to features from the backbone network, significantly extracted relevant information through focused attention mechanisms and adaptive average pooling, achieving feature dimensions at $256 \times 1 \times 1$. This enhancement led to a 5.58% increase in AUC for AD vs. NC and further gains of 4.68% and 4.55% for AD vs. MCI and MCI vs. NC, respectively. The results underscore the importance of last-stage features in diagnosing early-stage MCI of AD.

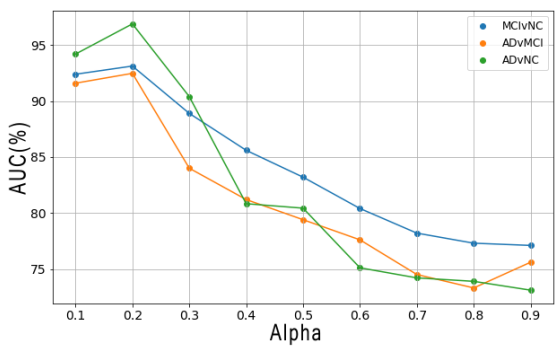

FIGURE 8. AUC Results vs. alpha for smoothing in each binary case.

2) EFFECT OF PARAMETER ALPHA (α)

Figure 8. illustrates how varying the α parameter in the SCE function affects the AUC results across different binary classifications. Increasing α values generally led to a degradation of results in a non-linear manner. For AD vs. NC, significant decreases highlight the distinct differences between these conditions, aligning with clinical findings. However, the variations for AD vs. MCI and MCI vs. NC were less pronounced, with optimal results achieved at an alpha of 0.2, emphasizing the sensitivity of the model to the parameter settings.

These ablation studies validate the effectiveness of each integrated component and parameter adjustment in our model, illustrating their collective impact on enhancing the diagnostic accuracy for Alzheimer's disease and its precursors.

IV. CONCLUSION

In this paper, we introduced the ResGLPyramid model, which is designed to diagnose early stages of Alzheimer's disease using 18F-FDG PET. The model incorporates TCT modules, including convolution operations, MobileViTv3, and skip connections, enabling it to capture local region specifics and global contextual information on metabolic activities across the brain. We employed softened cross entropy to mitigate overfitting and enhance the model's generalizability. Experimental results show the ResGLPyramid model achieved an accuracy of 92.75% in the classification of MCI and AD individuals, which is higher compared to the SOTA method. In the future, we will further refine the model's ability to distinguish among the three diagnostic classes by enhancing its feature extraction capabilities. Additionally, we will boost AD detection rates and explore the application of this approach across different imaging modalities. This future work will enhance the diagnostic precision and utility of the ResGLPyramid model in a clinical environment.

V. ACKNOWLEDGMENT

The data for this paper was sourced from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu). Although ADNI investigators contributed to the design and data provision of ADNI, they were not involved in the analysis or the writing of this paper. A complete list of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

REFERENCES

[1] 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, pp. 939–939, 1984.

[2] A. D. International, "World alzheimer report 2018. the state of the art of dementia research: new frontiers," Alzheimer's Disease International, p. 2018, 2018.

[3] Y.-X. Chen, N. Liang, X.-L. Li, S.-H. Yang, Y.-P. Wang, and N.-N. Shi, "Diagnosis and treatment for mild cognitive impairment: a systematic review of clinical practice guidelines and consensus statements," Frontiers in Neurology, vol. 12, p. 719849, 2021.

[4] S. Liu, A. V. Masurkar, H. Rusinek, J. Chen, B. Zhang, W. Zhu, C. Fernandez-Granda, and N. Razavian, "Generalizable deep learning model for early alzheimer's disease detection from structural mris," Scientific reports, vol. 12, no. 1, p. 17106, 2022.

[5] A. J. Chang, R. Roth, E. Bougioukli, T. Ruber, S. S. Keller, D. L. Drane, R. E. Gross, J. Welsh, A. Abrol, V. Calhoun et al., "Mri-based deep learning can discriminate between temporal lobe epilepsy, alzheimer's disease, and healthy controls," Communications Medicine, vol. 3, no. 1, p. 33, 2023.

[6] K. Ote, F. Hashimoto, Y. Onishi, T. Isobe, and Y. Ouchi, "List-mode pet image reconstruction using deep image prior," IEEE Transactions on Medical Imaging, 2023.

[7] N. Smailagic, L. Lafortune, S. Kelly, C. Hyde, and C. Brayne, "18f-fdg pet for prediction of conversion to alzheimer's disease dementia in people with mild cognitive impairment: an updated systematic review of test accuracy," Journal of Alzheimer's Disease, vol. 64, no. 4, pp. 1175–1194, 2018.

[8] K. A. Johnson, N. C. Fox, R. A. Sperling, and W. E. Klunk, "Brain imaging in alzheimer disease," Cold Spring Harbor perspectives in medicine, vol. 2, no. 4, p. a006213, 2012.

[9] C.-C. Liu and J. Qi, "Higher snr pet image prediction using a deep learning model and mri image," Physics in Medicine & Biology, vol. 64, no. 11, p. 115004, 2019.

[10] K. T. Chen, T. N. Toueg, M. E. I. Koran, G. Davidzon, M. Zeineh, D. Holley, H. Gandhi, K. Halbert, A. Boumis, G. Kennedy et al., "True ultra-low-dose amyloid pet/mri enhanced with deep learning for clinical interpretation," European journal of nuclear medicine and molecular imaging, vol. 48, pp. 2416–2425, 2021.

[11] M. Liu, D. Cheng, W. Yan, and A. D. N. Initiative, "Classification of alzheimer's disease by combination of convolutional and recurrent neural networks using fdg-pet images," Frontiers in neuroinformatics, vol. 12, p. 35, 2018.

[12] Y. Ding, J. H. Sohn, M. G. Kawczynski, H. Trivedi, R. Harnish, N. W. Jenkins, D. Lituiev, T. P. Copeland, M. S. Aboian, C. Mari Aparici et al., "A deep learning model to predict a diagnosis of alzheimer disease by using 18f-fdg pet of the brain," Radiology, vol. 290, no. 2, pp. 456–464, 2019.

[13] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.

[15] F. Zhang, Z. Li, B. Zhang, H. Du, B. Wang, and X. Zhang, "Multi-modal deep learning model for auxiliary diagnosis of alzheimer's disease," Neurocomputing, vol. 361, pp. 185–195, 2019.

[16] H. W. Kim, H. E. Lee, K. Oh, S. Lee, M. Yun, and S. K. Yoo, "Multi-slice representational learning of convolutional neural network for alzheimer's disease classification using positron emission tomography," BioMedical Engineering OnLine, vol. 19, no. 1, pp. 1–15, 2020.

[17] J. Song, J. Zheng, P. Li, X. Lu, G. Zhu, and P. Shen, "An effective multimodal image fusion method using mri and pet for alzheimer's disease diagnosis," Frontiers in digital health, vol. 3, p. 637386, 2021.

[18] Y. Chen, H. Wang, G. Zhang, X. Liu, W. Huang, X. Han, X. Li, M. Martin, and L. Tao, "Contrastive learning for prediction of alzheimer's disease using brain 18f-fdg pet," IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 4, pp. 1735–1746, 2022.

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[20] E. Jun, S. Jeong, D.-W. Heo, and H.-I. Suk, "Medical transformer: Universal brain encoder for 3d mri analysis," arXiv preprint arXiv:2104.13633, 2021.

[21] U. Khatri and G.-R. Kwon, "Explainable vision transformer with self-supervised learning to predict alzheimer's disease progression using 18f-fdg pet," Bioengineering, vol. 10, no. 10, p. 1225, 2023.

[22] C. Li, Y. Cui, N. Luo, Y. Liu, P. Bourgeat, J. Fripp, and T. Jiang, "Transresnet: Integrating transformers and cnns for alzheimer's disease classification," in 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI). IEEE, 2022, pp. 1–5.

[23] J. Jang and D. 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, 2022, pp. 20 718–20 729.

[24] Z. Hu, Z. Wang, Y. Jin, and W. Hou, "Vgg-tswinformer: Transformer-based deep learning model for early alzheimer's disease prediction," Computer Methods and Programs in Biomedicine, vol. 229, p. 107291, 2023.

[25] S. Mehta and M. Rastegari, "Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer," arXiv preprint arXiv:2110.02178, 2021.

[26] Y. Li, J. Tang, L. Li, X. Wang, W. Ding, X. Li, T. Yu, and X. Wu, "Mobilevit-based classification of alzheimer's disease," in 2023 IEEE 6th International Conference on Pattern Recognition and Artificial Intelligence (PRAI), 2023, pp. 443–448.

[27] Functional Imaging Laboratory, UCL, "Statistical parametric mapping (spm) software," http://www.fil.ion.ucl.ac.uk/spm/software/.

[28] K. J. Worsley, S. Marrett, P. Neelin, and A. Evans, "Searching scale space for activation in pet images," Human brain mapping, vol. 4, no. 1, pp. 74–90, 1996.
[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[30] S. N. Wadekar and A. Chaurasia, "Mobilevitv3: Mobile-friendly vision transformer with simple and effective fusion of local, global and input features," arXiv preprint arXiv:2209.15159, 2022.
[31] C. H. Song, H. J. Han, and Y. Avrithis, "All the attention you need: Global-local, spatial-channel attention for image retrieval," in Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2022, pp. 2754–2763.
[32] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?" Advances in neural information processing systems, vol. 32, 2019.
[33] R. Wightman, "timm: PyTorch image models," https://github.com/rwightman/pytorch-image-models, 2020.
[34] 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 Transactions on Medical Imaging, vol. 40, no. 1, pp. 81–92, 2021.
[35] X. Gao, F. Shi, D. Shen, and M. Liu, "Task-induced pyramid and attention gan for multimodal brain image imputation and classification in alzheimer's disease," IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 1, pp. 36–43, 2022.
[36] P. M. Tuan, T.-L. Phan, M. Adel, S. Bourennane, N. L. Trung, and E. Guedj, "C-atlas: A brain mapping based on fdg-pet images for alzheimer's disease diagnosis," in 2022 RIVF international conference on computing and communication technologies (RIVF). IEEE, 2022, pp. 150–155.
[37] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." Journal of machine learning research, vol. 9, no. 11, 2008.

ABDUL REHMAN received a B.Sc Degree in Electronics Engineering from the University of Engineering and Technology, Taxila, Pakistan. He is currently pursuing his master's degree in IT convergence engineering from Gachon University, South Korea. His research interests include computer vision, deep learning, and image processing.

MYUNG-KYU YI (Member, IEEE) received a Ph.D. in Computer Science and Engineering from Korea University in 2005. He is currently a Research Professor with Gachon University, and is a member of the personal health information standardization task force of the TTA U-Health project group. His research interests include healthcare, security, machine learning, deep learning, and human activity recognition.

ABDUL MAJEED (Member, IEEE) received a B.S. degree in Information Technology from the UIIT, PMAS-UAAR, Rawalpindi, Pakistan, in 2013, an M.S. degree in Information Security from the COMSATS University, Islamabad, Pakistan, in 2016, and a Ph.D. degree in Computer Information Systems & Networks from the Korea Aerospace University, Korea, in 2021. He worked as a Security Analyst with Trillium Information Security Systems (TISS), Rawalpindi, Pakistan, from 2015 to 2016. He is currently working as an Assistant Professor with the Department of Computer Engineering, Gachon University, Korea. His research interests include privacy-preserving data publishing, statistical disclosure control, privacy-aware analytics, data-centric artificial intelligence, and machine learning.

SEONG OUN HWANG (Senior Member, IEEE) received a B.S. in mathematics from Seoul National University in 1993, an M.S. in information and communications engineering from the Pohang University of Science and Technology in 1998, and a Ph.D. in computer science from the Korea Advanced Institute of Science and Technology in 2004. He worked as a Software Engineer for LG-CNS Systems, Inc. from 1994 to 1996. He also worked as a Senior Researcher with the Electronics and Telecommunications Research Institute (ETRI) from 1998 to 2007. He was a Professor with the Department of Software and Communications Engineering, Hongik University, from 2008 to 2019. He is currently a Professor with the Department of Computer Engineering, Gachon University. His research interests include cryptography, cybersecurity, and artificial intelligence. He is an Editor of ETRI Journal.