Biomedical Signal Processing and Control 101 (2025) 107263

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

An efficient vision transformer for Alzheimer's disease classification using magnetic resonance images

Si-Yuan Lu [a], Yu-Dong Zhang [b,c,*], Yu-Dong Yao [d,**]

[a] School of Communications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
[b] School of Computing and Mathematical Sciences, University of Leicester, Leicester LE1 7RH, UK
[c] Department of Information Technology, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[d] Department of Electrical and Computer Engineering, Stevens Institute of Technology, Hoboken, NJ 07030, USA

ARTICLE INFO

Keywords:
Computer-aided diagnosis
Alzheimer's disease
Magnetic resonance imaging
Vision transformer
Token compression

ABSTRACT

Alzheimer's disease (AD) is the most common dementia that is often seen among the elderly. AD can cause the loss of cognitive ability and memory, which can result in death as AD is progressive. The exact cause of AD is still in research, but it is believed to be related to genes, diet, and environment. One observation of AD is the shrinkage of the hippocampus and frontal lobe cortex. Magnetic resonance imaging (MRI) is often employed in the diagnosis of AD as it can produce clear images of the soft tissues. In this study, a new computer-aided diagnosis (CAD) method named RanCom-ViT, is proposed to interpret the brain MRI slices automatically and precisely for AD diagnosis with better global representation learning and efficiency. A pre-trained vision transformer (ViT) is chosen as the backbone because ViTs with attention modules can achieve better performance than convolutional neural networks on larger datasets. Then, a novel token compression block is proposed to improve the efficiency of the RanCom-ViT by removing the less important tokens. Further, the classification head of the RanCom-ViT is enhanced by a random vector functional-link structure to obtain better classification performance in AD diagnosis. A large public brain MRI dataset is utilized in the evaluation experiments of the proposed RanCom-ViT, and it achieved an overall accuracy of 99.54% with a double throughput than the benchmark model. The results reveal that the RanCom-ViT outperforms several existing state-of-the-art AD diagnosis methods in terms of accuracy, and the token compression method contributes to higher efficiency.

1. Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disorder that primarily affects the brain, resulting in memory loss, cognitive decline, and behavioral changes. It is the most common form of dementia, accounting for around 60–80 % of cases [1]. The symptoms of AD typically begin with mild forgetfulness and difficulty in remembering recent events or information. As the disease progresses, individuals may experience confusion, disorientation, language problems, mood swings, and difficulty in performing daily tasks. In the advanced stages, AD can severely impair memory, judgment, and reasoning, leading to a loss of independence. There is currently no cure for AD, and treatment options focus on managing symptoms and slowing down the progression of the disease. Medications may be prescribed to alleviate symptoms and improve cognitive function, while supportive therapies such as occupational therapy and cognitive training can help individuals maintain independence and quality of life. Early diagnosis and intervention are crucial in managing the disease and providing support for individuals and their families affected by this challenging condition [2].

To diagnose AD, a bunch of evaluation methods are used, including medical history, cognitive assessment, neurological examination, and brain imaging. So far, magnetic resonance imaging (MRI) and computed tomography (CT) are the top two methods in brain imaging [3]. The scans can be used to evaluate brain structure and rule out other causes of cognitive decline [4]. MRI is uniquely suited for AD diagnosis due to its ability to provide high-resolution, detailed images of brain structures. This capability makes it invaluable for detecting subtle changes associated with AD, such as hippocampal atrophy, cortical thinning, and

* Corresponding authors at: School of Computing and Mathematical Sciences, University of Leicester, Leicester, LE1 7RH, UK (Yu-Dong Zhang) & Department of Electrical and Computer Engineering, Stevens Institute of Technology, Hoboken, NJ 07030, USA (Yu-Dong Yao).
** Corresponding author.
E-mail addresses: 352888@njupt.edu.cn (S.-Y. Lu), yudongzhang@ieee.org (Y.-D. Zhang), Yu-Dong.Yao@stevens.edu (Y.-D. Yao).

https://doi.org/10.1016/j.bspc.2024.107263
Received 27 March 2024; Received in revised form 20 November 2024; Accepted 23 November 2024
Available online 27 November 2024

ventricular enlargement—key markers of disease progression [5]. Unlike CT, MRI avoids the use of ionizing radiation, making it safer for repeated scans in longitudinal studies and clinical settings. This is particularly advantageous when monitoring disease progression or evaluating the effects of therapeutic interventions. Furthermore, MRI is more accessible and cost-effective than functional imaging modalities like positron emission tomography (PET) or single-photon emission computed tomography (SPECT), which require the use of radioactive tracers and are often restricted to specialized facilities. Therefore, MRI can directly visualize structural atrophy in brain regions critical for memory and cognition, such as the hippocampus and medial temporal lobe, which are among the earliest areas affected in AD. This aligns with the pathological progression of the disease, where neurodegeneration begins in these regions and spreads to other parts of the brain. Additionally, MRI can quantify progressive loss of gray matter volume and disruptions in structural connectivity, both hallmarks of AD.

Manual interpretation of these brain images is a time-consuming task, which requires expertise and training. Meanwhile, manual interpretation is prone to human error, including misinterpretation or oversight of certain findings. This can result in missed diagnoses or incorrect assessments of the severity or progression of the condition. Also, there is a lack of standardized guidelines and criteria for manual interpretation of brain MRIs, leading to variability in interpretation practices among radiologists and clinicians. This lack of standardization can affect consistency and reliability in the diagnosis and monitoring of conditions. To overcome some of these drawbacks, there is ongoing research and development of computer-aided diagnosis (CAD) tools that can aid in the interpretation of brain images. These tools can provide quantitative measurements and objective assessments, improving accuracy and efficiency in the diagnosis and monitoring of AD.

The rapid development of deep learning, especially computer vision over the past ten years sheds light on the research on CAD tools for AD. Essentially, the goal of detecting AD in brain MRI scans can be viewed as an image classification problem, and it often involves the use of supervised learning models. Bi et al. [6] proposed two distinct deep models for learning relevant features. Initially, brain graphs were obtained from the brain MRI scans. Subsequently, a convolutional neural network (CNN) and a recurrent neural network were utilized to extract local and temporal features from these brain graphs, respectively. Finally, the feature vectors were concatenated and used to train an extreme learning machine for classification. The dataset utilized in their experiments encompassed three classes of brain MRI slices, namely normal, mild cognitive impairment (MCI), and AD. Zhu et al. [7] introduced a dual-attention CNN for the diagnosis of AD using brain structural MRI scans. They developed a spatial attention branch to capture local representations from the patches and utilized a multi-instance pooling structure to generate global representations based on the patches. The classification stage involved two fully connected layers. The experiments included binary classification tasks. Puente-Castro et al. [8] utilized a pre-trained ResNet as the representation extractor. The obtained features from the sagittal brain MRIs were fused with the sex and age information, which were used to train a support vector machine (SVM) for classification. Wang, et al. [9] utilized functional-MRI time series for AD detection. They designed an STNet to fuse spatial and temporal information from the input images using convolutional operations and long short-term memory (LSTM) blocks. The goal was to effectively capture both the spatial and temporal dynamics in the scans for accurate AD detection. Alorf and Khan [10] introduced a novel approach for Alzheimer's disease (AD) identification using functional MRI by leveraging a graph convolutional network (GCN). The brain network was constructed by extracting connected regions from the MRI slices. The GCN model was designed with graph path convolution, edge pooling, and node operations, and was trained for the purpose of classification. Additionally, a stacked sparse autoencoder was trained for comparison. The experimental results demonstrated that the GCN model outperformed the stacked sparse autoencoder in terms of accuracy. El-Sappagh et al. [11] employed a multi-modal dataset to recognize AD, MCI, and normal samples. Furthermore, they aimed to obtain accurate predictions on the conversion time values of the MCI samples. For both the classification and prediction tasks, LSTM demonstrated excellent performance and yielded promising results. Ilias and Askounis [12] leveraged transcript data for the recognition of AD. They applied transfer learning techniques, utilizing transformers, to classify AD from normal brain MRI scans. To detect AD in an interpretable manner, they enhanced siamese networks with a co-attention block. Furthermore, multi-task learning algorithms were employed to simultaneously label and assess the severity of dementia. The proposed models were extensively evaluated through a series of experiments. Loddo et al. [13] utilized three pre-trained CNN models, namely AlexNet, ResNet-101, and InceptionResNetV2, to classify AD based on brain scans. Each model underwent individual fine-tuning on the training set, and their predictions were combined using an average ensemble mechanism to produce the final labels. Their approach yielded promising classification results across four distinct datasets. Tanveer et al. [14] presented an ensemble-based deep model for AD recognition. They segmented the MRI scans to obtain the gray matter, white matter, and cerebrospinal fluid images. Those three types of images were employed to fine-tune three pre-trained VGG-16 models. A two-stage ensemble was implemented to gradually obtain the final outcomes. Cao et al. [15] developed an end-to-end CAD method to detect AD as well as locate dementia-related regions from brain MRI scans. A shallow CNN model was employed to generate initial global representations from the brain MRI scans, and a localization model was constructed based on an LSTM to obtain the dementia-related patches. Then, the feature vectors were extracted from these patches, which were used to train a classifier of fully-connected layers for AD identification. Two datasets were used in their evaluation experiments, and their model achieved satisfactory results in both classification and localization. Pradhan et al. [16] used three different brain MRI datasets to train three identical CNN models, respectively. The outcomes of the three CNNs were concatenated to produce the ensemble predictions. However, the datasets were all small. Qasim Abbas et al. [17] proposed to detect AD using 3D MRI scans. Firstly, pre-processing was implemented, including alignment correction, intensity correction, and registration. Then, the 3D brain MRIs were transformed into the Jacobian space to enhance disease-related feature learning. Finally, a conventional CNN model was trained to identify AD from normal samples. Shamrat et al. [18] compared five famous CNN models for AD classification using brain MRI scans and found InceptionV3 outperformed other CNNs. Then, their AlzheimerNet was developed based on the InceptionV3 by parameter modification. A large dataset containing 60,000 images of 6 different types was used for evaluation, and the classification performance of the AlzheimerNet was promising. Yao et al. [19] suggested using fuzzy C-means to rearrange pixel distribution in the brain MRI scans for better feature learning. They employed a VGG-Net with batch normalization for multi-class classification. Zhang et al. [20] converted the AD detection into a graph classification problem, and different regions of a brain MRI scan are connected with spatial and semantic weights. The graph convolutions can generate more useful information for classification, so their model gained accuracy improvements. Sait and Nagaraj [21] presented a Feature-Fusion model to leverage the features from a CNN and a ViT for AD classification in 2D MRI slices.

As we have seen from the analysis, most of the current CAD systems for AD detection use brain MRI and deep CNN models. This makes sense because CNN models have made great progress in computer vision in the last ten years. CNNs use convolutional layers to concentrate on local regions in an image, which helps them to capture spatial relationships and filter out unimportant factors, increasing efficiency and accuracy. Moreover, CNN models can identify objects in images no matter where they are or how they are oriented, thanks to the shared weights and pooling layers. This makes them resistant to translations and rotations, and suitable for tasks like object detection and image classification. Also,

the shared weights greatly lower the number of parameters compared with fully connected networks, making CNNs more memory-efficient and faster to train. Besides, transfer learning allows CNN models trained on large datasets to be used as a starting point for solving other image recognition tasks. By fine-tuning and adapting the pre-trained CNN to a new task, one can take advantage of the representation learning ability and achieve good performance within a few epochs. However, existing CAD systems for AD diagnosis using MRI face several challenges, including the localized learning nature of CNNs, which limits global feature extraction, and the computational overhead of advanced models in clinical settings. Moreover, detecting early-stage AD, especially very mild dementia, remains difficult due to class imbalance in datasets.

To overcome these problems and improve the classification performance of CAD methods for AD detection, a novel efficient model called RanCom-ViT is proposed in this paper. The vision transformer (ViT) is a variant of a transformer for computer vision tasks as vanilla transformers are designed for language understanding [22]. Unlike CNN models, the basic block of a ViT is the multi-head self-attention (MHSA) block, which can generate better contextual representations from the images. Nevertheless, the computational cost for MHSA is much higher than convolution operation, and for CAD models, only specific regions of an image are related to diseases. Therefore, we proposed to employ a token compression block (TCB) to drop the irrelevant tokens and reserve the contributive tokens adaptively in the MHSA block, which can improve computational efficiency. We further modified the head block using a random enhanced projection layer inspired by random vector functional-link (RVFL), and the random features are concatenated with the class token embedding for classification. In summary, the novelty of this paper is listed:

I. A RanCom-ViT is proposed for AD classification in brain MRI scans using a ViT as the backbone for better contextual representation learning.
II. A token compression block is proposed to improve the computational efficiency of the RanCom-ViT by dropping the inattentive tokens.
III. The head block of the RanCom-ViT is improved with a random enhanced feature projection to achieve better generalization performance.

The proposed RanCom-ViT is evaluated on a public brain MRI scan dataset, and extensive experiments are conducted. The remainder of this paper is organized as follows. Section 2 introduces the brain MRI dataset that we use for evaluation. Section 3 describes the proposed RanCom-ViT in detail. Section 4 presents the experiments and discussion, and Section 5 is about the conclusion and future research directions.

2. Materials

We evaluate the proposed RanCom-ViT on a large public brain MRI dataset for AD analysis, which can be accessed on Kaggle (https://www.kaggle.com/datasets/ninadaithal/imagesoasis, accessed on 15 September 2023). The dataset consists of 461 patients from the open access series of imaging studies-1 (OASIS-1) MRI cross-sectional dataset [23]. Each 3D MRI volume is split into 256 2D scans along the sagittal view, and only the scans ranging from the 100th to the 160th are added to the dataset. The dataset has four classes with a total of 86,437 MRI scans: 67,222 non-demented, 13,725 very mild dementia, 5002 mild dementia, and 488 moderate dementia. The dataset is imbalanced, but it reflects the real-world situation as moderate dementia patients are rare in clinical diagnosis. The slices have a size of 496×248 pixels, which are resized to 384×384 before being input to the RanCom-ViT. Fig. 1 shows some examples of brain MRI scans from the four classes in the dataset. Each class shows distinct differences in anatomical features that correspond to varying stages of AD:

- Non-Demented: The brain appears structurally intact, with well-defined cortical and subcortical regions. Minimal atrophy is observed, and the hippocampal and medial temporal regions remain preserved.
- Very Mild Dementia: Subtle signs of atrophy begin to emerge, particularly in the hippocampal and medial temporal regions. The ventricular spaces may show slight enlargement, indicating early structural changes associated with AD.
- Mild Dementia: Noticeable cortical thinning and hippocampal atrophy are observed, alongside increased enlargement of the lateral ventricles. These structural changes reflect moderate disease progression, with a more pronounced impact on memory and cognition.
- Moderate Dementia: Severe atrophy becomes evident in both cortical and subcortical regions. The lateral ventricles and sulci are significantly enlarged, reflecting advanced neurodegeneration. The structural integrity of critical regions such as the hippocampus is heavily compromised.

3. Methodology

The proposed RanCom-ViT for efficient AD classification employed a ViT as the backbone. Pre-training is a vital step for ViT models on downstream tasks because the number of parameters in ViT models is usually much larger than that in CNN models. Therefore, it is harder for ViT models to converge. Pre-trained weights in a ViT obtained from the ImageNet datasets are capable of generating complex and contributive representations from a variety of images. The diversity of images in big datasets is beneficial for better generalization as well as faster convergence. However, the computational complexity of ViT models is still higher than CNNs because MHSA is a memory-intensive global representation learning mechanism. To improve the efficiency during training

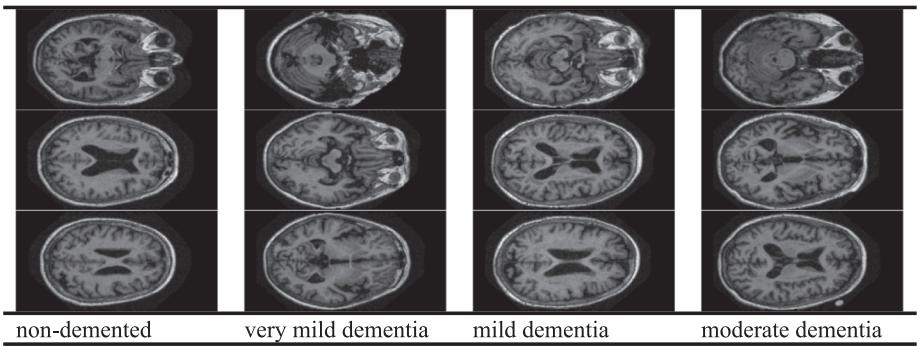

Fig. 1. Brain MRI samples of four classes. (The four classes are non-demented, very mild dementia, mild dementia, and moderate dementia.)

and testing, a token compression block is proposed to embed in the ViT. Depending on the importance of the tokens, the tokens can be compressed adaptively by dropping the inattentive tokens. In addition, a random enhanced feature projection is proposed to improve the classification performance of the RanCom-ViT and mitigate the overfitting problem. The embedding in the class token of the RanCom-ViT is randomly projected to obtain the enhanced feature vector, which is concatenated with the original embedding for final classification. A block diagram of the proposed RanCom-ViT is depicted in Fig. 2. The details of the proposed RanCom-ViT are shown in the rest of this section.

3.1. Backbone and transfer learning

The proposed RanCom-ViT uses a ViT as its backbone. Transformers are good at learning from sequential data, such as text and voice, so they excel in language understanding. Recently, they have also been applied to computer vision tasks. To do this, an image is split into non-overlapping patches that form a sequence, which is then fed into the transformers for training. The key idea behind the transformers is MHSA. MHSA allows each patch of the image to interact with all other patches, capturing the relationships between different regions of the image. This is different from CNNs, which have been the dominant models for image processing tasks. CNNs use convolution operations, which are limited in their ability to model long-range dependencies and global context in images. MHSA in transformers overcomes this limitation by computing weighted representations of the patches based on their relevance or importance to each other. This way, the model can extract high-level features that consider the global context of the image.

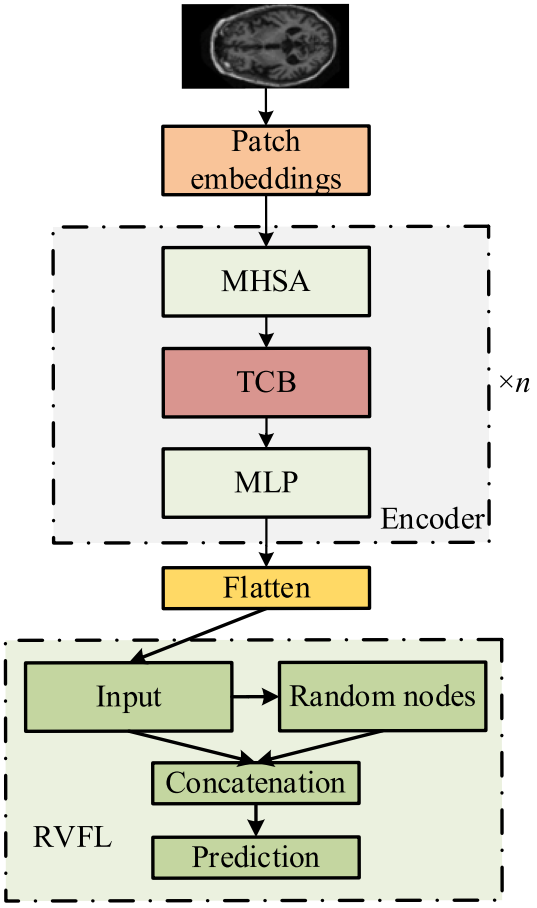

Fig. 2. A detailed block diagram of the RanCom-ViT architecture. (The diagram visually represents the key components, including the ViT backbone, token compression block (TCB), and RVFL-inspired classification head.)

An MHSA block consists of five key components: linear projection, dot-product attention, softmax and normalization, sum, and output transformation. An MHSA block has the same input and output dimensions, which allows multiple MHSA blocks to be stacked together to form a deep ViT model. This makes ViT models deeper than CNN models. However, a deeper structure also means more parameters, computation, and memory. It also poses a challenge for training a deep ViT directly on the brain MRI dataset for classification, as the dataset is too small and ViT models need long training epochs even on much larger datasets like ImageNet-21 K. To solve this problem, we use transfer learning to leverage the pre-trained ViT models.

Transfer learning allows us to use the knowledge and learned representations from a large dataset to a smaller one. We modify the output layer of the pre-trained ViT model to match the number of target classes before fine-tuning it on the brain MRI dataset. We also have the option to freeze some or all of the layers (except the head) of the pre-trained ViT model. Freezing the layers means that they do not change their weights during fine-tuning, while only the output layer and maybe some other layers are updated. Fine-tuning all layers means that the model can learn features that are specific to the target task. In this study, we fine-tune all the layers in the ViT model for AD classification in brain MRI scans, because we have over 80,000 brain MRI scans, which is a relatively large dataset for medical image processing, and brain MRI scans are not similar to the common images in ImageNet datasets, so we need to update the parameters in the early layers for feature extraction as well.

3.2. Token compression block

The MHSA blocks improve the global and contextual representation learning capability of ViT models, but the computational cost also increases. Specifically, in brain MRI scans, there are considerable background regions that are not associated with the AD classification task. Meanwhile, only a small number of parts in a brain are believed to be important for AD classification, such as the hippocampus, medial temporal lobe, and lateral ventricle. Therefore, the tokens in MHSA blocks can be compressed to reduce the computational intensity. To this end, a token compression block (TCB) is presented to improve the training and testing efficiency of the RanCom-ViT [24]. The core idea behind the TCB is to measure the significance of all the patches and remove the unimportant ones. In AD classification, the embedding from the class token is used as the feature vector, which is generated from the tokens of all the patches as

$$\boldsymbol{t}_{CLS} = \text{softmax}\left(\frac{\boldsymbol{q}_{CLS}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V} = \boldsymbol{w} \bullet \boldsymbol{V} \tag{1}$$

in which $\boldsymbol{t}_{CLS}$, and $\boldsymbol{q}_{CLS}$ stand for the embedding of class token and its query vector, d is the dimension of the query vector, and $\boldsymbol{K}$, and $\boldsymbol{V}$ denote the key and value matrixes from all the patch tokens. Intuitively, the class token can be regarded as a transformed value matrix using a weighting factor $\boldsymbol{w}$. Subsequently, the importance of the i^{th} patch token can be regarded by the w_i in the weighting vector $\boldsymbol{w}$. This is the example for vanilla self-attention, and for MHSA, there are different weighting vectors for different heads. The importance of the tokens can be measured by the averaged weighting vector conveniently. In practice, a hyper-parameter p is employed to denote the proportion of dropped tokens. For example, $p = 30$ % means that 30 % of the patch tokens are dropped while the rest 70 % of tokens are preserved for the following layers in the RanCom-ViT.

The TCB blocks can be inserted in the encoders after the MHSA operations, shown in Fig. 3, which can substantially improve the computational efficiency and maintain good classification.

3.3. Head improvement using random projection

Random features have been used in traditional machine learning

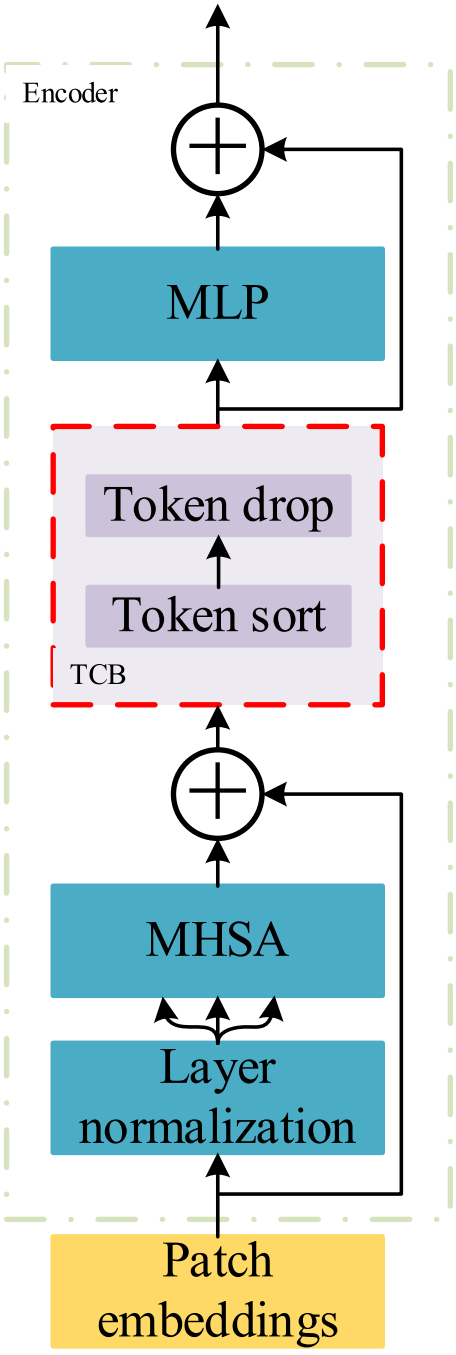

Fig. 3. Token compression in a transformer encoder (In an encoder, a token compression block is inserted after the multi-head self-attention operation to remove the least contributive tokens and improve the efficiency).

algorithms for decades, such as the extreme learning machine (ELM) and random vector functional-link (RVFL) [25]. The random projected features are found to be helpful in overcoming overfitting and achieving better classification performance [26], which often occurs with small datasets. To this end, instead of using a fully connected layer as the head in the RanCom-ViT, a structure like the RVFL is designed for classification, as shown in Fig. 4.

The input of the RVFL is the embedding of the class token $\boldsymbol{t}_{CLS} = [t_1, t_2, t_3, \ldots, t_d]^T$, which is projected into another feature space using the random weights $\boldsymbol{w}_i$, and these weights do not change once initialized. The random features are then joined with the original features to produce the final output. The original RVFL usually uses the pseudo-inverse

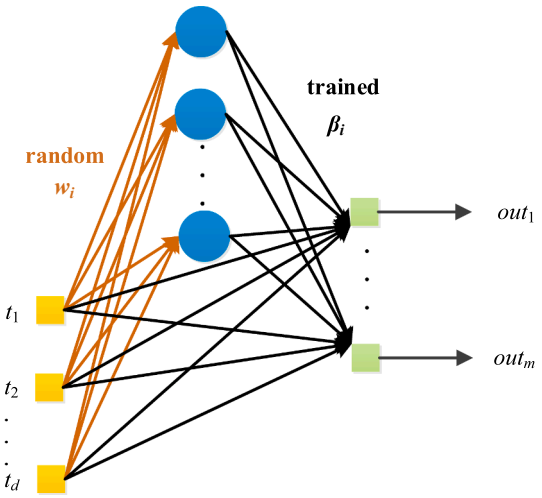

Fig. 4. Head in the RanCom-ViT. (The class token embedding is mapped into a hidden space with random weights, and concatenated with the input features to compute the predictions.)

algorithm for training, but we suggest training it directly with the entire model, as it is part of the RanCom-ViT's head. The important condition is to keep the input weights $\boldsymbol{w}_i$ from being updated during training. This training method allows for a better combination of the RVFL with the whole model.

Unlike other random neural networks such as the extreme learning machine, the RVFL has a direct connection from the input to the output layer, which makes the classification more stable.

3.4. Hyper-parameter settings

Table 1 shows some important hyper-parameters that affect the classification performance of the RanCom-ViT. We use the same values as the DeiT_base_patch16_384 backbone model for these parameters: input size, patch size, model depth, and embedding dimension. The embedding dimension is the feature dimension of the transformer encoder blocks, and it is also the input dimension for the RVFL classifier head. We set the number of enhanced nodes in the head to 768, which matches the feature dimension. The proportion for dropping unimportant tokens is defined as 30 %, which is the same for both training and testing in this study. The 30 % token drop rate used in our model is chosen as a balance between computational efficiency and classification performance. We use SGD (stochastic gradient descent with momentum) as the optimizer to fine-tune the RanCom-ViT. The learning rate is 1×10^{-4} and the momentum factor is 0.9, which helps to achieve convergence. The batch size is 32, which is suitable for the GPU memory of 24 GB. The maximum number of training epochs is 20, which is a small value to avoid overfitting. Deep models usually need hundreds of epochs to converge, but we choose a smaller value because of computational limitations and overfitting prevention.

Table 1
Hyper-parameters in the RanCom-ViT.

Hyper-parameter	Value
Number of enhanced nodes	768
Proportion for dropping tokens	30 %
Optimizer	SGD
Learning rate	1×10^{-4}
Momentum	0.9
Batch size	32
Max epoch	20

4. Experimental results and discussion

We tested the proposed RanCom-ViT model on a public brain MRI scan dataset using the PyTorch platform, and all the results were obtained with an NVIDIA RTX 3090 GPU. To measure the generalization performance of the RanCom-ViT, we split the dataset randomly into 80 % for training and 20 % for testing. This ratio is commonly used in machine learning studies as it provides a sufficient amount of data for training the model while retaining a reasonable sample size for testing. To ensure an equal and random distribution of each dementia stage (Non-Demented, Very Mild Dementia, Mild Dementia, and Moderate Dementia), we stratified the dataset by class during the split. This stratified splitting method guarantees that the proportions of each class in the training and testing sets are representative of the overall dataset distribution. It should be noted that the trained RanCom-ViT is portable and can run on different devices for inference, as long as the device meets the required specifications. We used sensitivity, specificity, precision, F1 score, and accuracy as performance metrics, which can be easily calculated from confusion matrixes.

4.1. Generalization performance of the RanCom-ViT

As shown in Fig. 5 and Table 2, the results highlight the exceptional performance of the proposed RanCom-ViT model in Alzheimer's disease classification across four categories: Mild Dementia, Moderate Dementia, Non-Demented, and Very Mild Dementia.

Table 2
Classification metrics of RanCom-ViT.

	Precision	Sensitivity	Specificity	F1 score	Accuracy
Mild dementia	99.39 %	97.90 %	99.96 %	98.64 %	99.54 %
Moderate dementia	**100.00 %**	98.97 %	**100.00 %**	99.48 %	
Non-demented	99.63 %	**99.88 %**	98.70 %	**99.75 %**	
Very mild dementia	99.16 %	98.51 %	99.84 %	98.83 %	
Average	99.55 %	98.82 %	99.63 %	99.18 %	

The provided accuracy curves show that the model continues to improve steadily up to the 19th epoch. While the accuracy growth begins to plateau after approximately 16 epochs, the model does not fully stabilize until the final epoch (19th). By the 19th epoch, the testing accuracy reaches approximately 99.54 %, as reported in the results section, indicating that the model requires nearly the full 20 epochs to achieve optimal performance.

For Mild Dementia, the model correctly classified 979 out of 985 samples, with only six misclassifications, including five as Very Mild Dementia and one as Moderate Dementia. This demonstrates the model's high sensitivity (97.90 %) and precision (99.39 %) for this category, reflecting its ability to accurately detect Mild Dementia cases with minimal false positives.

For Moderate Dementia, the model exhibits outstanding performance, correctly identifying all 96 cases with no false positives or

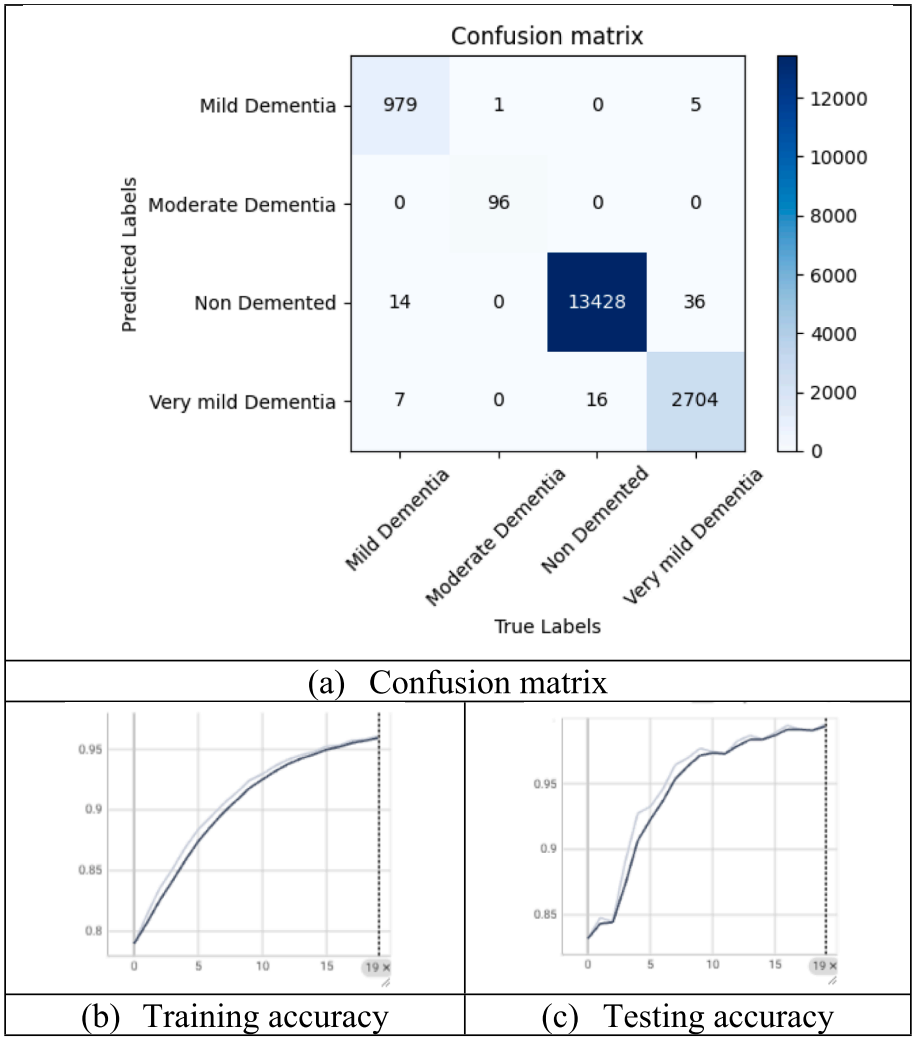

Fig. 5. Confusion matrix and accuracy plots of RanCom-ViT.

negatives. This leads to perfect precision (100 %) and specificity (100 %), and near-perfect sensitivity (98.97 %). These results emphasize the RanCom-ViT's robustness in detecting this less frequent category, a critical achievement given the challenges posed by smaller sample sizes for moderate cases.

In the Non-Demented category, which is the largest in the dataset, the model demonstrates remarkable accuracy, correctly classifying 13,428 samples. Despite a small number of misclassifications (14 as Mild Dementia and 36 as Very Mild Dementia), the metrics remain impressive with a sensitivity of 99.88 % and precision of 99.63 %. These results highlight the model's ability to handle a majority class effectively while maintaining robustness across the dataset.

The Very Mild Dementia category, representing early-stage Alzheimer's disease, presents the greatest challenge. While the model correctly classifies 2,704 cases, there are still 36 misclassifications into the Non-Demented class and seven into the Mild Dementia category. This results in slightly lower sensitivity (98.51 %) and F1-score (98.83 %) compared to other categories. Although the performance remains strong, the model's ability to detect early-stage AD can be further enhanced to ensure timely and accurate diagnosis.

Overall, the RanCom-ViT achieves exceptional average metrics, including a precision of 99.55 %, sensitivity of 98.82 %, F1-score of 99.18 %, and an overall accuracy of 99.54 %. These results demonstrate the model's reliability and robustness in distinguishing between AD categories, including the more challenging early stages.

However, a closer analysis reveals that improvements in detecting Very Mild Dementia are necessary. The misclassifications into the Non-Demented category highlight the need for strategies to address the imbalance in the dataset and improve sensitivity for early-stage AD. Future efforts could include augmenting the dataset, optimizing class-specific loss functions, or incorporating 3D MRI data to better capture subtle features indicative of early dementia.

In conclusion, the RanCom-ViT model demonstrates state-of-the-art performance across all categories, particularly excelling in Moderate Dementia classification. By addressing the remaining challenges in Very Mild Dementia detection, the model holds promise for an even greater impact in early and accurate Alzheimer's disease diagnosis.

4.2. Ablation study on pre-training

The classification results of the RanCom-ViT trained from scratch are provided in Fig. 6 and Table 3 to compare with the RanCom-ViT using pre-trained weights from the ImageNet dataset. The RanCom-ViT model trained from scratch has significantly more misclassifications than the RanCom-ViT model using pre-trained weights. The RanCom-ViT model trained from scratch fails to recognize any moderate dementia cases, with 0.00 % precision and sensitivity, and has a low overall accuracy of 79.65 %. This shows that the RanCom-ViT model trained from scratch cannot differentiate the four types of brain dementia within 20 epochs. Although the brain MRI scans are very different from the natural images in the ImageNet dataset, the pre-trained weights help the RanCom-ViT model converge faster on the brain MRI dataset. We think that the large and diverse ImageNet dataset can provide a good generalization ability that is useful for downstream tasks and that there may be some common features in the latent representation space between the two datasets.

4.3. Ablation study on token compression

We provide the classification results of the RanCom-ViT without

Table 3
Classification metrics of RanCom-ViT trained from scratch.

	Precision	Sensitivity	Specificity	F1 score	Accuracy
Mild dementia	41.82 %	2.30 %	99.80 %	4.36 %	79.65 %
Moderate dementia	0.00 %	0.00 %	**100.00 %**	0.00 %	
Non-demented	**82.22 %**	**97.21 %**	26.42 %	**89.09 %**	
Very mild dementia	50.64 %	24.63 %	95.47 %	33.14 %	
Average	43.67 %	31.04 %	80.42 %	31.65 %	

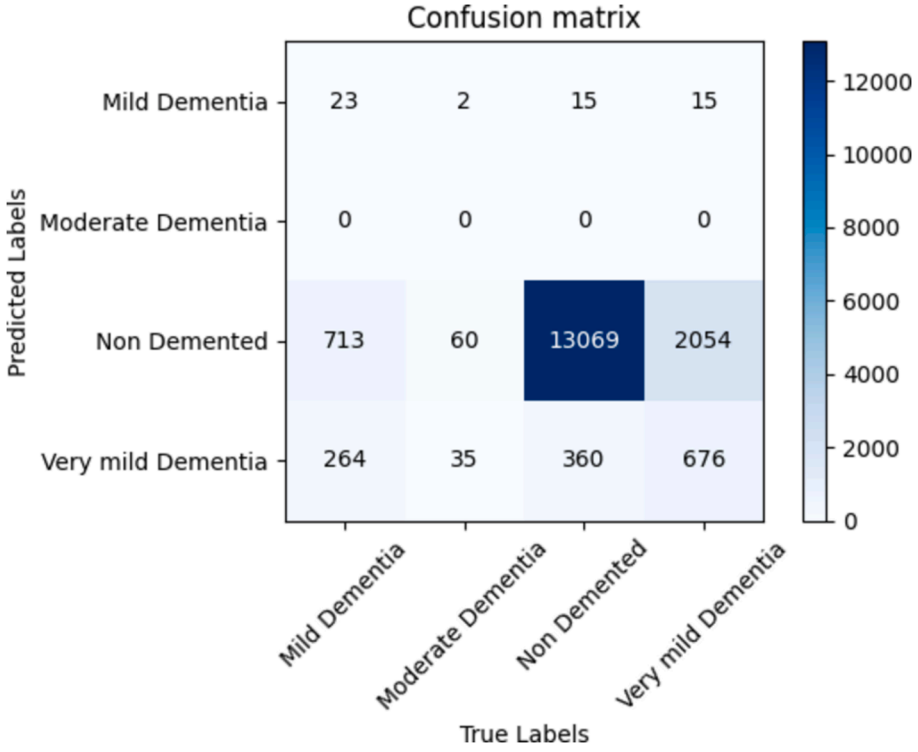

Fig. 6. Confusion matrix of RanCom-ViT trained from scratch.

TCB, as shown in Fig. 7 and Table 4, to evaluate the effectiveness of the TCB, and the throughput comparison is listed in Table 5. With no tokens dropped, the number of very mild dementia samples classified as non-demented increases from 36 to 49, and the number of misclassifications for non-demented samples is also higher than the RanCom-ViT with TCB. As the early diagnosis of AD is more important, more attention should be paid to the classification performance for very mild dementia. Both models achieve an accuracy of over 99 %, and the difference between the two models for all the five metrics is marginal though the RanCom-ViT with TCB outperforms the other model. However, the TCB contributes to less training and testing time (both 20 epochs), and the inference throughput grows by a factor of more than two. Therefore, the TCB is effective in improving the computational cost.

4.4. Ablation study on classification head

The ablation study for the head of the RanCom-ViT is performed, and the results of the model without randomly projected features are given in Fig. 8 and Table 6. The model without randomly projected features has some advantages over the model with randomly projected features, such as correctly classifying all the moderate dementia samples and reducing the number of very mild dementia samples misclassified as non-demented. However, the model without randomly projected features also has some drawbacks, such as lower precision, specificity, and F1-score for very mild dementia. Moreover, the overall accuracy is higher for the model with randomly projected features. Therefore, we conclude that the randomly projected features in the head of the RanCom-ViT model are beneficial for AD classification.

4.5. Token compression visualization

Visualization is a significant method to understand the predictions from deep models because medical practitioners cannot interpret the meanings of the weights. The TCB in the RanCom-ViT can improve the efficiency of the model during both training and inference stages, and it is also important to know which tokens are dropped. Hence, a visualization of the patch tokens is illustrated in Fig. 9. As the proportion for dropping unimportant tokens is 30 %, only 70 % of tokens are preserved in every transformer encoder, so the number of tokens will reduce with the forward propagation. It can be observed that the regions containing more textual information are preserved in the encoder block 1 while the patches near the corners of the background are often dropped. However, in the encoder block 2, more regions associated with AD are dropped, which is inevitable because there will always be 30 % tokens that have to be dropped. Moreover, the classification performance of the RanCom-ViT is satisfactory without the dropped tokens.

Table 4
Classification metrics of RanCom-ViT without TCB.

	Precision	Sensitivity	Specificity	F1 score	Accuracy
Mild dementia	99.20%	98.90%	**99.95%**	99.05%	99.40%
Moderate dementia	**100.00%**	95.88%	100.00%	97.90%	
Non-demented	99.61%	**99.72%**	98.65%	**99.66%**	
Very mild dementia	98.39%	98.11%	99.70%	98.25%	
Average	99.30%	98.15%	99.58%	98.72%	

Table 5
Throughput comparison.

	Training and testing time	Throughput for inference
RanCom-ViT with TCB	6.5 h	86.6 images/s
RanCom-ViT without TCB	10.1 h	40.2 images/s

4.6. Comparison with SOTA models

The proposed RanCom-ViT model is evaluated on a public dataset of brain MRI scans and compared with state-of-the-art (SOTA) methods for AD classification, as shown in Table 7. Most of the existing methods are

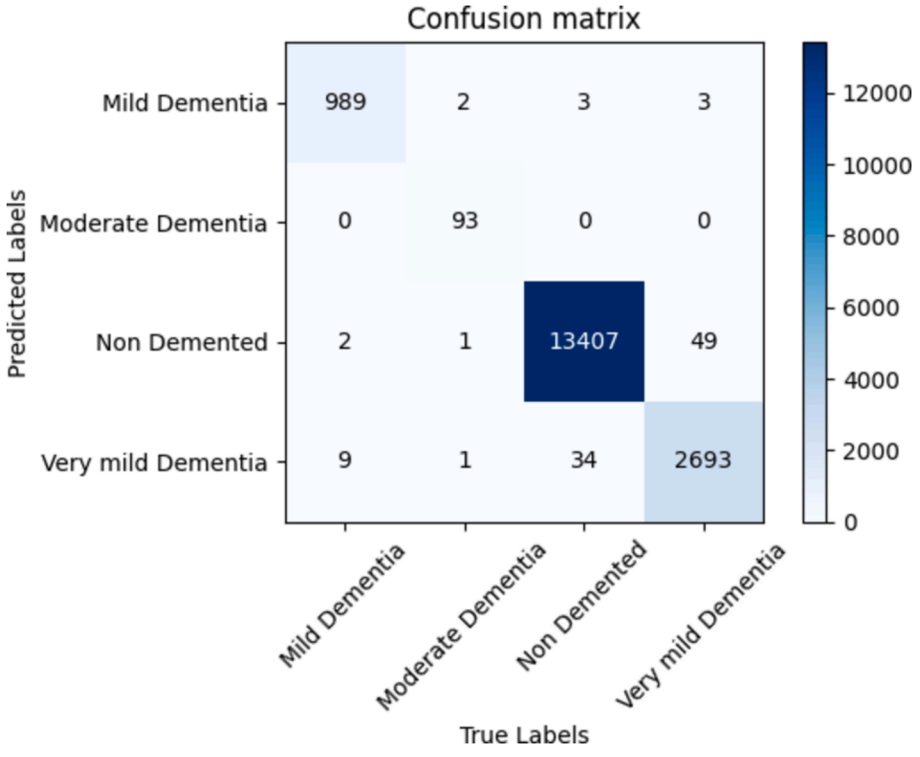

Fig. 7. Confusion matrix of RanCom-ViT without TCB.

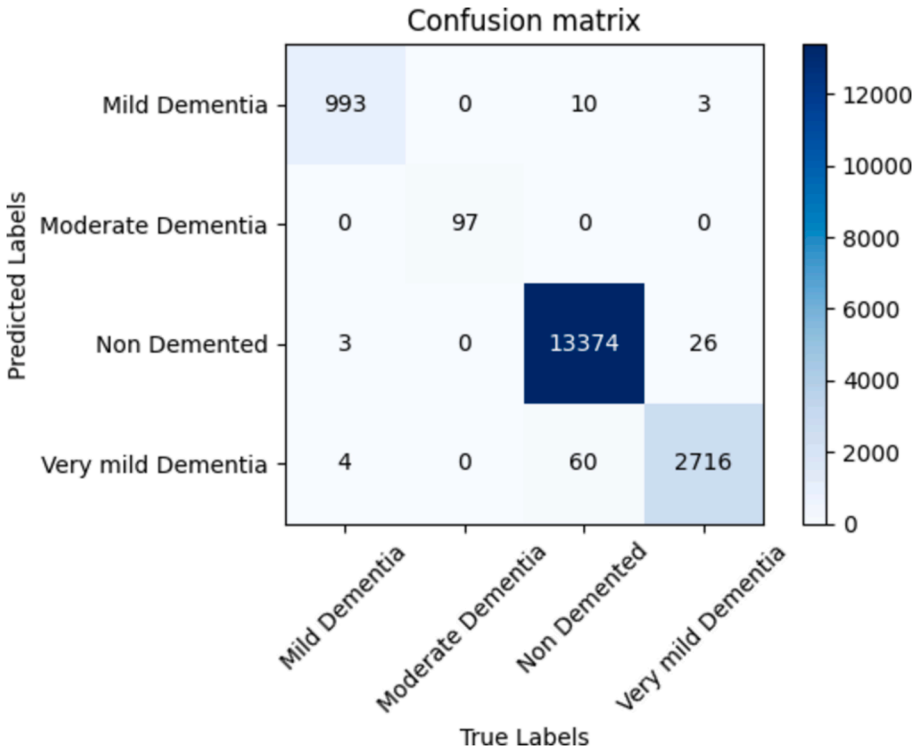

Fig. 8. Confusion matrix of RanCom-ViT without randomly projected features.

Table 6
Classification metrics of RanCom-ViT without randomly projected features.

	Precision	Sensitivity	Specificity	F1 score	Accuracy
Mild dementia	98.71 %	99.30 %	99.92 %	99.00 %	99.39 %
Moderate dementia	**100.00 %**	**100.00 %**	**100.00 %**	**100.00 %**	
Non-demented	99.78 %	99.48 %	99.25 %	99.63 %	
Very mild dementia	97.70 %	98.94 %	99.56 %	98.32 %	
Average	99.05 %	99.43 %	99.68 %	99.24 %	

based on CNN or recurrent neural network architectures, which have limitations in capturing the global and local features of brain images. One exception is BC-GCN [10], which uses a graph convolutional network to classify six types of scans. However, the dataset used by BC-GCN is very small, with only 706 samples, which may affect its generalization ability. Two other methods, deep ensemble [13] and deep transfer ensemble [14], report very high accuracies, but they have some drawbacks. The deep transfer ensemble only deals with binary classification, while the deep ensemble has a low sensitivity. LSTM [11] and STNet [9] are two methods that exploit the spatial information of 3D brain MRI, which is more difficult than 2D image classification. However, their accuracies are much lower than the proposed RanCom-ViT. Moreover, the 2D images in the dataset are selected automatically, without human intervention, which adds to the challenge of the classification task. The Feature-Fusion model [21] was evaluated on the same dataset as ours, and its performance is slightly worse than our model in terms of all the metrics. The proposed RanCom-ViT model incorporates a token compression block (TCB), which is a novel approach compared to existing methods. While conventional ViTs process all tokens equally, our method adaptively drops less relevant tokens based on their contributions to the classification task. This selective token reduction substantially enhances computational efficiency without compromising accuracy, which sets our method apart from standard ViTs and CNN-based models. Unlike existing models that rely on fixed token pruning strategies or global pooling, the TCB dynamically assesses token importance using attention weights generated within the transformer layers. This ensures that only the most informative tokens are retained, resulting in more efficient global feature extraction tailored for AD diagnosis.

Traditional CNNs and attention-based methods typically lack mechanisms to adaptively prioritize input features. For example, CNNs rely on localized filters that do not leverage global context effectively. Existing ViT-based models, such as DeiT and Swin Transformer, do not include an integrated mechanism for token compression and instead focus on fixed hierarchical down-sampling.

5. Conclusion

This paper presents a novel ViT model, named RanCom-ViT, for AD classification using 2D MRI slices. ViT models have achieved remarkable results on various computer vision tasks, but they also require more computation than CNN models. To overcome this limitation, the RanCom-ViT introduces a novel token merging block within its self-attention blocks, which reduces the number of tokens and improves computational efficiency. Moreover, an RVFL structure is used as the classification head, which leverages random projection to enhance the generalization performance. The proposed RanCom-ViT is evaluated on a large public dataset of brain MRI slices with four different scan types. The experimental results show that the RanCom-ViT outperforms the state-of-the-art methods for AD classification, achieving an overall accuracy of more than 99 %. This indicates the effectiveness of the RanCom-ViT as a powerful CAD system for AD diagnosis.

The RanCom-ViT model holds significant potential for translation into clinical research and radiological practice. Its ability to achieve high accuracy (99.54 %) and computational efficiency makes it a valuable tool for augmenting the diagnostic workflow in Alzheimer's disease (AD). Specifically, this technology could be employed to automate the initial screening of MRI scans, enabling faster identification of patients with possible cognitive impairment and flagging cases that require further evaluation. This would save time for radiologists and allow them

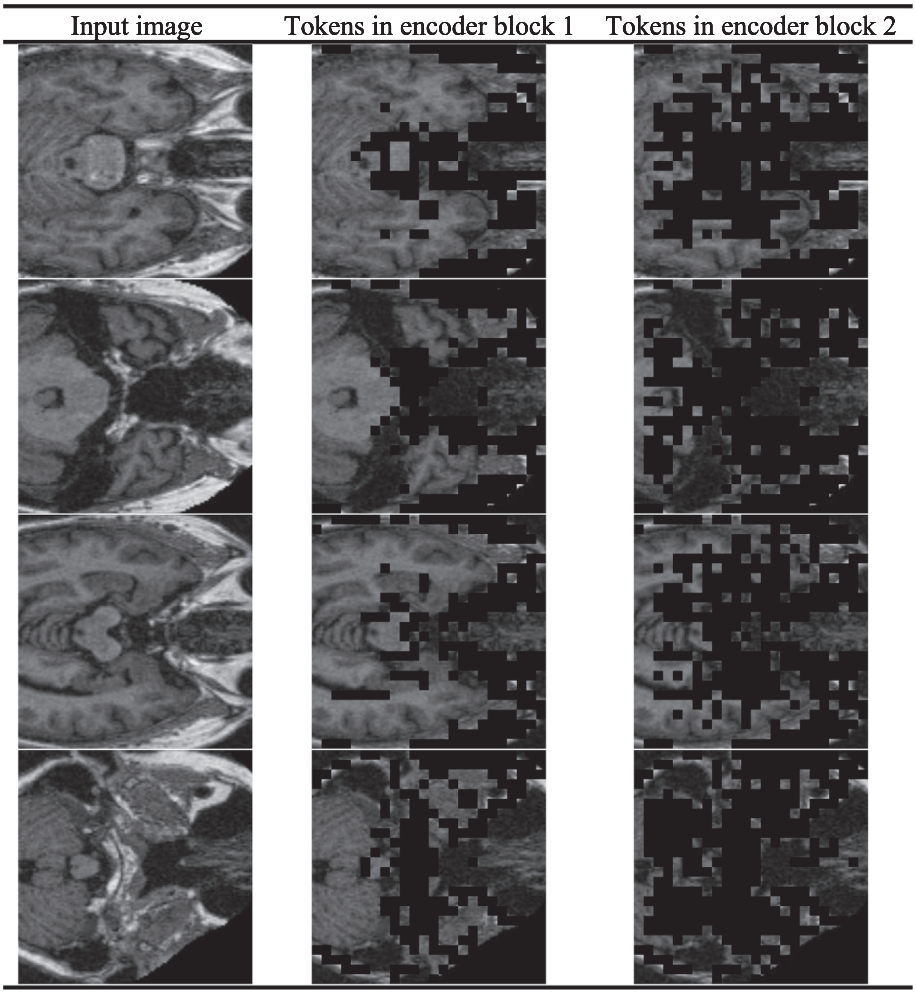

Fig. 9. A visualization of patch tokens in the TCB.

Table 7
Comparison with state-of-the-art methods.

Method	Precision	Sensitivity	Specificity	F1 score	Accuracy	# of classes	# of samples
STNet [9]	–	–	–	–	60.67 %	4	563
BC-GCN [10]	–	–	–	–	84.03 %	**6**	706
LSTM [11]	78.74 %	76.39 %	–	82.80 %	82.80 %	4	1,371
Deep-ensemble [13]	–	93.05 %	97.31 %	–	98.24 %	4	2,229
Deep transfer ensemble [14]	–	97.32 %	99.37 %	–	98.71 %	2	813
Feature-Fusion [21]	97.9 %	99.1 %	97.5 %	98.0 %	98.8 %	4	**86,437**
RanCom-ViT (proposed)	**99.55 %**	**98.82 %**	**99.63 %**	**99.18 %**	**99.54 %**	4	**86,437**

to focus on more complex cases requiring detailed interpretation. Moreover, the model's capacity to analyze large volumes of MRI data consistently and accurately reduces the risk of diagnostic errors associated with manual interpretation. Variability among radiologists is a known challenge, and an AI-driven tool like RanCom-ViT could provide objective, standardized assessments, serving as a second opinion to support clinical decisions. This is especially relevant in the detection of early-stage AD, where subtle structural changes such as hippocampal atrophy might be overlooked during visual assessment. In clinical research, the RanCom-ViT model could facilitate the exploration of longitudinal changes in brain structure by processing large MRI datasets efficiently, enabling researchers to study the progression of AD across different patient cohorts. Its computational efficiency, bolstered by the token compression block, ensures scalability for such tasks without compromising accuracy.

Despite the promising performance of the proposed RanCom-ViT, certain limitations remain that need to be acknowledged. First, the model shows reduced sensitivity in detecting very mild dementia, the earliest stage of AD. This limitation is likely due to the inherent class imbalance in the dataset, as very mild dementia cases are underrepresented compared to normal ones. Additionally, while our method processes 2D MRI slices effectively, it does not leverage the full spatial information available in 3D MRI data, which may limit the model's ability to capture certain structural patterns critical for early-stage AD diagnosis. Another limitation lies in the computational intensity of the ViT backbone. Although the token compression block significantly improves efficiency, the model's complexity remains a potential challenge for deployment in resource-constrained clinical environments.

To address these limitations, future research will focus on several key areas. First, efforts will be made to collect and incorporate more data for early-stage AD, aiming to improve the model's sensitivity and robustness in detecting very mild dementia. Furthermore, we plan to extend

the RanCom-ViT framework to process 3D MRI data. By capturing spatial–temporal features, the enhanced model could provide deeper insights into early structural changes associated with AD. Lastly, we aim to refine the token compression mechanism and explore lightweight transformer architectures, enabling the model to maintain high performance while reducing computational demands, which will facilitate broader clinical adoption and real-time application.

CRediT authorship contribution statement

Si-Yuan Lu: Writing – review & editing, Writing – original draft. **Yu-Dong Zhang:** Project administration, Funding acquisition. **Yu-Dong Yao:** Supervision.

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.

Acknowledgement

This paper is partially supported by MRC (MC_PC_17171); Royal Society (RP202G0230); BHF (AA/18/3/34220); Hope Foundation for Cancer Research (RM60G0680); GCRF (P202PF11); Sino-UK Industrial Fund (RP202G0289); LIAS (P202ED10, P202RE969); Data Science Enhancement Fund (P202RE237); Fight for Sight (24NN201); Sino-UK Education Fund (OP202006); BBSRC (RM32G0178B8); Natural Science Research Start-up Foundation of Recruiting Talents of Nanjing University of Posts and Telecommunications (Grant No. XK0020923211); Jiangsu Provincial Distinguished Postdoctoral Program.

Data availability

Data will be made available on request.

References

[1] J. Hu, et al., Diagnostic performance of magnetic resonance imaging–based machine learning in Alzheimer's disease detection: a meta-analysis, Neuroradiology 65 (3) (2022) 513–527, https://doi.org/10.1007/s00234-022-03098-2.
[2] S.L. Warren, A.A. Moustafa, Functional magnetic resonance imaging, deep learning, and Alzheimer's disease: a systematic review, J. Neuroimaging 33 (1) (2022) 5–18, https://doi.org/10.1111/jon.13063.
[3] Y. Zhang, et al., Vascular-water-exchange MRI (VEXI) enables the detection of subtle AXR alterations in Alzheimer's disease without MRI contrast agent, which may relate to BBB integrity, Neuroimage 270 (2023), https://doi.org/10.1016/j.neuroimage.2023.119951.
[4] G. Macin, et al., An accurate multiple sclerosis detection model based on exemplar multiple parameters local phase quantization: ExMPLPQ, Appl. Sci. 12 (2022) 4920, https://doi.org/10.3390/app12104920.
[5] A. J. Chang et al., MRI-based deep learning can discriminate between temporal lobe epilepsy, Alzheimer's disease, and healthy controls, Commun. Med. 3(1) (2023), doi: 10.1038/s43856-023-00262-4.
[6] X. Bi, et al., Functional brain network classification for Alzheimer's disease detection with deep features and extreme learning machine, Cogn. Comput. 12 (3) (2019) 513–527, https://doi.org/10.1007/s12559-019-09688-2.
[7] W. Zhu, et al., Dual attention multi-instance deep learning for alzheimer's disease diagnosis with structural MRI, IEEE Trans. Med. Imag. 40 (9) (2021) 2354–2366, https://doi.org/10.1109/tmi.2021.3077079.
[8] A. Puente-Castro, et al., Automatic assessment of Alzheimer's disease diagnosis based on deep learning techniques, Comput. Biol. Med. 120 (2020) (2020) 1–7, https://doi.org/10.1016/j.compbiomed.2020.103764.
[9] M. Wang, et al., Spatial-temporal dependency modeling and network hub detection for functional mri analysis via convolutional-recurrent network, IEEE Trans. Biomed. Eng. 67 (8) (2020) 2241–2252, https://doi.org/10.1109/tbme.2019.2957921.
[10] A. Alorf, M.U.G. Khan, Multi-label classification of Alzheimer's disease stages from resting-state fMRI-based correlation connectivity data and deep learning, Comput. Biol. Med. 151 (2022), https://doi.org/10.1016/j.compbiomed.2022.106240.
[11] S. El-Sappagh, et al., Two-stage deep learning model for Alzheimer's disease detection and prediction of the mild cognitive impairment time, Neural Comput. Appl. 34 (17) (2022) 14487–14509, https://doi.org/10.1007/s00521-022-07263-9.
[12] L. Ilias, D. Askounis, Explainable identification of dementia from transcripts using transformer networks, IEEE J. Biomed. Health Inform. 26 (8) (2022) 4153–4164, https://doi.org/10.1109/jbhi.2022.3172479.
[13] A. Loddo, et al., Deep learning based pipelines for Alzheimer's disease diagnosis: A comparative study and a novel deep-ensemble method, Comput. Biol. Med. 141 (2022), https://doi.org/10.1016/j.compbiomed.2021.105032.
[14] M. Tanveer, et al., Classification of Alzheimer's disease using ensemble of deep neural networks trained through transfer learning, IEEE J. Biomed. Health Inform. 26 (4) (2022) 1453–1463, https://doi.org/10.1109/jbhi.2021.3083274.
[15] G. Cao, et al., End-to-end automatic pathology localization for Alzheimer's disease diagnosis using structural MRI, Comput. Biol. Med. 163 (2023), https://doi.org/10.1016/j.compbiomed.2023.107110.
[16] N. Pradhan, et al., Analysis of MRI image data for Alzheimer disease detection using deep learning techniques, Multimedia Tools Appl. (2023), https://doi.org/10.1007/s11042-023-16256-2.
[17] S. Qasim Abbas et al., Transformed domain convolutional neural network for Alzheimer's disease diagnosis using structural MRI, Pattern Recogn. 133 (2023), doi: 10.1016/j.patcog.2022.109031.
[18] F.M.J.M. Shamrat, et al., AlzheimerNet: an effective deep learning based proposition for Alzheimer's disease stages classification from functional brain changes in magnetic resonance images, IEEE Access 11 (2023) 16376–16395, https://doi.org/10.1109/access.2023.3244952.
[19] Z. Yao, et al., Fuzzy-VGG: a fast deep learning method for predicting the staging of Alzheimer's disease based on brain MRI, Inf. Sci. 642 (2023) (2023) 1–9, https://doi.org/10.1016/j.ins.2023.119129.
[20] J. Zhang, et al., Multi-relation graph convolutional network for Alzheimer's disease diagnosis using structural MRI, Knowl.-Based Syst. 270 (2023) (2023) 1–15, https://doi.org/10.1016/j.knosys.2023.110546.
[21] A. R. W. Sait, R. Nagaraj, A feature-fusion technique-based alzheimer's disease classification using magnetic resonance imaging, Diagnostics (Basel) 14(21) (2024), doi: 10.3390/diagnostics14212363.
[22] A. Dosovitskiy et al., An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale, Presented at the International Conference on Learning Representations (ICLR 2021), Virtual, 2021.
[23] L.N. Koenig, et al., Select Atrophied Regions in Alzheimer disease (SARA): an improved volumetric model for identifying Alzheimer disease dementia, NeuroImage: Clin. 26 (2020) (2020) 1–12, https://doi.org/10.1016/j.nicl.2020.102248.
[24] Y. Liang et al., Not All Patches Are What You Need: Expediting Vision Transformers Via Token Reorganizations, Presented at the IEEE/CVF International Conference on Learning Representations, Virtual, 2022.
[25] P.N. Suganthan, Letter: On non-iterative learning algorithms with closed-form solution, Appl. Soft Comput. 70 (2018) 1078–1082, https://doi.org/10.1016/j.asoc.2018.07.013.
[26] S. Baek et al., Face detection in untrained deep neural networks, Nat. Commun. 12 (1) (2021) 7328, doi: 10.1038/s41467-021-27606-9.