Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

Alzheimer's Disease Detection Using Vision Transformers: A survey

Nasrallah Asem AL-Sultani [a,*], Alaa Taima Albu-Salih [b], Osama Majeed Hilal [c]

[a] *Computer Science and Information Technology, University of Al-Qadisiyah, Iraq. Email: cm.post23.17@qu.edu.iq*

[b] *Computer Science and Information Technology, University of Al-Qadisiyah, Iraq. Email: alaa.taima@qu.edu.iq*

[c] *Computer Science and Information Technology, University of Al-Qadisiyah, Iraq. Email: osama.m@qu.edu.iq*

ARTICLE INFO

Article history:
Received: 06/06/2025
Rrevised form: 25 /06/2025
Accepted : 29/06/2025
Available online: 30/06/2025

Keywords:

Alzheimer's disease,

Vision transformer,

Optimization,

Deep Learning,

ViT.

ABSTRACT

Alzheimer's disease is a progressive neurodegenerative disorder that primarily affects individuals aged 65 and older, leading to irreversible memory loss and cognitive decline. Early detection plays a critical role in managing the disease and improving patient outcomes. In recent years, numerous studies have investigated the development of automated systems to identify the stages of Alzheimer's disease using advanced deep learning methods. This paper provides a structured literature review focused on the use of Vision Transformers (ViTs) and metaheuristic optimization algorithms for early diagnosis. The reviewed studies demonstrate that ViT-based models outperform traditional approaches in extracting spatial and temporal features from brain imaging data, achieving classification accuracies exceeding 96% on widely used datasets such as ADNI and OASIS. Additionally, the review addresses key challenges in processing 3D medical images and highlights ongoing efforts to develop hybrid architectures that integrate the strengths of Convolutional Neural Networks (CNNs) and Vision Transformers (ViT). The paper also explores how collaborative learning strategies can enhance model training while preserving patient privacy, making these techniques more suitable for real-world clinical applications.

MSC..

https://doi.org/10.29304/jqcsm.2025.17.22229

1. Introduction

Here Alzheimer's disease (AD) is regarded as the most common form of dementia, accounting for an estimated 60-80% of dementia cases [1]. Although the majority of individuals with dementia are over 65 years old [5], other types, such as Parkinson's disease, Huntington's disease, and vascular dementia also affect a significant number of people [2]. Globally, the annual cost of supporting individuals with AD alone is estimated to be around 605 billion USD, illustrating the considerable financial strain tied to condition [2].

Research typically categorize individuals into three groups: Cognitive normal (CN), Mild cognitive Impairment (MCI) with Alzheimer's disease (AD). Patients with MCI share some symptoms with AD patients but can generally carry out their daily routines without significant difficulty. An individual with MCI may remain cognitively stable for several years before developing a form of dementia, most often Alzheimer's disease. Each year, an estimated 10-

∗Corresponding author: Nasrallah Asem AL-Sultani

Email addresses: *cm.post23.17@qu.edu.iq*

Communicated by 'sub etitor'

30% of patients with MCI progress to Alzheimer's disease, while only about 1-2% of cognitively normal individuals develop MCI AD [3] [4].

Recent findings indicate that initiating treatment for MCI at an early stage can help slow its progression to AD [5]. Consequently, investigating and anticipating MCI in its initial phase is of significant importance.

Medical imaging helps detect degenerative tissue changes associated with neurological disorders. These changes can manifest long before clinical symptoms of the disease appear [6]. The most popular imaging techniques in this context are: magnetic resonance imaging (MRI), single-photon emission computed tomography (SPECT), and positron emission tomography (PET). Although most research efforts have focused on the use of CNNs as [7-11]. The Vision Transformers (ViT) models have received extensive attention in the field of image processing due to their high scalability and computational efficiency, which have also proven effective for image classification tasks [12]. The ViT architecture for solving computer vision (CV) problems is based on the transformer architecture. A notable feature of this model is its reliance on an attention mechanism in the processing process[13].

Models based on the ViT architecture have shown significant outperformance over CNNs on the ImageNet dataset [14], [15]. They have also held the lead on several other image datasets [16]. Although ViT models perform well on computer vision tasks, their large-scale architecture requires massive amounts of data for training, along with higher computational requirements compared to their CNN counterparts. This limits their use. The self-attention mechanism in the ViT effectively captures the connections between distant brain regions. Given that the brain is an intricate network with significant connections between distant regions, this feature makes the ViT better suited for processing MRI images [17].

This survey is organized as follows: Section 1 introduces Alzheimer's disease and emphasizes the importance of early diagnosis. Section 2 reviews vision transformers models for AD detection. Section 3 focuses on recent development in vision transformer, Section 4 a discussion. Section 5 identifies the key challenges and limitations, Section 6 outlines future research direction and recommendations, while Section 7 concludes with a conclusion.

2. Alzheimer’s Disease Detection Using Vision Transformer

Vision Transformers (ViTs) represent a paradigm shift in the field of computer vision by leveraging the transformer framework. Unlike traditional convolution based models, ViTs can capture long-range dependencies in visual data, allowing them to excel in tasks such as segmentation, object detection, and image classification, especially when trained on large-scale datasets. Numerous optimizations have been introduced to improve their accuracy and efficiency over time. A thorough analysis of research and developments in the use of Vision Transformers is given in Table 1.

In [18], the authors proposed a cross-domain transfer learning strategy that leverages insights from natural images to address the issue of limited brain imaging data. They used the ImageNet-21K dataset to train the Vision Transformer (ViT). They also developed a spline-based convolutional embedding method that improved the standard fitting process of vanilla ViT. When fine-tuning (FT) was performed on the dataset, the ViT model had the best accuracy (96.8%) and the best F1 score (94.9%). They also conducted a series of tests to see how the sample size of the CN/AD samples affected the model's performance. The results demonstrate that their method works well for using information from real-world images to improve brain scans. Furthermore, the researchers noted that the enhanced model not only outperformed traditional techniques but also exhibited greater robustness against variations in image quality. This advancement could pave the way for more accurate diagnostic tools in medical imaging and ultimately improve patient outcomes.

Z Zilun et al. [19] presented two models for 3D-MRI classification. The first model, CVVT, is inspired by the ViT architecture and modified to receive 3D inputs and explore non-local relationships between different parts of the brain. The second model, convNet3D, is characterized by its simple architecture. The result showed that the second model outperformed, achieving an accuracy of 98%, while the CVVT model's accuracy was 86%. This indicates that a simple model has high performance in the presence of scarce data.

Saman Sarraf et al. [20] proposed an improved vision transformer architecture called the Optimized Vision Transformer (OVT), This architecture aimed to reduce the number of trainable parameters compared to the vanilla transformer, while maintaining its effectiveness in extracting relevant patterns from the dataset. Their model

outperforms most CNN-based deep learning models, with a 30% reduction in the number of trainable parameters compared to the original ViT model. The model obtained 97% F1 scores. However, the study relies on a complex data-processing pipeline that may incur information loss, and the model's performance is highly dependent on the accuracy of this processing, making it sensitive to any glitches in its steps.

In [21] a recent study by Shin and colleagues tested the effectiveness of the ViT model in analyzing brain PET using the radioactive substance 18F-florbetaben, which is used to detect beta-amyloid protein accumulation, an early indicator associated with Alzheimer's disease. ViT gave more accurate results than the VGG19 model when used in binary classification (AD and CN), but did not perform as well in ternary classification (AD, CN, MCI), achieving an accuracy of 80%. The study also indicated that using data augmentation techniques through image rotation did not improve the model's class imbalance. The researchers recommended the use of more diverse data and improved model hyperparameters to achieve more accurate performance.

In [22], Mohammad H. Alshayeji proposed a framework using Vision Transformers (ViTs) for precise stage identification and accurate AD diagnosis. The illness was divided into four categories. (non-affected, very mild dementia, mild dementia, and moderate dementia). They used transfer learning, applying effective strategies to address the data imbalance problem using data augmentation techniques. The dataset consisted of 6400 MRI scans from Kaggle. 99.88% specificity, 99.69% sensitivity, and 99.83% accuracy.

Maram Almufareh et al. In [23] suggested Alzheimer's disease (AD) diagnosis from MRI pictures with the use of an attention-driven mechanism that uses the Vision Transformer approach. The MRI images must first undergo preprocessing as part of the suggested process before being sent into the Vision Transformer for categorization. Using dataset from OASIS1, which comprises 80,000 MRI pictures and classified into four phases. Clinical Dementia Rating (CDR) scores and the relevant information were used to categorize the patients. By attaining 99.06% accuracy, the suggested approach demonstrated exceptional performance.

Khatri et al. [24] proposed a self-learning model using a vision transformer (ViT) to automatically extract features from unlabeled FDG-PET images. The model relied on the DINO algorithm in the pre-training phase, along with an ELM classifier, and achieved state-of-the-art performance with an accuracy of 92.31%. The model also provided interpretability by identifying brain regions that influence the classification. However, the study noted potential diagnostic overlap due to similar PET patterns in other dementias such as DLB and FTD, and was limited to MCI without including other groups, which may restrict generalizability. The authors recommended incorporating multimodal imaging data such as fMRI and sMRI in future work to enhance predictive accuracy.

Marwa Zaabi et al, in [25], have proposed using Vision Transformers (ViTs) combined with Transfer Learning. Their approach leverages the transformer's ability to capture temporal correlations between different image patches. As an initial step, they focused on the most crucial brain regions containing the hippocampus, rather than the entire image. Their investigation showed that Transformers significantly outperformed Convolutional Neural Networks (CNNs). Using the OASIS dataset, which includes 250 brain scans, they tested two transformer models, ViT-B16 and ViT-B32, with ViT-B32 outperforming ViT-B16 by 1%. Transfer learning with transformer models outperformed transfer learning with CNN models by 4% and traditional CNN models without transfer learning by 8%. The ViT-B32 model attained an accuracy of 96.25%, while CNN with transfer learning reached 92.86%, and CNN without transfer learning achieved 88.10%.

In [26]. Odusami and colleagues presented a new approach to handling Alzheimer's disease medical imaging data by processing it using a hybrid approach that combines frequency analysis techniques and deep learning models. They applied an unconventional approach to fusion MRI and PET images at the frequency level using wavelet transform, then transformed these fused data into unified images that were fed directly into a vision transformer model. What distinguishes this work is not only the precise fusion mechanism but also its ability to leverage the complementary properties of both MRI and PET without the need to train the model from scratch. They achieve this by optimizing a pre-trained ViT with fine-tuning. In addition, the study lacks a comparative analysis with existing modern methods.

In [27]. Uttam Khatri et al. Have proposed the RMTnet (Recurrence, Encounter, and Transformer Network) model, where a comprehensive model was designed between Vision Transformers (ViT). This approach uses cross-domain

transfer learning techniques, where a slow stream provides contextual information to a fast stream, which is processed through an attention bottleneck. The integration of regression neural networks (RNNs) allows the ViT to focus on local details. When evaluated in an AD/CN classification task using FDG-PET data, positron emission tomography (PET) scans measure decreased glucose concentration in the temporoparietal association cortex. The proposed approach achieved classification results comparable to leading studies, with an accuracy of 91.08%.

Xin Xing et al. [28] presented a model called ADVIT, which used PET images from two different types (PET-AV45 and PET-FDG) to diagnose the disease. Unlike traditional methods based on 3D convolution networks (3D CNNs), The researchers replaced the underlying structure with vision transformer, leveraging their ability to capture global relationships between image segments through a self-attention mechanism. To reduce the compulation complexity association with 3D images, a transformation module was developed that projects 3D images into 2D images, allowing the use of ViT pre-trained on the ImageNet dataset. The model demonstrated superior performance over 3DCNN model in terms of accuracy (91.34%) and area under the curve (AUC) (95.22%). the study also demonstrated that using multimodality data enhances diagnostic accuracy compared to using only a single modality.

Pinky Sherwani et al. [29] suggested work three distinct vision transformers: Class Attention in Image Transformer (CaiT), Deep Vision Transformer (DeepViT), and Vanilla Vision Transformer (Vanilla ViT).The results demonstrated that ViTs exhibit an inductive bias that depends more on model regularization or data augmentation (AgReg). The ViT models surpassed CNNs in both accuracy and computational efficiency by a factor of four, with DeepViT achieving the highest accuracy of 90.2%. This superior performance is attributed to DeepViT's Reattention mechanism, which regenerates attention maps to reduce similarities and increase diversity across layers, thereby improving image classification accuracy. Consequently.

Nikhil J. Dhinagar et al. In [30] evaluated several variants of the ViT architecture for various neuroimaging tasks, focusing on sex classification and AD classification using 3D images. AUCs for sex classification and AD classification were 98.7% and 89.2%, respectively. The models were assessed using OASIS3, ADNI, and UKBB data. The researchers found that fine-tuning the ViT models led to enhance performance 5% in models pre-trained on actual MRI images and 9–10%, for models using synthetic MRI data.

Ramesh Poonia et al. In [31] proposed an ensemble model that combines transfer learning techniques (VGG19, ResNet50, InceptionV3) and Vision Transformer, explainable AI (EXAI). Their results demonstrated that the collaborative model outperformed traditional models, with the combined InceptionV3+ViT model achieving superior accuracy (96%) compared to the maximum accuracy achieved by the individual models. The use of Grad-CAM also helped identify the brain regions most influential in the classification process, enhancing the transparency of the model. The results confirm the effectiveness of their model and suggest that incorporating further machine learning techniques could improve prediction accuracy in the future.

Shah, S. M. et al in [32] proposed an architecture called the Bi-Vision Transformer (BiViT), designed to classify stages of Alzheimer's disease using two-dimensional MRI (2D MRI) images.. Mutual Latent Fusion (MLF) and Parallel Paired Encoding Strategy (PCES), two innovative modules of the BiViT architecture, are intended to improve feature learning. A pair of datasets were employed. The initial dataset included mild and moderate dementia, among other phases of AD. Samples from individuals with AD and other cognitive problems, such as mild, early, and moderate impairment, were included in the second dataset, on the other hand. For comparison, a deep autoencoder and several transfer learning algorithms were trained on both datasets. The BiViT model obtained 96.38% accuracy, 97.87% recall, 88.28% precision, 92.84% F1 score. When applied to the dataset of cognitive disorders, however, the accuracy decreased to less than 96%, most likely as a result of the smaller dataset and uneven data representation.

Yue Yin et al. In [33] proposed the SMIL-DeiT network, which employs the DeiT architecture as the backbone vision transformer model and integrates a MIL (Multiple Instance Learning) head along with a multi-layer perceptron (MLP) in the classification head. The Vision Transformer is the core structure of their model. Multiple Instance-Learning (MIL) was used to complete the final classification job, whereas DINO, a self-supervised learning technique, was used for pre-training. The model was evaluated using the ADNI dataset, with accuracy being the primary focus. With an accuracy of 93.2%, their approach outperformed CNN's 90.8% and the Vision Transformer's 90.1%. The imbalance in the dataset affect performance cognitive disorders, and the complex nature of the model made it difficult to interpret decision-making, an issue that requires improvement for practical healthcare applications.

Maryam Akhavan et al. [34] presented a new classification model based on a pre-trained hierarchical vision transformer (PVT). This model is characterized by its ability to extract fine-grained and broad spatial features within a unified architectural framework, giving it an advantage over traditional methods based on CNNs or ViTs. They achieved an accuracy of 97.7%. The study recommends further development of the model using advanced optimization algorithms, such as stochastic optimization and multi-objective search techniques, to reduce the risks associated with overlearning and achieve more stable and reliable performance in future clinical applications.

Fei Huang et al. In [35] propose a new method known as the Monte Carlo Ensemble Vision Transformer (MC-ViT), which combines a single Vision Transformer (ViT) model with an ensemble strategy. Instead of relying on multiple learners as in traditional ensemble methods, their approach leverages Monte Carlo sampling that generates a wide range of classification outcomes. This enhances the performance of the MC-ViT. This technique effectively addresses the limitations Only a part of the brain's structure is captured by 3D convolutional neural networks (CNNs), opening the door for a neural network that can recognize connections between 3D features. When evaluated on the ADNI dataset with 7,199 scans and the OASIS-3 dataset with 1,992 scans, this method achieved 90% accuracy in Alzheimer's disease classification with minimal preprocessing, outperforming both 2D-slice CNNs and 3D CNNs.

Fei Huang et al. [35] proposed the Monte Carlo Ensemble Vision Transformer (MC-ViT), a framework that enhances AD classification by leveraging Monte Carlo sampling to generate multiple perturbed versions of a single Vision Transformer (ViT) model's input, producing a diverse set of predictions that are aggregated to improve robustness and reduce overfitting without requiring multiple independent models. This method addresses the limited receptive field of 3D convolutional neural networks (CNNs) by enabling better capture of spatial dependencies in neuroimaging data. Evaluated on the ADNI dataset with 7,199 scans and the OASIS-3 dataset with 1,992 scans, MC-ViT achieved approximately 90% accuracy with minimal preprocessing, outperforming traditional 2D-slice and 3D CNN approaches.

In [36], the authors used CNN and ViT and applied the SMOTE algorithm to address the imbalanced sample distribution to learn and classify brain magnetic resonance imaging (MRI) of Alzheimer's disease (AD) using the public Alzheimer's disease dataset Kaggle. They then compared the accuracy of three models: CNN, ViT, and CNN + ViT combined on the Alzheimer's disease dataset. In the end, their proposed model achieved 90.95% accuracy and Recall 91.81%.

Table 1. Vision Transformer for Alzheimer's Disease Detection

Ref / Cite	Data Type	DL	Approaches	Effectiveness	Limitation
Lyu, Y et-al.2022 [18]	MRI	VIT	A cross-domain transfer learning method using Vision Transformer (ViT) pre-trained on ImageNet-21K and fine-tuned on brain imaging data.	Accuracy 96.8%, F1-scores 94.9%	Limited generalization on clinical data
Zilun Zhang et al, 2022 [19]	3D MRI	CNN& ViT	Proposed two models: Convolutional Voxel Vision Transformer (CVVT) and ConvNet3D-4, a shallow 3D CNN-based model.	ConvNet3D-4 achieved 98% accuracy, and CVVT's 86%.	The CVVT model has relatively low performance compared to traditional CNN models in 3D image processing.
Saman Sarraf et al 2023 [20]	fMRI & MRI	OViT	The ViT architecture has been	F1-scores of 97%	Heavily dependent on complex preprocessing that may introduce information loss

			modified to reduce complexity while maintaining efficiency in feature extraction.		
Hyunji Shin 2023 [21]	PET	VGG19&ViT	Comparison between ViT and VGG19 for binary and ternary classification.	Accuracy: 80% with vit, 56.67% (Ternary classification) with VGG19	Poor performance on multi-class classification
Alshayeji, M. H. 2024 [22]	MRI	ViT	Transfer learning and data augmentation to handle imbalance.	Accuracy 99.83% Sensitivity 99.69% Specificity 99.88%	The model has not been evaluated on external datasets, which weakens confidence in its generalizability.
Maram Almufareh et al 2023 [23]	MRI	ViT	Have proposed an attention-based mechanism using a vision transformation approach	Accuracy 99.06% F1-score 99.1%	Limited generalizability
Uttam Khatri et al 2023[24]	PET	ViT	The method merges Vision Transformer (ViT) with DINO for feature extraction and classifies using Extreme Learning Machine (ELM).	Accuracy 92.31%, specificity 90.21%, sensitivity 95.50%	Diagnostic overlap with other dementias; focusing only on the transition of MCI to AD,
Marwa Zaabi et al 2024 [25]	MRI	ViT-B16 ViT-B32	Comparison between ViT variants and CNNs; ViT-B32 showed superior results.	ViT-B32 achieved 96.25% accuracy,	Using data containing 250 scans, only the hippocampus was analyzed without other brain regions.
Modupe Odusami et al 2023 [26]	sMRI& PET	ViT+ Wavelet Transformer	A multimodal fusion approach	MRI accuracy: 81.25%, PET accuracy: 93.75%, Fused (MRI+PET): AD/EMCI 98.50%, AD/LMCI 99.58%.	The DWT fusion technique is not optimized enough, which may affect the image quality.
Uttam Khatri et al 2024 [27]	FDG-PET	RMTnet (ViT+RNN)	Augmenting ViT with RNN to improve representation and efficiency	Accuracy 91.08%	Limited data volume
Xin Xing et al 2023 [28]	(PET-AV4 ,PETFDG)	ADViT	Have proposed use the 3D-2D module to enable ViT to process 3D data effectively.	Accuracy 91.34% AUC 95.22%	Loss of medical details when converting 3D images to 2D, and the data used is limited.
Pinky Sherwani et al 2023 [29]	MRI	Vanilla ViT; DeepViT, CaiT	Evaluation of three models on a small dataset.	DeepViT 90.2%. Vanilla ViT 56.41% CaiT128-16 80.27%	Sensitivity of the model to parameters it requires careful tuning of parameters.
Nikhil J. Dhinagar et al 2023 [30]	3D MRI	ViT	Focuses on pre-training, data augmentation, and learning rate strategies to enhance model performance with limited data.	AUC of 98.7% for sex & 0.892 for AD	Using synthetic images in training may affect the realism of the results, requiring additional testing on real data.

				classification	
Ramesh Chandra Poonia et al 2023 [31]	MRI	ViT & CNNs models	An ensemble model that combines different ViT learning models with explainable AI.	Accuracy 96%, F1-score 92 %	Need to improve interpretability and explainability. Increasing computational complexity and resource requirements.
Shah, S et al 2024 [32]	2D MRI	BiViT	The transformer consists of two novel modules: Mutual Latent Fusion (MLF) and Parallel Coupled Encoding Strategy (PCES) for effective feature learning.	Accuracy 96.38%, Recall 97.87% Precision 88.28% F1-score 92.84%	BiViT relies on a complex hybrid architecture that requires precise parameter tuning, making it difficult to use without advance expertise and extensive experimentation.
Yue Yin et al 2022 [33]	MRI	SMIL- DeiT	they proposed SMIL-DeiT network, which use DeiT as the vision Transformer model and use MIL Head together with MLP in the classification head to form the classification head.	Accuracy 93.2% Precision 93.2% Recall 92.6% F1-score 92.8%	Information loss from 3D images due to conversion to 2D images, model complexity, limited comparability, and a lack of interpretability.
Maryam Akhavan et al 2024 [34]	sMRI	PVTAD	This approach combines the strengths of CNNs and ViT to capture both local and global features of AD from WM coronal middle slices.	Accuracy 97.7% Sensitivity 97.15% Specificity 98.16% F1-score 97.6%	The problem of the imbalance in the number of samples between the two categories (AD and CN) was not addressed. Furthermore, the classification was limited to only two cases without distinguishing between the different stages of disease progression.
Fei Huang et al 2024 [35]	MRI	MC- ViT	Suggest a novel method known as the Monte Carlo Ensemble Vision Transformer (MC-ViT), which combines a single Vision Transformer (ViT) model with an ensemble approach.	Accuracy 90%	High computational requirements for Monte Carlo
Xing Mu et al 2023 [36]	MRI	CNN+ ViT	Combined CNN and ViT models with the SMOTE technique to address data imbalance.	Accuracy 97.43%	Still affected by data imbalance despite using SMOTE

3. Alzheimer's Disease Detection Using Optimized Vision Transformer

Several studies have relied on metaheuristic algorithms to improve deep learning models through feature selection and hyperparameter tuning, which accelerates convergence and determines optimal weights and biases. Combining Vision Transformers with these algorithms is a promising approach to enhance the accuracy of Alzheimer's disease

detection. However, given the recent emergence of ViT in medical classification, studies on these two techniques are still extremely rare. Table 2 lists all systematic reviews that have addressed the improvement of deep learning models using metaheuristic algorithms and the application of ViT in medical classification.

Anuvab Sen et al.[37] In have proposed using ViT model-based metaheuristic algorithms for the detection of dementia at various stages. The DE, GA, ACO, and PSO metaheuristic algorithms were used. To choose hyper-parameters and categorize patient data into AD, MCI, and HC groups, metaheuristic algorithms (MH) are employed. The ADNI dataset was the one in question. The suggested approach using the Differential Evolution algorithm (DE), the best metaheuristic algorithm, yields 96.8% accuracy, 94% recall, 95% precision, and 96% F1 score.

In [39], Authors Andrew Becker and collaborators propose the NSGA-ViT model, using Neural Architecture Search (NAS) to combine convolutional layers with multi-head self-attention for improved performance and computational complexity. NSGA-ViT achieves 92.35% accuracy on CIFAR-10, outperforming the 8-head ViT by 12%. However, on ImageNet, NSGA-ViT underperforms ViT, likely due to fewer pixels processed by each attention head.

Jiangning Zhang et al.[40] In introduced a novel Vision Transformer model inspired by evolutionary algorithms to enhance image classification. By integrating both global and local interactions and aggregating information from multiple scales, the model achieves a more nuanced understanding of image details. Key components include the Global and Local Interaction (GLI) module and the Multi-Scale Region Aggregation (MSRA) module, which together improve the model's accuracy significantly. On the ImageNet-1K dataset, the EATFormer-Base model achieved an accuracy of 83.9%, showcasing its competitive performance.

TABLE 2. Optimized Vision Transformer for Alzheimer's Disease Detection

Ref / Cite	Dataset	Data Type	DL	Meta-heuristic Algorithm	Approaches	Effectiveness
Anuvab Sen et al 2023 [37]	ADNI	MRI	ViT	DE GA PSO ACO	have proposed using ViT model-based metaheuristic algorithms for the detection of dementia at various stages	Accuracy with DE 96.8% Accuracy with GA 91% Accuracy with PSO 92% Accuracy with ACO 94%
Becker, A. 2023[39]	CIFAR-10, ImageNet-1k	Low-resolution image classification. & High resolution image classification.	Deep Learning (DL)	Non-dominated Sorting Genetic Algorithm II (NSGA-II)	design a network architecture that combines Convolutional Layers and Multi-head Self-Attention.	CIFAR-10, NSGA-ViT achieves up to 92.95% accuracy with 44M parameters, outperforming ViT
Zhang et al 2024 [40]	ImageNet-1K	Image	ViT	EA	The "EATFormer" model enhances Vision Transformers using concepts from	accuracy of 83.9%

					evolutionary algorithms.	

4. Discussions

The application of Vision Transformer (ViT) models is an example of technical progress in the field of medical image analysis using deep learning techniques. [17] demonstrated that adopting transfer learning strategies with the ViT model. Which is initially trained on the ImageNet database and then recalibrated to fit medical imaging data, resulted in a classification accuracy of 96.8, as confirmed by the results reported in [24]. In addition, demonstrates the model's ability to capture distant relationships between brain regions using the self-attention mechanism. However, converting 3D data to a 2D representation to reduce computational complexity, as reported in [27], may result in the loss of some important spatial information, posing a challenge in maintaining the accuracy of the results while reducing the computational burden.

While studies based on positron emission tomography (PET) images have shown lower accuracy compared to other imaging techniques, this is primarily due to its limited spatial resolution, as well as technical and physiological challenges, as described in [24, 27, 28]. Therefore, it is often recommended to combine PET with other imaging methods, such as MRI, to achieve more accurate and reliable results. Additionally, large datasets, such as those available on the Kaggle dataset, often yield high accuracy; however, this performance is typically obtained under idealized conditions and may not reflect real-world clinical settings. This highlights the importance of validating models on diverse and representative datasets to ensure their generalizability.

While recent efforts have focused on improving the performance of VIT by combining it with various deep learning models or optimization algorithms to improve Alzheimer's disease (AD) classification. Selecting an appropriate deep learning architecture is essential for reliable early detection. Optimization techniques such as Differential Evolution (DE), Particle Swarm Optimization (PSO), and Ant Colony Optimization (ACO) have shown promise in enhancing the performance of Vision Transformer (ViT) models. As reported in [37], integrating these algorithms has led to improvements in precision and recall, highlighting the importance of combining deep learning with optimization strategies to address training variability and enhance model stability.

In conclusion, AI models face a significant challenge in generalizing their results to different data sets outside of traditional training sets such as ADNI and OASIS, which reduces their reliability in real-world clinical use. Therefore, it is essential to test these models on diverse data from different environments, in addition to improving their interpretation capabilities to help clinicians understand their decisions and increase confidence in their diagnosis.

5. Challenges and Limits

- **Limited Medical Data:** Obtaining brain images classified across all disease stages faces ethical and medical challenges (preserving patient privacy and a limited number of participants). Some studies suffer from limited or unbalanced data samples, which may affect the generalizability of the results and make the model more susceptible to bias.

- **Computational Complexity:** ViT models require significant computational resources due to their complexity and large number of parameters. This poses a barrier to their application in environments that lack advanced infrastructure.

- **Complexity of Data Processing:** A fundamental challenge facing the application of Vision Transformer techniques in medical image analysis is the complexity of processing raw data to ensure that critical features are highlighted without losing essential information. This increases the complexity of the computational process and requires extensive processing expertise, especially for brain images.

- **Limited External Generalization Tests:** Many of the proposed models have not been tested on data from external sources or on people from different backgrounds. There is still a pressing need to ensure the robustness of performance in real clinical settings.

6. Future Work

- **Developing hybrid models combining ViT and CNN:** Design a binary model containing CNN layers to extract spatial features, followed by ViT layers to process contextual relationships, then fuse the outputs using an attention-based fusion mechanism or BiFusion modules.

- **Improving ViT using adaptive metaheuristic algorithms:** Enhance generalization and reduce overlearning by dynamically tuning hyperparameters. Novel optimization algorithms, such as Hippopotamus Optimization Algorithm or other recent algorithms, can be explored.

- **Multimodal imaging fusion:** To improve diagnostic accuracy by combining fMRI, PET, and sMRI images to obtain more comprehensive representations, the ViT-based model, supported by a late fusion or co-attention fusion is used to process each type of data separately and then fuse them at a later stage to preserve temporal and spatial coherence.

- **Enabling physicians to understand model decisions:** By clarifying the brain regions that influence the decision, we are introducing interpretive intelligence (XAI) techniques to increase transparency, such as Attention Rollout, which contributes to transforming experimental results into practical clinical tools.

- **Designing 3D ViT models:** Future efforts may focus on developing a dedicated 3D Vision Transformer model that leverages 3D attention mechanisms or incorporates spatio-temporal encoding units to preserve spatial structure between slices.

7. Conclusion

This paper has thoroughly reviewed the techniques and methods used in Alzheimer's Disease (AD) detection. The results indicated that vision transformer models, especially enhanced transformers (OViTs), achieved exceptional performance on key datasets. The OViTAD model achieved an F1 rate of 97% on the ADNI dataset, and when ViT was optimized using metaheuristic algorithms to adjust hyperparameters, an accuracy of 96.8% was achieved using the differential evolution (DE) algorithm. Similarly, attention-driven mechanisms applied to OASIS-1 data (80,000 MRI) demonstrated outstanding results, obtaining a 99.06% accuracy rate and a 99.1% F1 score. The review shows that ViTs are better at dealing with complicated imaging data because their self-attention mechanisms can effectively recognize important details over long distances in brain MRI images. This review highlights the advances in vision transducer technologies (ViTs) in diagnosing Alzheimer’s disease.

REFERENCES

[1] A. L. Sosa-Ortiz, I. Acosta-Castillo, and M. J. Prince, “Epidemiology of Dementias and Alzheimer’s Disease,” Arch. Med. Res., vol. 43, no. 8, pp. 600–608, Nov. 2012, doi: 10.1016/j.arcmed.2012.11.003.

[2] E. R. Danielsen, Magnetic Resonance Spectroscopy Diagnosis of Neurological Diseases, 1st ed. CRC Press, 1999. doi: 10.1201/9781482270105.

[3] P. Celsis, “Age-related cognitive decline, mild cognitive impairment or preclinical Alzheimer’s disease?,” Ann. Med., vol. 32, no. 1, pp. 6–14, Jan. 2000, doi: 10.3109/07853890008995904.

[4] R. C. Petersen, "Mild Cognitive Impairment:," Contin. Lifelong Learn. Neurol., vol. 22, no. 2, Dementia, pp. 404–418, Apr. 2016, doi: 10.1212/CON.0000000000000313.

[5] F. T. Hane, M. Robinson, B. Y. Lee, O. Bai, Z. Leonenko, and M. S. Albert, "Recent Progress in Alzheimer's Disease Research, Part 3: Diagnosis and Treatment," J. Alzheimer's Dis., vol. 57, no. 3, pp. 645–665, Apr. 2017, doi: 10.3233/JAD-160907.

[6] T. Gómez-Isla, J. L. Price, D. W. McKeel Jr., J. C. Morris, J. H. Growdon, and B. T. Hyman, "Profound Loss of Layer II Entorhinal Cortex Neurons Occurs in Very Mild Alzheimer's Disease," J. Neurosci., vol. 16, no. 14, pp. 4491–4500, Jul. 1996, doi: 10.1523/JNEUROSCI.16-14-04491.1996.

[7] A. Khvostikov, K. Aderghal, J. Benois-Pineau, A. Krylov, and G. Catheline, "3D CNN-based classification using sMRI and MD-DTI images for Alzheimer disease studies," 2018, arXiv. doi: 10.48550/ARXIV.1801.05968.

[8] A. W. Salehi, P. Baglat, B. B. Sharma, G. Gupta, and A. Upadhya, "A CNN Model: Earlier Diagnosis and Classification of Alzheimer Disease using MRI," in 2020 International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India: IEEE, Sep. 2020, pp. 156–161. doi: 10.1109/ICOSEC49089.2020.9215402.

[9] E. N. Marzban, A. M. Eldeib, I. A. Yassine, Y. M. Kadah, and for the Alzheimer's Disease Neurodegenerative Initiative, "Alzheimer's disease diagnosis from diffusion tensor images using convolutional neural networks," PLOS ONE, vol. 15, no. 3, p. e0230409, Mar. 2020, doi: 10.1371/journal.pone.0230409.

[10] A. Punjabi, A. Martersteck, Y. Wang, T. B. Parrish, A. K. Katsaggelos, and and the Alzheimer's Disease Neuroimaging Initiative, "Neuroimaging modality fusion in Alzheimer's classification using convolutional neural networks," PLOS ONE, vol. 14, no. 12, p. e0225759, Dec. 2019, doi: 10.1371/journal.pone.0225759.

[11] G. W. Kidder and C. W. Montgomery, "Oxygenation of frog gastric mucosa in vitro," Am. J. Physiol., vol. 229, no. 6, pp. 1510–1513, Dec. 1975, doi: 10.1152/ajplegacy.1975.229.6.1510.

[12] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," 2020, arXiv. doi: 10.48550/ARXIV.2010.11929.

[13] X. Li et al., "Deep Learning Attention Mechanism in Medical Image Analysis: Basics and Beyonds," Int. J. Netw. Dyn. Intell., pp. 93–116, Mar. 2023, doi: 10.53941/ijndi0201006.

[14] S. Cuenat and R. Couturier, "Convolutional Neural Network (CNN) vs Vision Transformer (ViT) for Digital Holography," in 2022 2nd International Conference on Computer, Control and Robotics (ICCCR), Shanghai, China: IEEE, Mar. 2022, pp. 235–240. doi: 10.1109/ICCCR54399.2022.9790134.

[15] M. Filipiuk and V. Singh, "Comparing Vision Transformers and Convolutional Nets for Safety Critical Systems.," in Safeai@ aaai, 2022. Accessed: Apr. 03, 2025. [Online]. Available: https://ceur-ws.org/Vol-3087/paper_31.pdf

[16] X. Chen, S. Xie, and K. He, "An Empirical Study of Training Self-Supervised Vision Transformers," in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada: IEEE, Oct. 2021, pp. 9620–9629. doi: 10.1109/ICCV48922.2021.00950.

[17] L. Yuan et al., "Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet," in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada: IEEE, Oct. 2021, pp. 538–547. doi: 10.1109/ICCV48922.2021.00060.

[18] Y. Lyu, X. Yu, D. Zhu, and L. Zhang, "Classification of Alzheimer's Disease via Vision Transformer: Classification of Alzheimer's Disease via Vision Transformer," in Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments, Corfu Greece: ACM, Jun. 2022, pp. 463–468. doi: 10.1145/3529190.3534754.

[19] Z. Zhang and F. Khalvati, "Introducing Vision Transformer for Alzheimer's Disease classification task with 3D input," 2022, arXiv. doi: 10.48550/ARXIV.2210.01177.

[20] S. Sarraf, A. Sarraf, D. D. DeSouza, J. A. E. Anderson, M. Kabia, and The Alzheimer's Disease Neuroimaging Initiative, "OViTAD: Optimized Vision Transformer to Predict Various Stages of Alzheimer's Disease Using Resting-State fMRI and Structural MRI Data," Brain Sci., vol. 13, no. 2, p. 260, Feb. 2023, doi: 10.3390/brainsci13020260.

[21] H. Shin, S. Jeon, Y. Seol, S. Kim, and D. Kang, "Vision Transformer Approach for Classification of Alzheimer's Disease Using 18F-Florbetaben Brain Images," Appl. Sci., vol. 13, no. 6, p. 3453, Mar. 2023, doi: 10.3390/app13063453.

[22] M. H. Alshayeji, "Alzheimer's disease detection and stage identification from magnetic resonance brain images using vision transformer," Mach. Learn. Sci. Technol., vol. 5, no. 3, p. 035011, Sep. 2024, doi: 10.1088/2632-2153/ad5fdc.

[23] M. F. Almufareh, S. Tehsin, M. Humayun, and S. Kausar, "Artificial Cognition for Detection of Mental Disability: A Vision Transformer Approach for Alzheimer's Disease," Healthcare, vol. 11, no. 20, p. 2763, Oct. 2023, doi: 10.3390/healthcare11202763.

[24] 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, Oct. 2023, doi: 10.3390/bioengineering10101225.

[25] M. Zaabi, M. I. Khedher, and M. A. El-Yacoubi, "Improving Alzheimer's Diagnosis Using Vision Transformers and Transfer Learning," in 2024 16th International Conference on Human System Interaction (HSI), Paris, France: IEEE, Jul. 2024, pp. 1–6. doi: 10.1109/HSI61632.2024.10613527.

[26] M. Odusami, R. Maskeliūnas, and R. Damaševičius, "Pixel-Level Fusion Approach with Vision Transformer for Early Detection of Alzheimer's Disease".

[27] U. Khatri and G.-R. Kwon, "RMTnet: Recurrence meet Transformer for Alzheimer's disease diagnosis using FDG-PET," Jan. 22, 2024, Preprints. doi: 10.36227/techrxiv.170594596.62106929/v1.

[28] X. Xing, G. Liang, Y. Zhang, S. Khanal, A.-L. Lin, and N. Jacobs, "Advit: Vision Transformer On Multi-Modality Pet Images For Alzheimer Disease Diagnosis," in 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), Kolkata, India: IEEE, Mar. 2022, pp. 1–4. doi: 10.1109/ISBI52829.2022.9761584.

[29] P. Sherwani, P. Nandhakumar, P. Srivastava, J. Jagtap, V. Narvekar, and H. R, "Comparative Analysis of Alzheimer's Disease Detection via MRI Scans Using Convolutional Neural Network and Vision Transformer," in 2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF), Chennai, India: IEEE, Jan. 2023, pp. 1–9. doi: 10.1109/ICECONF57129.2023.10084260.

[30] N. J. Dhinagar, S. I. Thomopoulos, E. Laltoo, and P. M. Thompson, "Efficiently Training Vision Transformers on Structural MRI Scans for Alzheimer's Disease Detection," in 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Sydney, Australia: IEEE, Jul. 2023, pp. 1–6. doi: 10.1109/EMBC40787.2023.10341190.

[31] R. C. Poonia and H. A. Al-Alshaikh, "Ensemble approach of transfer learning and vision transformer leveraging explainable AI for disease diagnosis: An advancement towards smart healthcare 5.0," Comput. Biol. Med., vol. 179, p. 108874, Sep. 2024, doi: 10.1016/j.compbiomed.2024.108874.

[32] S. M. A. H. Shah, M. Q. Khan, A. Rizwan, S. U. Jan, N. A. Samee, and M. M. Jamjoom, "Computer-aided diagnosis of Alzheimer's disease and neurocognitive disorders with multimodal Bi-Vision Transformer (BiViT)," Pattern Anal. Appl., vol. 27, no. 3, p. 76, Sep. 2024, doi: 10.1007/s10044-024-01297-6.

[33] Y. Yin, W. Jin, J. Bai, R. Liu, and H. Zhen, "SMIL-DeiT:Multiple Instance Learning and Self-supervised Vision Transformer network for Early Alzheimer's disease classification," in 2022 International Joint Conference on Neural Networks (IJCNN), Padua, Italy: IEEE, Jul. 2022, pp. 1–6. doi: 10.1109/IJCNN55064.2022.9892524.

[34] M. A. Aghdam, S. Bozdag, and F. Saeed, "Pvtad: Alzheimer's Disease Diagnosis Using Pyramid Vision Transformer Applied to White Matter of T1-Weighted Structural Mri Data," in 2024 IEEE International Symposium on Biomedical Imaging (ISBI), Athens, Greece: IEEE, May 2024, pp. 1–4. doi: 10.1109/ISBI56570.2024.10635541.
[35] F. Huang and A. Qiu, "Ensemble Vision Transformer for Dementia Diagnosis," IEEE J. Biomed. Health Inform., vol. 28, no. 9, pp. 5551–5561, Sep. 2024, doi: 10.1109/JBHI.2024.3412812.
[36] X. Mu et al., "Alzheimer Classification Based on Convolutional Neural Network and Vision Transformer," in 2023 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML), Chengdu, China: IEEE, Nov. 2023, pp. 329–334. doi: 10.1109/ICICML60161.2023.10424819.
[37] A. Sen, U. Sen, and S. Roy, "A Comparative Analysis on Metaheuristic Algorithms Based Vision Transformer Model for Early Detection of Alzheimer's Disease," in 2023 IEEE 15th International Conference on Computational Intelligence and Communication Networks (CICN), Dec. 2023, pp. 200–205. doi: 10.1109/CICN59264.2023.10402213.
[38] A. Becker, "NSGA-VIT: AN EVOLUTIONARY APPROACH TO VISION TRANSFORMER ARCHITECTURE DESIGN".
[39] J. Zhang et al., "EATFormer: Improving Vision Transformer Inspired by Evolutionary Algorithm," Int. J. Comput. Vis., vol. 132, no. 9, pp. 3509–3536, Sep. 2024, doi: 10.1007/s11263-024-02034-6.