ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

Deep vision transformers in neurodegenerative disease diagnosis using ^{18}F-fluorodeoxyglucose positron emission tomography scans and anatomical brain atlas

Pooriya Khorramyar*, Amira Soliman, Farzaneh Etminani, and Stefan Byttner

Center for Applied Intelligent Systems Research in Health (CAISR Health), The School of Information Technology, Halmstad University, Halmstad, Halland, Sweden

(This article belongs to the *Special Issue: Artificial intelligence for diagnosing brain diseases*)

Abstract

This research explores adapting vision transformers (ViTs) to classify neurodegenerative diseases while ensuring their decision-making process is interpretable. We developed a model to classify ^{18}F-fluorodeoxyglucose (^{18}F-FDG) positron emission tomography (PET) brain scans into three categories: cognitively normal (CN), mild cognitive impairment (MCI), and Alzheimer's disease (AD). The dataset utilized in this research contains 580 samples of ^{18}F-FDG PET scans obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI). The proposed model obtained an F1 score of 81% (macro-average of all classes) on the test dataset, a significant performance improvement compared to the literature. Furthermore, we combined the model's attention maps with the Automated Anatomical Atlas 3 (AAL3), which represents a digital brain map, to identify the most influential areas on the model's predictions and to conduct a regions' importance study as a step toward explainability. We demonstrated that ViTs can achieve competitive performance compared to convolutional neural networks while enabling the development of explainable models without extra computations due to the attention mechanism.

Keywords: Vision transformer; Neurodegenerative disease; ^{18}F-FDG PET; Medical image analysis; Brain scan; Deep neural network

***Corresponding author:**
Pooriya Khorramyar
(pookho20@student.hh.se)

Citation: Khorramyar P, Soliman A, Etminani F, Byttner S. Deep vision transformers in neurodegenerative disease diagnosis using 18F-fluorodeoxyglucose positron emission tomography scans and anatomical brain atlas. *Artif Intell Health*. 2025;2(4):33-46. doi: 10.36922/AIH025140026

Received: March 31, 2025

1st revised: April 12, 2025

2nd revised: May 22, 2025

Accepted: May 26, 2025

Published online: June 19, 2025

1. Introduction

Neurodegenerative diseases (NDDs) lead to progressive deterioration and death of neurons, damaging the nervous system and brain. Affecting more than 55 million patients with a yearly increase rate of 10 million new cases worldwide, NDDs are a prominent cause of disability and death.[1] In addition, Alzheimer's disease (AD), as the most widespread form, accounts for 70% of NDD cases and plays a significant role in these statistics.[1] Although NDDs have a heavy impact on healthcare systems and patients' lives, they remain incurable as of today.[1] However, timely diagnosis is pivotal in disease management and improving the patient's quality of life.[2]

Diagnosing NDDs is exceedingly demanding and requires years of training and experience. Hence, according to some studies, it has been estimated that 75% of NDD cases are undiagnosed worldwide due to various reasons, including the diagnosis complexity.[3] Astoundingly, this number rises to 90% in low- and middle-income countries, according to the same analysis.[3] Moreover, the growing number of NDD cases could devastate healthcare systems in coming years, according to this study.[3] Therefore, innovative and affordable methods are needed to assist doctors and decrease this diagnosis gap.

The rapid progress of artificial intelligence (AI) and its sub-fields has led to outstanding results in different domains, including medical image processing. Thus, researchers attempted to harness the power of deep neural networks (DNNs) in diagnosing NDDs and demonstrated that they could have competitive performance compared to human experts.[4,5]

The advent of vision transformers (ViTs) resulted in distinguished performance in various computer vision tasks, surpassing traditional approaches like convolutional neural networks (CNNs).[6] Therefore, their application in NDD diagnosis has been a trending research subject and the focal point of various studies, including this paper. We developed our model based on vanilla ViT, proposed by Dosovitskiy *et al.*[6], and trained it using ^{18}F-fluorodeoxyglucose (^{18}F-FDG) positron emission tomography (PET) brain scans provided by the Alzheimer's Disease Neuroimaging Initiative (ADNI).[7] The motivation behind our work is as follows:

- Dosovitskiy *et al.*[6] achieved exceptional results in image classification tasks by applying standard transformers,[8] utilized in natural language processing (NLP), directly to images with the least possible modifications. In addition to its notable performance, this approach enables vision models to benefit from advancements in the NLP domain, including large language models, because of architectural similarities. Consequently, vanilla ViT[6] was a rational and sustainable foundation due to its design, performance, and simplicity for investigating what transformer-based vision models accomplish in diagnosing NDDs.
- ^{18}F-FDG PET scans, which reveal metabolic activities of various brain regions by measuring their glucose consumption, are considered pivotal in diagnosing and discriminating different NDDs, including mild cognitive impairment (MCI) and AD.[9] Although other brain imaging technologies such as computed tomography (CT) and magnetic resonance imaging (MRI) can expose NDDs too, PET scans have proved to be superior in exposing these brain conditions as soon as possible and earlier than other methods.[10,11]

Understanding the model's logic is the key to obtaining explainability in the medical domain, as human users must comprehend the reasoning behind each prediction before considering it. Therefore, we combined ViT's attention maps and the Automated Anatomical Atlas 3 (AAL3)[12] brain atlas to develop an explainable model that provides the most critical brain regions in the classification. The proposed model also delivers a heatmap of the input scan, in which the brightness of each pixel corresponds to its significance in the model's decision, overlaid on the original image, allowing the user to investigate pivotal regions further.

Our model achieves an F1 score of 81% (macro-average of all classes) on the test dataset, surpassing other approaches by a considerable gap. Please note that we only analyze our results against comparable studies regarding classes and the type of input brain scans. Furthermore, our proposed ViT has remarkable performance, in contrast to other models, in distinguishing MCI, which has proved to be one of the most challenging brain conditions to diagnose due to its prodromal nature. MCI is a transition stage between cognitively normal (CN) and AD. Consequently, MCI patients may experience some common NDD symptoms, such as memory loss or language problems, but the extent is such that they do not impede daily life.[13] Therefore, differentiating MCI cases from other categories can be inherently complicated.

Finally, we conducted experiments to reveal the contribution of different brain regions to the model's decisions. Although NDDs can affect various areas, this study showed that some brain regions are significantly more critical in the model's predictions.

To summarize, the contribution and novelty of this research is as follows:

- Introducing a complete data pre-processing and reshaping pipeline for 3D PET scans and brain atlases, allowing for fine-tuning of pre-trained ViTs on this type of data. This step is crucial since most ViTs are pre-trained on natural three-channel RGB images. Therefore, resizing and reshaping 3D data into three channels are essential to match the model's input shape.
- Obtaining competitive performance in ternary NDD classification (CN/MCI/AD) utilizing ^{18}F-FDG PET brain scans and vanilla ViT.[6] This approach is beneficial since vanilla ViT mostly shares the same architecture as the standard transformer,[8] used in NLP. Therefore, these architectural similarities could enable future studies to leverage advancements in NLP.
- Outperforming previous approaches by a noticeable margin (specifically in predicting MCI cases) in

the ternary classification of NDDs solely based on ^{18}F-FDG PET scans.
- Combining the model's attention maps and the AAL3 brain atlas for improved model explainability. Apart from the predicted label, our model provides a heatmap overlaid on the original input scan, highlighting the most influential brain regions to the model's prediction. Furthermore, the model delivers names of the key areas with the assistance of the AAL3 brain atlas.
- Performing a comprehensive brain regions' importance analysis by combining the model's attention maps and AAL3 atlas to find the most influential areas in the model's predictions. This study aims to enhance the model's explainability and suggest key areas in distinguishing various brain conditions.

2. Related works

Deep learning algorithms have shown outstanding results and potential in solving intricate tasks, motivating researchers to employ them for various medical image analysis tasks, including NDD classification.

Before the emergence of transformer-based vision models, such as ViTs,[6] most researchers had focused on employing CNNs for NDD classification.[4,5] Etminani *et al.*[4] proposed a comprehensive data pre-processing pipeline and a 3D CNN model based on VGG16[14] for NDD classification using ^{18}F-FDG PET scans. The authors demonstrated that 3D CNN algorithms could obtain competitive results compared to human readers, outperforming experienced nuclear medicine physicians independently and their consensus.[4] Furthermore, Etminani *et al.*[4] focused on explainability and dedicated a part of their research to interpreting the suggested model using an occlusion experiment.[15] Ding *et al.*[5] developed a CNN established on inception-v3[16] to classify NDDs through brain ^{18}F-FDG PET scans. The authors also compared their model's performance to radiology readers' using a subset of the ADNI and an independent test dataset, which resulted in the model's superior results in both cases. Furthermore, Ding *et al.*[5] employed the saliency map approach[17] for the model interpretation and analysis. Lozupone *et al.*[18] utilized 2D CNNs and a new explainable AI strategy to develop an interpretable model for classifying NDDs; however, the authors aimed for a two-class classification in their research and used 3D MRI brain scans for designing the model.

The advent of ViTs[6] and their cutting-edge performance in various computer vision tasks convinced researchers to investigate utilizing them in the medical domain and NDD diagnosis. Khatri and Kwon[19] focused on designing an explainable ViT utilizing self-supervised learning and ^{18}F-FDG PET scans for binary classification of two MCI sub-categories, namely convertible MCI (MCI-c) and stable MCI (MCI-s), to predict MCI progression to AD. The authors also studied attention regions for model explainability. Shin *et al.*[20] proposed applying ViTs on ^{18}F-florbetaben scans for binary and ternary classification of NDDs. Although this type of PET scan, which demonstrates beta-amyloid (β-amyloid) plaques in the brain, has proved beneficial in identifying NDDs, it is often used in research settings.[21] Therefore, ^{18}F-FDG has remained the most commonly used brain PET imaging technique.[11] Xing *et al.*[22] developed a multi-modal ViT by combining two types of PET brain scans (^{18}F-FDG and ^{18}F-AV45) for the binary classification of NDDs. Specifically, the proposed model includes two ViTs, each specialized in extracting features of a specific PET type. Then, the extracted features are concatenated and fed into a classifier for the final prediction.[22] Similarly, Odusami *et al.*[23] suggested an approach for binary classification of NDDs by fusing MRI and PET brain scans.

Most studies have concentrated on applying ViTs to MRI data. Unlike PET scans, which expose metabolic activities and functions, MRI is supposed to reveal the brain's structure. Therefore, MRI is usually beneficial in diagnosing NDDs at later stages, when the disease causes abnormalities in the physical brain's structure. Lyu *et al.*[24] developed a ViT solely based on an MRI dataset for a binary classification task; however, the authors added convolutional layers to their model to obtain better results. Sarraf *et al.*[25] proposed OVITAD, an optimized ViT architecture trained on a combination of functional MRI and structural MRI (sMRI) to classify NDDs. Furthermore, the authors used attention maps to achieve better model interpretation. Hoang *et al.*[26] focused their study on predicting MCI cases that could potentially progress into AD; therefore, the authors trained their ViT on sMRI data for a binary classification.[26] Aghdam *et al.*[27] applied a pre-trained pyramid ViT[28] to sMRI data to classify CN and AD cases. Kushol *et al.*[29] designed Addformer, which utilizes a new fusion transformer block that combines sMRI data in spatial and frequency domains to improve binary classification accuracy. They also visualized the model's attention maps, similar to most ViT-based studies, to gain model explainability. Shah *et al.*[30] introduced the multi-modal Bi-vision Transformer (BiViT), a ViT that includes two modules of mutual latent fusion and parallel coupled encoding strategy to enhance feature learning. The authors also utilized MRI data and demonstrated tokens for a better model understanding.

While we aimed for NDD classification in this work, there are some key differences compared to the literature, as follows:
- Achieving competitive performance in ternary NDD

classification while developing the ViT solely on ^{18}F-FDG PET scans
- Integrating a brain atlas with ViT's attention maps to gain model explainability and provide more information to the user.

3. Data and methods

3.1. Data acquisition

Data used in the preparation of this article were obtained from the ADNI database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Michael W. Weiner, MD, as the principal investigator. The primary goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD.

Figure 1 depicts a 3D raw ^{18}F-FDG PET scan selected from the ADNI dataset before our pre-processing steps along axial, sagittal, and coronal axes. A thorough description of technical details for each imaging session and phase is available in the ADNI documentation.[31]

The following criteria in choosing ^{18}F-FDG PET scans from ADNI were considered, similar to Etminani *et al.*:[4]
- CN and AD: We solely selected the most recent scan for each subject if more than one was available
- MCI: We exclusively chose the cases that later developed into AD during the ADNI studies. Considering this and the first criterion, the dataset includes the last MCI scan before progression to AD.

This sample selection procedure resulted in a dataset of size 580. Table 1 provides the details about the dataset split ratios and the number of samples.

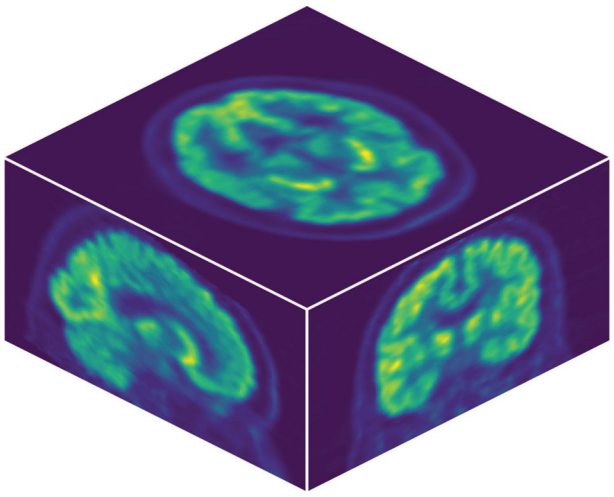

Figure 1. A 3D raw ^{18}F-FDG PET scan from the ADNI dataset along axial, sagittal, and coronal axes. The ADNI scans differ in voxel intensities, image size, and number of channels since they are obtained using a diverse range of scanners on different sites. Also, each scan contains the subject's skull, which does not provide beneficial information for our research. Therefore, these scans need pre-processing before utilizing them for model training.
Abbreviations: ADNI: Alzheimer's Disease Neuroimaging Initiative; ^{18}F-FDG: ^{18}F-fluorodeoxyglucose; PET: Positron emission tomography.

3.2. Brain imaging technologies and techniques

There are several brain imaging technologies with their unique advantages and disadvantages. Thus, in this part, we discuss the rationale behind utilizing ^{18}F-FDG PET scans in our research.

A PET scan imaging session starts after injecting slight amounts of a radioactive tracer into the subject's veins, which spreads to the body through the blood flow. The tracer enables the PET scanning device to capture metabolic activities in various tissues and organs, including the subject's brain.

Although all brain imaging technologies can reveal NDDs when sufficiently developed, PET scans are the best choice for detecting brain conditions at the earliest stages.[10,11] The reason is that NDDs usually cause abnormal metabolic patterns in some parts of the brain from the very early phases.[10] Therefore, PET imaging often exposes NDDs before other brain imaging technologies, including CT and MRI, due to its focus on the brain's metabolism.[10,11]

There are three well-known PET imaging types, namely amyloid, tau, and FDG, each suited for demonstrating special metabolic activities or changes in the brain using different tracers and procedures. Amyloid and tau PET scans, although showing promising results in NDD diagnosis, are commonly used in research settings at the time of writing.[21] Consequently, ^{18}F-FDG PET scans that show the brain's glucose (energy) usage are the most accessible and standard option in NDD diagnosis.

A central objective of our research was to propose a model and set of methods that enable rapid clinical diagnosis of NDDs. Consequently, ^{18}F-FDG PET scans were the most reasonable choice compared to other imaging technologies since they usually allow for early identification of NDDs.

Table 1. The number of samples per class and data split ratios

Class	Training	Validation	Test
CN	140	20	20
MCI	160	20	20
AD	160	20	20
Sum	460	60	60

Abbreviations: AD: Alzheimer's disease; CN: Cognitively normal; MCI: Mild cognitive impairment.

3.3. Data pre-processing

The ^{18}F-FDG PET scans in the ADNI dataset were acquired utilizing different types of imaging devices and during various phases. Therefore, these scans vary significantly in their properties, including the image size, number of channels, and intensities of voxels. Furthermore, they include the subject's skull, which does not deliver beneficial information regarding NDD diagnosis. Furthermore, raw scans may contain noise or blur due to the patient's movement or other technical issues. Consequently, we use the pre-processing procedure developed by Etminani *et al.*,[4] which employs MATLAB[32] and statistical parametric mapping (SPM12)[33] to ensure all scans have the same properties.

The pre-processing steps for each sample are as follows:

- We converted the scan to the NIfTI format
- It was crucial to place the brain approximately in the center of the scan. Therefore, we reoriented and repositioned the brain to set the volume's origin at the anterior commissure region
- Our dataset included scans of various shapes. Hence, we normalized the scan to ensure all samples had identical spatial size and number of channels
- Using the tissue probability map of SPM12, the brain was segmented
- The last pre-processing stage removed the subject's skull from the scan. Consequently, we used segmentation maps obtained from the previous step with a filter for skull-stripping.

The pre-processing procedure led to skull-stripped scans of size 79 × 95 × 79, representing channels, height, and width, respectively. Then, the values of voxels were normalized using a min-max scaler across the channels in Python. Finally, we dismissed the initial ten and last nine channels of the 3D scan since they included a tiny fraction of the brain, resulting in 3D scans with the shape of 60 × 95 × 79.

3.4. Data reshaping

According to our experiments and the literature,[6] pre-training on large amounts of data is crucial to achieving the best performance using ViTs. However, most large computer vision datasets include natural images with three RGB channels. Therefore, we reshaped the samples to 3 × 570 × 950 in our dataset to utilize transfer learning and available pre-trained models. This procedure constructs a three-channel image, in which every channel depicts the brain along a unique axis (sagittal, coronal, and axial). Figure 2 shows the result of data pre-processing and reshaping steps on a single scan.

3.5. Model architecture and training

Our proposed model has a similar architecture to vanilla ViT, as suggested by Dosovitskiy *et al.*[6] After training different models with and without transfer learning, we concluded that pre-training is crucial to obtaining excellent results. Therefore, we employed the Hugging Face[34] Transformer API and model hub for development. Specifically, the foundation of our model is a base-sized ViT, pre-trained on ImageNet-21k[35] and ImageNet 2012,[36] respectively.[37] Finally, we fine-tuned the model on our ^{18}F-FDG PET scan dataset to classify NDDs.

Figure 3 illustrates the model's diagram, inspired by Dosovitskiy *et al.*[6] First, the scan was resized to 3 × 384 × 384 to match the model's input shape. Then, the scan was divided into patches of 3 × 32 × 32, flattened, and supplied to a standard transformer along with position embeddings holding the spatial information. Finally, a multilayer perceptron head translated the model's final hidden state into the probability of classes for the classification task. Table 2 summarizes the model's specifications.

We employed an AdamW optimizer (learning rate = 5e-5, weight decay = 0.15) for model development. Furthermore, an exponential learning rate decay (γ = 0.9999 per epoch) was used during training. Finally, we selected a weighted cross-entropy as the loss function, in which the weight of each class was the inverse of its frequency during training, as shown below:

$$l(x,y) = L = \{l_1 \dots l_N\}^\intercal \quad \text{(I)}$$

$$l_n = -\sum_{c=1}^{C} w_c \log \frac{\exp(x_{n,c})}{\sum_{i=1}^{C} \exp(x_{n,i})} y_{n,c} \quad \text{(II)}$$

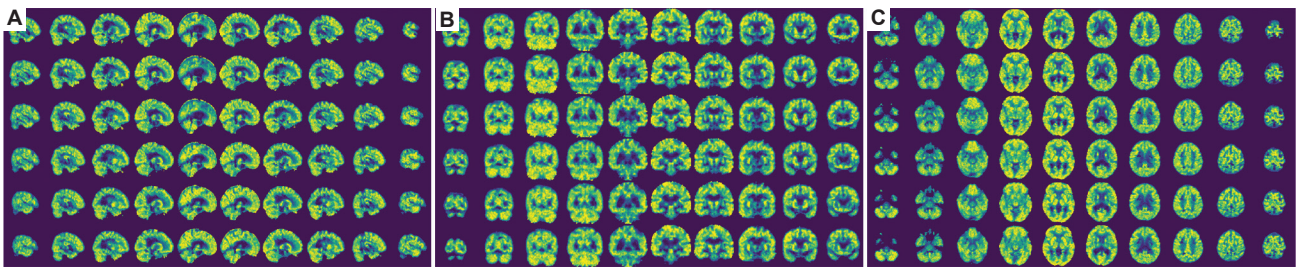

Figure 2. After the initial data pre-processing, we reshaped each scan of 60 × 95 × 79 to a three-channel image of 3 × 570 × 950, in which every channel illustrates the brain along a unique axis. This data reshaping was crucial to utilize transfer learning and pre-train the model on large computer vision datasets that contain natural three-channel RGB images. (A) Sagittal, (B) Coronal, (C) Axial.

$$w_c = \frac{1}{Class\ Frequency} \rightarrow w_{CN} = \frac{1}{140}, w_{MCI} = \frac{1}{160}, w_{AD} = \frac{1}{160} \quad \text{(III)}$$

Finally, Algorithm 1 shows the data augmentation process for model training.

Algorithm 1. The data augmentation procedure used in the model training

```
t1←GaussianBlur (kernel_size= (3, 3), sigma = (0.1,2))
t2←GaussianNoise (mean=0, std=0.05)
t3←ColorJitter (brightness=0.1)
t4←ColorJitter (contrast=0.1)
t5←ColorJitter (saturation=0.1)
random_choice←RandomChoice([t1, t2, t3, t4, t5])
transforms←RandomApply([random_choice], p=0.7)
```

3.6. Explainability

Explainability is vital in healthcare since experts should understand the reason behind the model's predictions before considering or counting them. Therefore, we combined the model's attention maps and the AAL3 brain atlas to discover the most impactful brain regions on the model's conclusions. During the inference mode, our model follows these steps to provide various details to the user:

- The model extracts the attention map of each input scan and overlays this data on the original image. The outcome of this step is a heatmap of the brain regions, in which pixel values correspond to their influence on the model's decision
- The model illuminates pixels with values exceeding 95% of the maximum value using red rectangles. This step enables the user to examine and analyze all key areas in the input scan
- Ultimately, the model overlays the heatmap, extracted in the first step, on the AAL3 atlas and locates pixels with the highest intensity to provide the name of the brain regions that encompass them. Providing these areas' names is crucial to the user since they are most influential in the model's prediction.

We reshaped the AAL3 atlas to 3 × 950 × 570 to fit the size of our input scans using the following procedure; the result of which is in Figure 4:

- By overlaying the AAL3 atlas on a pre-processed sample in MRIcron,[38] we first reshaped AAL3 to 79 × 95 × 79
- It was crucial to verify that both the reshaped atlas and the input scan followed the same coordinate system. Therefore, we loaded the resulting new atlas and the input scan into MRIcron again and compared their coordinate system side-by-side. This step ensured that corresponding coordinates referred to the same brain area in both files
- We discarded the first ten and last nine slices from AAL3, similar to the input scans, resulting in a shape

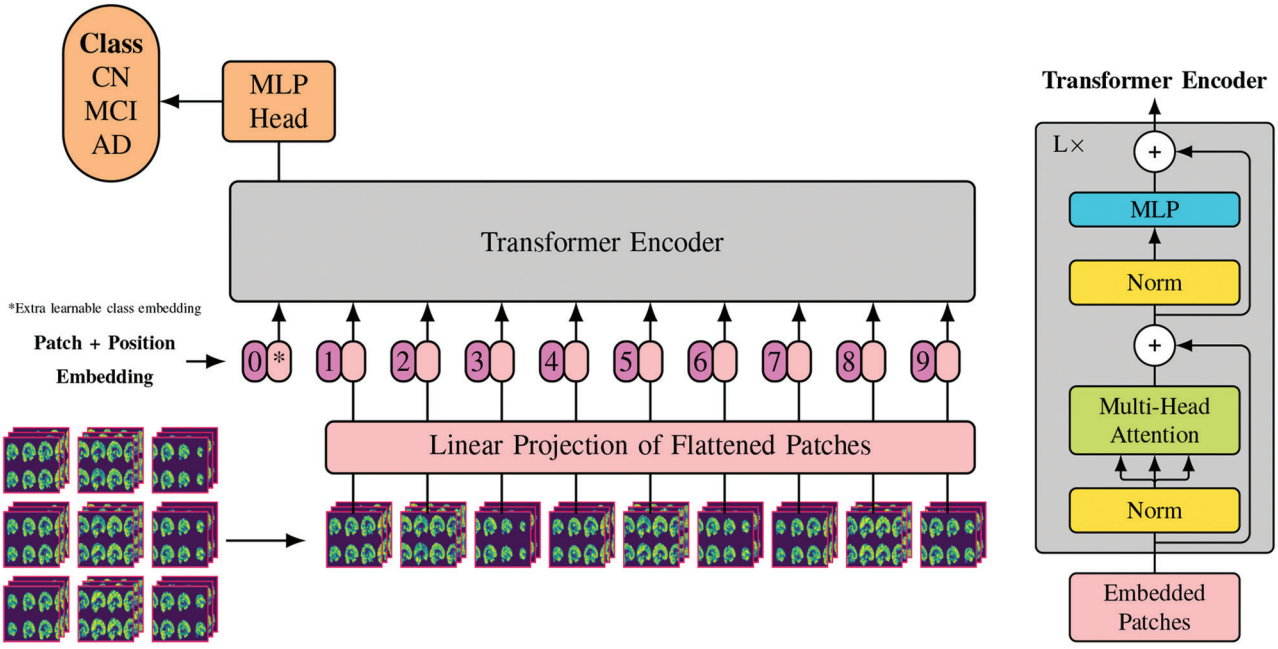

Figure 3. The model architecture is identical to the ViT-Base introduced by Dosovitskiy *et al.*[6] First, the scan is reshaped into 3 × 384 × 384 to fit the model's input. Then, it is split into patches of shape 3 × 32 × 32, flattened, and provided to a standard transformer along with position embeddings that contain spatial information. At the last stage, an MLP acts as the classification head to map the final hidden state into the probability of classes. The illustration of the model's architecture was inspired by Dosovitskiy *et al.*[6]
Abbreviations: AD: Alzheimer's disease; CN: Cognitively normal; MCI: Mild cognitive impairment; MLP: Multilayer perceptron.

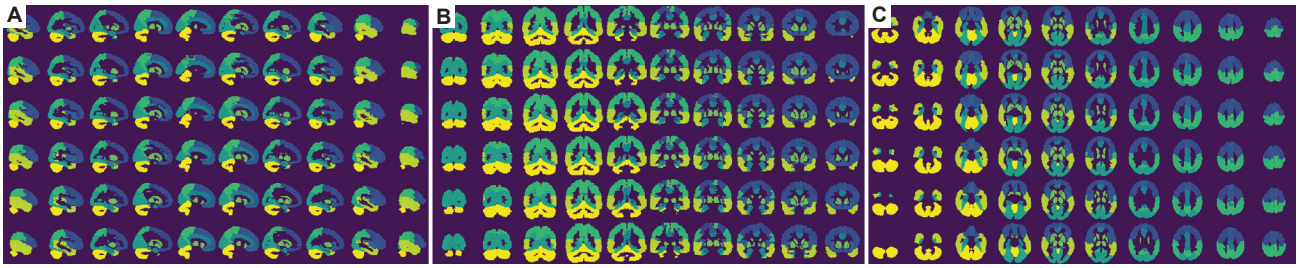

Figure 4. The resized AAL3 brain atlas (3 × 950 × 570), where channels illustrate regions from three different perspectives. Reshaping the atlas was a critical step since its dimensions should match that of input scans. Each color indicates a different brain area. (A) Sagittal, (B) Coronal, (C) Axial. Abbreviation: AAL3: Automated Anatomical Atlas 3.

of 60 × 95 × 79
- The final stage entailed projecting AAL3 into three channels along different axes, resulting in an image of shape 3 × 950 × 570.

3.7. Regions' importance study

A vital component of our research was identifying the most critical brain regions to the model's predictions. Apart from achieving better explainability, this study can help researchers and clinicians pay special attention to these key areas during their examinations.

As mentioned, our model utilizes the AAL3 atlas to provide the attention map and the most critical brain region for each input scan. Therefore, we conducted our study in the following manner:
- We combined the training, validation, and test sets to form a dataset of 580 scans
- After feeding all samples to the model, we saved the attention maps and suggested critical regions for each correctly classified scan in a database
- We considered the occurrence rate of each region in the database as a metric to show its importance in the model's predictions
- Finally, we calculated the mean of all attention maps to generate a heatmap of brain areas.

4. Results

4.1. Model performance

Figure 5 illustrates the confusion matrix of the model's predictions. The model performed the best in distinguishing between CN and AD cases with no error. However, classifying MCI cases was challenging for the model, similar to human experts, due to their prodromal state. Specifically, differentiating MCI and AD cases needed the most enhancement with an error of 25%. This error might be due to selecting the last scan of each MCI case that progressed into AD later, which made distinguishing these two classes more challenging. In addition, Table 3 demonstrates the performance of the proposed model in detail using several metrics. Similar to the confusion matrix, this table reveals the challenge of classifying MCI cases.

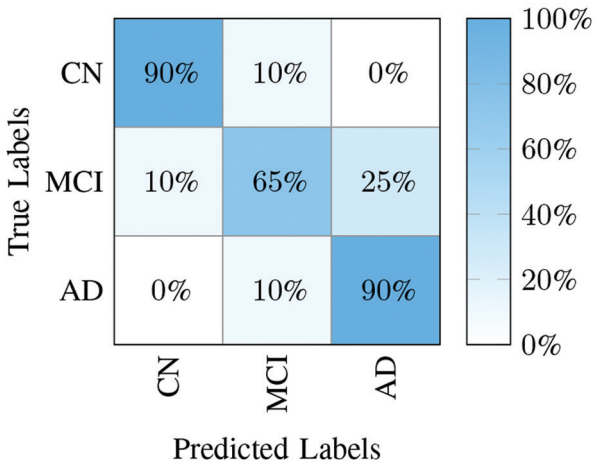

Figure 5. The model's confusion matrix, illustrating its performance on the test dataset, with values normalized over true labels. The model can perfectly distinguish CN and AD cases with no error. However, classifying MCI is challenging due to its prodromal nature. Abbreviations: AD: Alzheimer's disease; CN: Cognitively normal; MCI: Mild cognitive impairment.

Table 2. Specifications of the proposed vision transformer

Parameter	Value
Input shape	3×384×384
Patch size	3×32×32
Layers	12
Hidden size	768
Multilayer perceptron size	3072
Heads	12
Hidden dropout	0.1

Finally, to comprehend the model's representation of the learned data, we conducted a principal component analysis (PCA) on the last hidden state before SoftMax. Figure 6 illustrates the results of this analysis for the training and test datasets individually.

Table 3. The performance of the proposed model per class using different classification metrics

Class	Sensitivity	Specificity	Precision	F1 score	Accuracy (95% CI)
CN	0.90	0.95	0.90	0.90	82% (72%, 91%)
MCI	0.65	0.90	0.76	0.70	
AD	0.90	0.88	0.78	0.84	
Macro-average	0.82	0.91	0.81	0.81	

Abbreviations: AD: Alzheimer's disease; CI: Confidence interval; CN: Cognitively normal; MCI: Mild cognitive impairment.

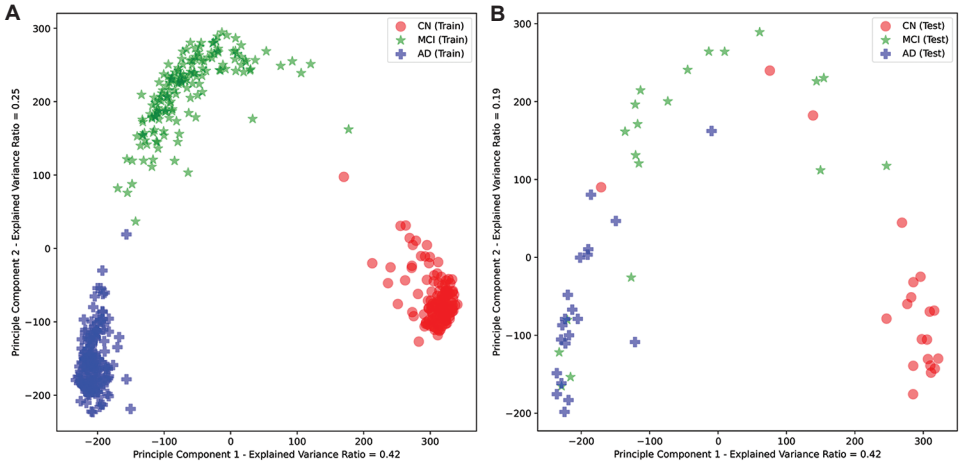

Figure 6. The result of dimensionality reduction on the model's last hidden state before SoftMax using PCA with two principal components. This analysis is beneficial in gaining insight into the learned representation of data and the model's performance in distinguishing between classes. The figure illustrates the true labels. (A) Train, (B) Test.
Abbreviations: AD: Alzheimer's disease; CN: Cognitively normal; MCI: Mild cognitive impairment; PCA: Principal component analysis.

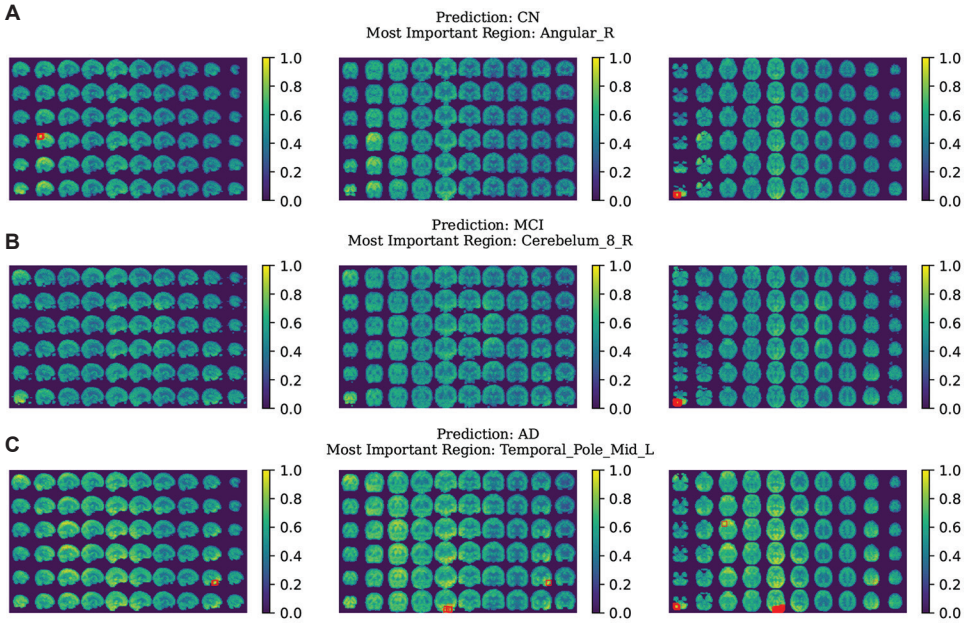

Figure 7. The model's inference output for three correctly classified samples. Our model provides extra information during inference to obtain explainability and to assist the user in making a diagnosis. (A) CN, (B) MCI, (C) AD.
Abbreviations: AD: Alzheimer's disease; CN: Cognitively normal; MCI: Mild cognitive impairment.

4.2. Explainability

Figure 7 illustrates the prediction results of three correctly classified scans. As depicted, the model provides the following information to the user in the inference mode:

- The predicted label
- The brain region that has the most influence on the model's prediction. This information results from locating the pixel with the highest intensity value in an overlay of the attention map and the AAL3 brain atlas
- An overlay of the attention map and the input scan, in which the brightness of each pixel is analogous to its significance in the model's conclusion. Red rectangles also illustrate regions with attention values greater than 95% of the maximum attention.

In addition to the predicted label, this information enables domain experts to find out the model's logic and examine the brain's key areas further.

4.3. Regions' importance study

Figure 8 illustrates the overall importance of different regions and for predicting each label independently. Please note that Figure 8 only contains the AAL3 regions that our model suggested as crucial at least once, ignoring all other areas without any occurrence during inference. Our model suggests the angular gyrus, known to be heavily affected by MCI and AD,[39-41] as the most critical region in distinguishing the CN class. In addition, the temporal pole is the key area in classifying AD, aligning with previous studies that found all AD patients experience atrophy and other complications in this brain region.[42] Finally, the proposed model defines the cerebellum as the essential area for MCI classification. Traditionally, this part of the brain did not play a pivotal role in diagnosing NDDs.[43] However, recent studies have revealed the significance of the cerebellum in diagnosing MCI and various stages of AD.[43] Further investigations also indicate that AD progression causes cerebellar transformations, and this region is central to obtaining significantly better performance in classification tasks.[44]

Figure 9 shows the brain heatmaps, where the brightness of a pixel signifies its impact on the model's decisions. As indicated in both figures, some brain regions play a substantial role in diagnosing various classes.

5. Discussion

Affecting millions of lives, NDDs are a leading cause of death and disability worldwide.[1] Although remaining mostly incurable, early diagnosis of such conditions is a key to better disease management and enhancing the patient's quality of life.[2]

Diagnosing NDDs is challenging, even for proficient nuclear medicine physicians, and requires substantial

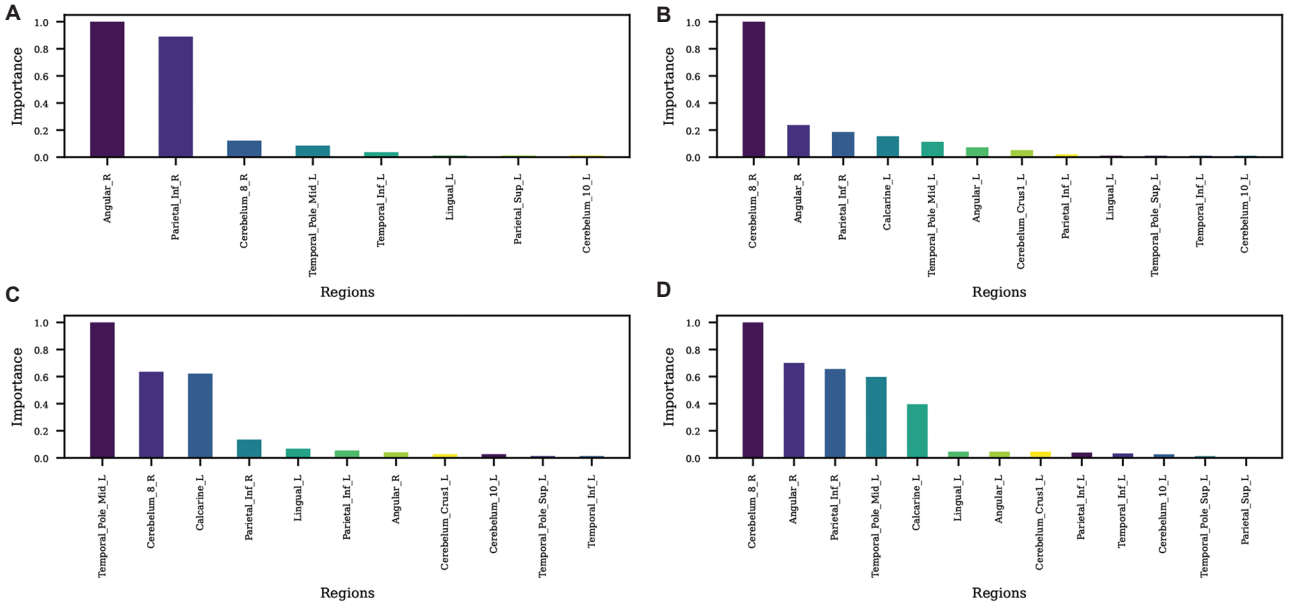

Figure 8. The significance of the AAL3 regions in predicting each class and their overall contributions during our regions' importance study. After combining the training, validation, and test datasets, we fed the resulting dataset of 580 samples to the model and saved the suggested crucial region for correctly classified scans. Then, we considered the occurrence rate of each region as a metric to show its importance in the model's diagnoses. Please note we only included the areas suggested by the model as critical at least once, ignoring all other parts without any occurrence during inference. (A) CN, (B) MCI, (C) AD, (D) Overall. Abbreviations: AAL3: Automated Anatomical Atlas 3; AD: Alzheimer's disease; CN: Cognitively normal; MCI: Mild cognitive impairment.

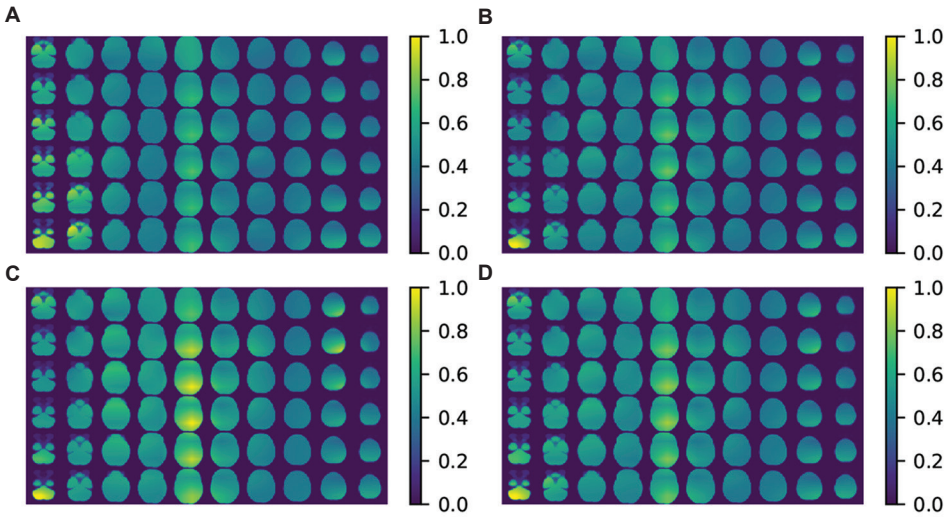

Figure 9. The mean attention maps of brain areas in our regions' importance study, where the brightness of a pixel translates to its significance in the model's predictions. This figure results from merging the training, validation, and test datasets and feeding the final dataset of 580 samples to the model. Then, we saved the model's attention maps for all correctly classified scans. Finally, we computed the mean of saved attention maps to generate a mean attention map for the whole dataset and each class. (A) CN, (B) MCI, (C) AD, (D) Overall.
Abbreviations: AD: Alzheimer's disease; CN: Cognitively normal; MCI: Mild cognitive impairment.

knowledge and training. Therefore, this complexity in disease diagnosis, together with other factors, has left about 75% of NDD patients undiagnosed worldwide, and this number rises to about 90% in low- and middle-income countries, according to some studies.[3] In addition to revealing a considerable diagnosis gap, these investigations predict that the fast-growing number of NDD patients could strain healthcare systems in the future.[3] Consequently, novel and affordable tools and techniques are required for the final diagnosis of NDD and/or for assisting healthcare providers in this task.

While the rapid progress of AI and its sub-fields revolutionized our lives, researchers have attempted to harness the power of these new technologies in the healthcare domain, including NDD diagnosis. Although most of these research projects utilized long-established approaches and architectures like CNNs, the emergence of ViTs and their groundbreaking performance convinced us to explore the potential of employing this new architecture in NDD classification.

In this work, we developed a model to classify ^{18}F-FDG PET brain scans into CN, MCI, and AD. Specifically, we designed the model based on the vanilla ViT-Base, introduced by Dosovitskiy *et al.*,[6] and trained on the ADNI dataset.[7] Combining the proposed data, pre-processing procedure, training recipe, and transfer learning enabled our model to achieve an F1 score of 81% (macro-average of all classes), significantly outperforming previous approaches. To comprehend the model's performance, we also compare our results with those presented in other papers, as listed in Table 4. However, a few points worth mentioning for a fair comparison:

- While several papers examined NDD classification using DNNs, we exclusively selected studies that aimed for ternary classification (CN/MCI/AD) using ^{18}F-FDG PET scans
- Although all chosen studies employed ADNI as their primary dataset, the authors may have used different subsets to train and test the models
- As shown in Table 4, DNNs can surpass physicians when NDD diagnosis is solely based on brain scans. However, domain experts usually consider a comprehensive collection of information for the clinical diagnosis, including the patient's medical history, genetics, blood tests, and cognitive and physical evaluations. Consequently, instead of solely relying on brain scans, doctors and nuclear medicine physicians consider various factors, a practice that poses a substantial advantage over AI models.

According to Table 4, in addition to a significantly higher F1 score, our model excels in distinguishing MCI cases, which proved to be the most challenging condition to classify, compared to human experts and other models.

Apart from improving the model's performance, developing an explainable model was a pivotal goal of this research. Therefore, we integrated the AAL3 brain atlas information into the attention maps. This method resulted in a model that provides the brain region with the highest

Table 4. The performance comparison of our model with others in the literature

Model	Class	Sensitivity	Specificity	Precision	F1 score	F1 score (micro-average)
Ding *et al.*[5] (model CNN)	CN	0.59	0.75	0.60	0.59	0.64
	MCI	0.54	0.68	0.55	0.55	
	AD	0.81	0.94[a]	0.76	0.78	
Etminani *et al.*[4] (model CNN)*	CN	0.88	0.90	0.81	0.84	0.63
	MCI	0.17	0.94[a]	0.20	0.18	
	AD	0.91[a]	0.92	0.83[a]	0.87[a]	
Etminani *et al.*[4] (consensus of human readers)*†	CN	0.70	0.81	0.64	0.67	0.45
	MCI	0.25	0.75	0.08	0.12	
	AD	0.47	0.90	0.68	0.56	
Our model (ViT)	CN	0.90[a]	0.95[a]	0.90[a]	0.90[a]	0.81[a]
	MCI	0.65[a]	0.90	0.76[a]	0.70[a]	
	AD	0.90	0.88	0.78	0.84	

Notes: All results were obtained using ^{18}F-FDG PET brain scans from the ADNI dataset. Besides acquiring a significantly higher F1 score, our model outperforms others in classifying MCI cases by a considerable margin. *The authors considered an additional DLB class in their paper and tested the model and human readers in a four-class classification task (CN, MCI, AD, DLB). †Four professional nuclear medicine physicians with 3, 8, 13, and 16 years of experience. [a]The highest values per each metric and class.

Abbreviations: AD: Alzheimer's disease; ADNI: Alzheimer's Disease Neuroimaging Initiative; CN: Cognitively normal; CNN: Convolutional neural network; DLB: Dementia with Lewy bodies; ^{18}F-FDG: ^{18}F-fluorodeoxyglucose; MCI: Mild cognitive impairment; PET: Positron emission tomography; ViT: Vision transformer.

impact on its decision and a heatmap of the input scan, in which the pixels' intensities illustrate their importance in the model's prediction. Furthermore, we conducted a study to reveal regions' significance in the model's decisions, showing some brain areas are of utmost importance in predicting various conditions.

5.1. Limitations

A key goal of this study was to ensure that the model can distinguish and classify all stages of AD. Therefore, we decided to develop the model solely on MCI cases that later progressed to AD. However, this sample selection might introduce some bias in the dataset regarding the MCI class and result in the model becoming a prognostic MCI-to-AD classifier. In addition, we exclusively relied on the ADNI dataset in this research, which may restrict the model's out-of-distribution generalization.

Although the critical regions suggested by the proposed model are consistent with the literature to a significant extent, our findings require more examination and validation by medical domain experts.

Finally, we should stress that while DNNs demonstrate promising results in NDD classification, they are limited to their datasets, substantially affecting their generalization performance. In addition, unlike clinical procedures, DNNs do not consider all medical factors and base their predictions on limited data. Therefore, vast improvements and training on diverse datasets are critical to designing robust and clinically applicable models that can assist experts in NDD diagnosis.

6. Conclusion

We believe this research showcases the extraordinary potential of the ViT architecture in NDDs classification, which surpasses other methods, including CNNs. Apart from their excellent performance, ViTs allow computer vision researchers to benefit from advancements in NLP due to the sharing of the same transformer architecture. Furthermore, ViTs make developing explainable models more feasible by leveraging the attention mechanism.

Acknowledgments

None.

Funding

Data collection and sharing for this project was funded by the ADNI (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Dec 5, 2024 12:30 PM Genentech,

Inc.; Fujirebio; GE Healthcare; IXICO Ltd.;Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research provided funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (https://www.fnih.org/). The grantee organization is the Northern California Institute for Research and Education, and the study was coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data were disseminated by the Laboratory for Neuro Imaging at the University of Southern California. Farzaneh Etminani and Amira Soliman are supported by Center for Applied Intelligent Systems Research in Health (CAISR Health) funded by Knowledge Foundation (grant no.: 20200208 01H).

Conflict of interest

The authors declare no conflicts of interest.

Author contributions

Conceptualization: All authors
Formal analysis: Pooriya Khorramyar, Amira Soliman
Investigation: All authors
Methodology: Pooriya Khorramyar, Amira Soliman
Writing–original draft: Pooriya Khorramyar
Writing–review & editing: All authors

Ethics approval and consent to participate

Please visit the ADNI documentation at https://adni.loni.usc.edu/help-faqs/adni-documentation/.

Consent for publication

Please visit the ADNI documentation at https://adni.loni.usc.edu/help-faqs/adni-documentation/.

Availability of data

To download the dataset, please visit the ADNI's web page at https://adni.loni.usc.edu/. The code is available on the project's GitHub repository (https://github.com/Pooriya-Kh/NDD_ViT).

References

1. World Health Organization (WHO). *Dementia Key Facts*. Available from: https://www.who.int/news-room/fact-sheets/detail/dementia [Last accessed on 2025 Mar 30].
2. Alzheimer Society of Canada. *The 10 Benefits of Early Diagnosis*. Available from: https://alzheimer.ca/en/about/dementia/do/i/have/dementia/how/get/tested/dementia/tips/individuals-families-friends/10 [Last accessed on 2025 Mar 30].
3. Alzheimer's Disease International. *Over 41 Million Cases of Dementia go Undiagnosed Across the Globe - World Alzheimer Report Reveals*. Available from: https://www.alzint.org/news/events/news/over/41/million/cases/of/dementia/go/undiagnosed-across-the-globe-world-alzheimer-report-reveals [Last accessed on 2025 Mar 30].
4. Etminani K, Soliman A, Davidsson A, *et al*. A 3D deep learning model to predict the diagnosis of dementia with lewy bodies, Alzheimer's disease, and mild cognitive impairment using brain 18F-FDG PET. *Eur J Nucl Med Mol Imaging*. 2022;49(2):563-584.
   doi: 10.1007/s00259-021-05483-0
5. Ding Y, Sohn JH, Kawczynski MG, *et al*. A deep learning model to predict a diagnosis of alzheimer disease by using ^{18}F-FDG PET of the brain. *Radiology*. 2019;290(2):456-464.
   doi: 10.1148/radiol.2018180958
6. Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. [ArXiv Preprint]; 2020.
7. *The Alzheimer's Disease Neuroimaging Initiative. Alzheimer's Disease Neuroimaging Initiative (ADNI)*. Available from: https://adni.loni.usc.edu [Last accessed on 2025 Mar 30].
8. Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. *Advances in Neural Information Processing Systems*. 2017. p. 30.
9. Nobili F, Arbizu J, Bouwman F, *et al*. European association of nuclear medicine and European academy of neurology recommendations for the use of brain ^{18}F-fluorodeoxyglucose positron emission tomography in neurodegenerative cognitive impairment and dementia: Delphi consensus. *Eur J Neurol*. 2018;25(10):1201-1217.
   doi: 10.1111/ene.13728
10. Mayo Foundation for Medical Education and Research (MFMER). *Positron Emission Tomography Scan*. Available from: https://www.mayoclinic.org/tests/procedures/pet/scan/about/pac-20385078 [Last accessed on 2025 Mar 30].
11. Cleveland Clinic. *PET Scan*. Available from: https://my.clevelandclinic.org/health/diagnostics/10123-pet-scan [Last accessed on 2025 Mar 30].
12. Rolls ET, Huang CC, Lin CP, Feng J, Joliot M. Automated anatomical labelling atlas 3. *Neuroimage*. 2020;206:116189.
   doi: 10.1016/j.neuroimage.2019.116189
13. Mayo Foundation for Medical Education and Research (MFMER). *Mild Cognitive Impairment (MCI)*. Available

from: https://www.mayoclinic.org/diseases/conditions/mild/cognitive/impairment/symptoms/causes/syc-20354578 [Last accessed on 2025 Mar 30].

14. Simonyan K, Zisserman A. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. [ArXiv Preprint]; 2014.

15. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*. Berlin: Springer; 2014. p. 818-833.

16. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016. p. 2818-2826.

17. Simonyan K, Vedaldi A, Zisserman A. *Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps*. [ArXiv Preprint]; 2013.

18. Lozupone G, Bria A, Fontanella F, Meijer FJ, De Stefano C. *AXIAL: Attention-Based Explainability for Interpretable Alzheimer's Localized Diagnosis using 2D CNNs on 3D MRI Brain Scans*. [ArXiv Preprint]; 2024.

19. Khatri U, Kwon GR. Explainable vision transformer with self-supervised learning to predict Alzheimer's disease progression using 18F-FDG PET. *Bioengineering (Basel)*. 2023;10(10):1225.

doi: 10.3390/bioengineering10101225

20. Shin H, Jeon S, Seol Y, Kim S, Kang D. Vision transformer approach for classification of Alzheimer's disease using 18f-florbetaben brain images. *Appl Sci*. 2023;13(6):3453.

doi: 10.3390/app13063453

21. The National Institutes of Health. *How Biomarkers Help Diagnose Dementia*. Available from: https://www.nia.nih.gov/health/alzheimers/symptoms/and/diagnosis/how/biomarkers-help-diagnose-dementia [Last accessed on 2025 Mar 30].

22. Xing X, Liang G, Zhang Y, Khanal S, Lin AL, Jacobs N. Advit: vision transformer on multi-modality pet images for Alzheimer disease diagnosis. In: *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. United States: IEEE; 2022. p. 1-4.

doi: 10.1109/ISBI52829.2022.9761584

23. Odusami M, Maskeliūnas R, Damaševičius R. Pixel-level fusion approach with vision transformer for early detection of Alzheimer's disease. *Electronics*. 2023;12(5):1218.

doi: 10.3390/electronics12051218

24. Lyu Y, Yu X, Zhu D, Zhang L. 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. PETRA '22*. United States: Association for Computing Machinery; 2022. p. 463-468.

doi: 10.1145/3529190.3534754

25. Sarraf S, Sarraf A, DeSouza DD, Anderson JAE, Kabia M, 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*. 2023;13(2):260.

doi: 10.3390/brainsci13020260

26. Hoang GM, Kim UH, Kim JG. Vision transformers for the prediction of mild cognitive impairment to Alzheimer's disease progression using mid-sagittal sMRI. *Front Aging Neurosci*. 2023;15:1102869.

doi: 10.3389/fnagi.2023.1102869

27. Aghdam MA, Bozdag S, Saeed F, Alzheimer's Disease Neuroimaging Initiative. PVTAD: Alzheimer's disease diagnosis using pyramid vision transformer applied to white matter of T1-weighted structural MRI data. *Proc IEEE Int Symp Biomed Imaging*. 2024;2024:10.

doi: 10.1109/isbi56570.2024.10635541

28. Wang W, Xie E, Li X, *et al*. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. United States: IEEE; 2021. p. 568-578.

29. Kushol R, Masoumzadeh A, Huo D, Kalra S, Yang YH. Addformer: Alzheimer's disease detection from structural MRI using fusion transformer. In: *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE; 2022. p. 1-5.

doi: 10.1109/ISBI52829.2022.9761421

30. Shah SMAH, Khan MQ, Rizwan A, Jan SU, Samee NA, Jamjoom MM. Computer-aided diagnosis of Alzheimer's disease and neurocognitive disorders with multimodal Bi-vision transformer (BiViT). *Pattern Anal Appl*. 2024;27(3):76.

doi: 10.1007/s10044-024-01297-6

31. *The Alzheimer's Disease Neuroimaging Initiative. ADNI Documentation*. Available from: https://adni.loni.usc.edu/help-faqs/adni-documentation [Last accessed on 2025 Mar 30].

32. The MathWorks Inc. *Matlab R*; 2016a. Available from: https://www.mathworks.com [Last accessed on 2025 Mar 30].

33. UCL Queen Square Institute of Neurology. *Statistical Parametric Mapping*. Available from: https://www.fil.ion.ucl.ac.uk/spm [Last accessed on 2025 Mar 30].

34. Wolf T, Debut L, Sanh V, *et al*. *HuggingFace's Transformers: State-of-the-Art Natural Language Processing*; 2020. Available from: https://arxiv.org/abs/1910.03771 [Last accessed on

2025 Mar 30].

35. Ridnik T, Baruch EB, Noy A, Zelnik-Manor L. *ImageNet-21K Pretraining for the Masses*. CoRR. 2021. Available from: https://arxiv.org/abs/2104.10972 [Last accessed on 2025 Mar 30].

36. Russakovsky O, Deng J, Su H, *et al*. ImageNet large scale visual recognition challenge. *Int J Comput Vis IJCV*. 2015;115(3):211-252.

    doi: 10.1007/s11263-015-0816-y

37. Google. *Vision Transformer (Base-Sized Model)*. Available from: https://huggingface.co/google/vit-base-patch32-384 [Last accessed on 2025 Mar 30].

38. Rorden C, Brett M. Stereotaxic display of brain lesions. *Behav Neurol*. 2000;12(4):191-200.

    doi: 10.1155/2000/421719

39. Li Y, Wang X, Li Y, *et al*. Abnormal resting-state functional connectivity strength in mild cognitive impairment and its conversion to Alzheimer's disease. *Neural Plast*. 2016;2016:4680972.

    doi: 10.1155/2016/4680972

40. Talwar P, Kushwaha S, Chaturvedi M, Mahajan V. Systematic review of different neuroimaging correlates in mild cognitive impairment and Alzheimer's disease. *Clin Neuroradiol*. 2021;31(4):953-967.

    doi: 10.1007/s00062-021-01057-7

41. Salmon E, Collette F, Bastin C. Cerebral glucose metabolism in Alzheimer's disease. *Cortex*. 2024;179:50-61.

    doi: 10.1016/j.cortex.2024.07.004

42. Arnold SE, Hyman BT, Van Hoesen GW. Neuropathologic changes of the temporal pole in Alzheimer's disease and pick's disease. *Arch Neurol*. 1994;51(2):145-150.

    doi: 10.1001/archneur.1994.00540140051014

43. Yang C, Liu G, Chen X, Le W. Cerebellum in Alzheimer's disease and other neurodegenerative diseases: An emerging research frontier. *MedComm (2020)*. 2024;5(7):e638.

    doi: 10.1002/mco2.638

44. Bruchhage MMK, Correia S, Malloy P, Salloway S, Deoni S. Machine learning classification identifies cerebellar contributions to early and moderate cognitive decline in Alzheimer's disease. *Front Aging Neurosci*. 2020;12:524024.

    doi: 10.3389/fnagi.2020.524024