# Machine learning approaches in comparative studies for Alzheimer's diagnosis using 2D MRI slices

Zhen ZHAO
s2143852@siswa.um.edu.my

Joon Huang CHUAH
jhchuah@um.edu.my

Chee-Onn CHOW
cochow@um.edu.my

Kaijian XIA
kjxia@suda.edu.cn

Yee Kai TEE
teeyeekai@gmail.com

*See next page for additional authors*
Follow this and additional works at: https://journals.tubitak.gov.tr/elektrik

Part of the Computer Engineering Commons, Computer Sciences Commons, and the Electrical and Computer Engineering Commons

# Machine learning approaches in comparative studies for Alzheimer's diagnosis using 2D MRI slices

## Authors

Zhen ZHAO, Joon Huang CHUAH, Chee-Onn CHOW, Kaijian XIA, Yee Kai TEE, Yan Chai HUM, and Khin Wee LAI

Research Article

# Machine learning approaches in comparative studies for Alzheimer's diagnosis using 2D MRI slices

**Zhen ZHAO[1]** [iD]**, Joon Huang CHUAH[1]** [iD]**, Chee-Onn CHOW[1]** [iD]**, Kaijian XIA[2]** [iD]**,**
**Yee Kai TEE[3]** [iD]**, Yan Chai HUM[3]** [iD]**, Khin Wee LAI[4]\*** [iD]

[1]Department of Electrical Engineering, Faculty of Engineering, Universiti Malaya,
Kuala Lumpur, Malaysia
[2]Changshu Hospital Affiliated to Soochow University (Changshu No. 1 People's Hospital),
1 Shuyuan St, Changshu, Suzhou, Jiangsu, PR China
[3]Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, Selangor, Malezya
[4]Department of Biomedical Engineering, Faculty of Engineering, Universiti Malaya,
Kuala Lumpur, Malaysia

**Abstract:** Alzheimer's disease (AD) is an illness that involves a gradual and irreversible degeneration of the brain. It is crucial to establish a precise diagnosis of AD early on in order to enable prompt therapies and prevent further deterioration. Researchers are currently focusing increasing attention on investigating the potential of machine learning techniques to simplify the automated diagnosis of AD using neuroimaging. The present study involved a comparison of models for the detection of AD through the utilization of 2D image slices obtained from magnetic resonance imaging brain scans. Five models, namely ResNet, ConvNeXt, CaiT, Swin Transformer, and CVT, were implemented to learn features and classify AD based on various perspectives of 2D image slices. A series of experiments were conducted using the dataset from the Alzheimer's Disease Neuroimaging Initiative. The results showed that ConvNeXt outperformed ResNet, CaiT, Swin Transformer, and CVT. ConvNeXt exhibited an average accuracy, precision, recall, and F1 score of 95.74%, 96.71%, 95.74%, and 96.14%, respectively, when applied to a 3-way classification task involving AD, mild cognitive impairment, and normal control subjects. The results suggest that the utilization of ConvNeXt may have potential in the identification of AD using 2D slice images.

**Key words:** Alzheimer's disease, convolutional neural network, transformer, classification, magnetic resonance imaging

## 1. Introduction

Cognitive decline, memory loss, and other intellectual impairments are symptoms of Alzheimer's disease (AD), a degenerative brain disease that commonly affects individuals of advanced age. Dementia is a term used to represent a variety of symptoms associated with a mental decline sufficiently severe to cause daily difficulties. AD is a prominent cause of dementia, but it is still unclear what causes AD exactly. It is reported that in 2019 there were 57.4 million dementia sufferers worldwide; by 2050, that number might reach 152.8 million [1]. A considerable amount of research in recent years has focused on comprehending the underlying reasons for AD and creating efficient treatments. There is presently no medicine that can completely cure AD; instead, it can only temporarily alleviate symptoms. Effective clinical intervention and reducing disease progression depend on early detection [2].

Conventional diagnostic techniques for AD can involve a tedious and complex procedure of assessing symptoms, performing cognitive tests, and obtaining a medical history. Moreover, these approaches frequently rely on subjective evaluations, such as memory and cognitive tests, whose results might be influenced by anxiety, depression, and stress, making them less trustworthy. The diagnosis and treatment of AD could be dramatically impacted by automatic diagnosis. Automatic AD diagnosis can increase diagnostic accuracy by evaluating various data sources using machine learning and other artificial intelligence techniques. Effective disease management and therapy depend on early detection, which can also reduce the disease's progression.

The field of medicine has experienced rapid progress in the realm of artificial intelligence (AI), which led to the successful integration of various AI-assisted applications, particularly those related to classification [3], localization [4], and segmentation [5, 6]. Applying machine learning and other AI techniques for automatic AD diagnosis has become progressively more common. Early diagnosis is essential for effective disease management and therapy; therefore, the present research was motivated by the need for more precise and practical techniques for classifying individuals with AD. Creating algorithms that can interpret medical imaging data, like structural magnetic resonance imaging (MRI), to identify AD is considered a crucial area of research within this field.

MRI is a noninvasive imaging method that prevents the patient from being exposed to contrast agents or ionizing radiation. High-resolution images of the brain are provided by MRI, which possesses the capability to detect subtle alterations in brain structure associated with AD. AD can be identified using information on the structural, functional, and metabolic aspects of the brain that can be obtained from MRI. Repeated MRI scans make it possible to track an illness's development and evaluate a treatment's effectiveness. Further, in order to create models that can automatically detect and quantify the alterations in cerebral morphology and cognitive processes associated with AD, researchers are employing various machine learning methodologies, such as deep learning techniques, in their investigations.

Compared with 3D models, 2D ones usually have fewer parameters and need less time for learning [7]. In order to train a more generalized model, an affine transformation data augmentation technique was implemented in the present study. The primary disadvantage of the 2D slice-level technique was that, in contrast to 3D MRI, the slices of one subject were subjected to independent examination using 2D convolutional filters in most cases. As a result, among the slices of a subject, spatial information that can be crucial in classification may be lost. The aforementioned issue can be addressed by integrating data from several multiple slices.

The aim of the present research was to evaluate and compare the precision and effectiveness of models based on convolutional neural networks (CNNs), models based on vision transformers (ViTs), and their incorporated models for the purpose of diagnosing AD through the utilization of 2D MRI slices. The present study contains two research questions: How do CNN-based models compare to ViT-based models and their hybrid models with regard to accuracy for diagnosing AD adopting 2D MRI slices and what is the influence of varying perspectives, such as axial, coronal, and sagittal views, on the diagnosis of AD through the utilization of 2D MRI slices?

AD is a prevalent neurological disease that impacts a significant number of individuals globally. Timely intervention and treatment are contingent upon precise and prompt diagnosis. Through comparative analysis of various deep learning models, the optimal methodology for precise diagnosis of AD via 2D MRI slices can be identified. The utilization of ViT-based models has garnered considerable interest in diverse computer vision tasks. However, their implementation in the diagnosis of AD through the utilization of 2D MRI slices remains relatively underexplored. The integration of CNN-based models and ViT-based models in hybrid models holds promise for enhancing the diagnostic accuracy in AD. The effectiveness of hybrid approaches and

optimal configurations for future development and deployment can be assessed by comparing their accuracy and efficiency with those of standalone models. Through the use of 2D MRI slices, our comparative study seeks to provide important insights into the effectiveness of several deep-learning-based models in the diagnosis of AD. The findings of the present study could be a useful resource for researchers, clinicians, and developers seeking to identify the optimal strategy for achieving a precise and efficient diagnosis of AD. Ultimately, these insights have the potential to enhance patient care and improve health outcomes.

The novelties of the paper include: the application of a ViT-based model for AD diagnosis, a thorough examination of CNN-based, ViT-based, and their hybrid models, and an in-depth evaluation of the axial, coronal, and sagittal views of the MRI scan utilizing 2D MRI slices to diagnose AD. The manuscript is organized as follows: Section 1 presents a comprehensive overview of the contextual background pertaining to the diagnosis of AD. Section 2 offers a review of the pertinent literature that has been published in recent years. Section 3 includes a thorough examination of the utilization of the dataset, the processing of the data, the adoption of networks, and the implementation of experiments. Section 4 outlines the evaluation metrics employed in the present study. Section 5 highlights the findings and outcomes of the experiments we conducted. Section 6 outlines the constraints of the present study and provides insight into potential avenues for future research.

## 2. Related work

CNNs have started to be widely used in medical fields, which goes hand in hand with the prominence of deep learning in computer vision. Existing CNNs with outstanding success for natural image classification are of benefit in medical diagnosis. In particular, numerous reliable pretrained 2D CNN models can be employed in transfer learning, such as VGG [8], ResNet [9], DenseNet [10], and GoogLeNet [11]. In particular, various research has investigated the use of deep learning models in the analysis of 2D MRI slices in the identification of AD. In order to diagnose AD using 2D MRI slices, this section gives a thorough assessment of recent research on CNN- and ViT-based models. Previous research has shown the effectiveness of CNN-based models in analyzing medical images, including MRI scans. Valliani and Soni employed a CNN consisting of a single convolutional layer and two fully connected (FC) layers [12]. For each subject, only one axial slice was employed. The authors also adopted transfer learning through pretraining their CNN network on ImageNet [13]. In [14], Wen et al. performed a series of experiments on three distinct datasets: ADNI, Australian Imaging Biomarkers and Lifestyle Study of Ageing (AIBL) [15], and Open Access Series of Imaging Studies (OASIS) [16]. They took pretrained ResNet as the backbone, added an FC layer on top of it, and achieved an accuracy of 79%. In our research, the sagittal slices were retrieved and replicated into three channels of a fake red, green, and blue (RGB) image for each patient. Puente-Castro et al. proposed a ResNet-SVM hybrid model to classify sagittal slices of MRIs from ADNI and OASIS datasets [17]. SVM and ResNet were used as the classifier and the feature extractor. The features extracted from ResNet were concatenated with sex and age and then fed into the SVM. To create a three-channel image, each slice was replicated three times. Then the network was trained using the images generated and it achieved an average accuracy of 86.47% on OASIS and 78.72% on ADNI. Lim et al. examined a custom CNN, VGG, and ResNet to perform 3-way classification using pictures of the brain taken from the axial perspective of the MRI image [18]. The highest accuracy in their study was 80.66% achieved by VGG. Additionally, following the preprocessing stage, the data were transformed into a series of two-dimensional images. This process significantly decreases the size of the dataset from 37 GB to 260 MB.

These studies demonstrate the capability of CNN-based models in accurately diagnosing AD using 2D MRI slices. ViT-based models have been investigated in recent research as a potential alternative to CNN-based models for medical image analysis. Bedel et al. presented a transformer-based model BolT with cross-window attention and regularization for fMRI blood–oxygen-level-dependent response analysis [19]. BolT has high efficiency in extracting features that range from local to global, which enables effective performance in detecting tasks. Sarraf et al. proposed an optimized vision transformer (OViTAD) based on a vision transformer for AD prediction using 2D MRI axial slices [20]. OViTAD achieves the same level of performance but uses a reduced number of parameters in contrast to the vanilla vision transformer. The OViTAD model achieved an average accuracy of 89.48% in a 3-way classification task. Their research indicates that models based on vision transformers possess the capability to offer an alternative method for diagnosing AD by utilizing 2D MRI slices. Despite the limited use of ViT-based models in the diagnosis of AD, current research efforts have started to look at their potential benefits.

In general, the existing literature demonstrates the efficacy of CNN-based models in the context of diagnosing AD through the utilization of 2D MRI slices. Additionally, it is conceivable that ViT-based models or Vit–CNN hybrid models could potentially function as replacements for CNNs within this domain.

## 3. Materials and methods

### 3.1. Dataset
The ADNI dataset (https://adni.loni.usc.edu/) was utilized in our research. Its objective is to better understand how MRI, PET, and other biological indications, in addition to clinical and neuropsychological testing, can be used to diagnose MCI and early AD. The dataset included 188 AD, 401 mild cognitive impairment (MCI), and 229 normal control (NC) subjects. In total, 4174 MRI scans of 818 participants from the database were used in the study. Only the standard 1.5 T T1-weighted sMRI data were used. The original dimensions of the raw MRI images were $256 \times 256 \times 256$.

### 3.2. Preprocessing
A standard pipeline of preprocessing was implemented to preprocess the MRI images [21, 22]. The preprocessing pipeline includes orientation, registration, skull stripping, bias field correction, image enhancement, and intensity normalization. The overall preprocessing workflow is shown in Figure 1. Using the orientation tool in the FMRIB Software Library (FSL) [23], an image can be rotated to make it align with the orientation of the common template images (MNI152 template), making them appear to be "the same way around." In order to ensure the spatial correspondence of anatomy across distinct images, image registration enables multiple images to be aligned into one integrated image. Image rotation, skew, and scale are common issues when overlaying images that can be resolved by registration. FLIRT (FMRIB's Linear Image Registration Tool) in FSL was utilized for brain image registration. Skull stripping is an essential step in the process of identifying brain concerns. It involves separating brain tissue from other tissue types on an MRI brain scan. Accurate skull stripping is the key to performing the subsequent neuroimage analysis. The Brain Extraction Tool (BET) in FSL was utilized in the present study for skull removal [24]. Bias field correction is a method that has been developed to eliminate this intensity gradient from the image. The N4 technique for bias field correction is frequently applied for addressing the bias field in MRI image data. The N4 algorithm offered by ANTs was utilized in the present research [25]. Image enhancement is the technique of modifying an image to improve its visual impact by modifying the brightness levels of the pixels. A few image enhancement techniques were

implemented to provide better input for the model. First, a median filter was used to remove noise from images, then 0.5% and 99.5% of the value of each image were taken as the minimum and maximum pixel value for rescaling, and lastly histogram equalization was used to improve contrast in the images. An image is scaled and shifted during normalization so that each pixel has a mean and variance of 0 and 1, respectively.
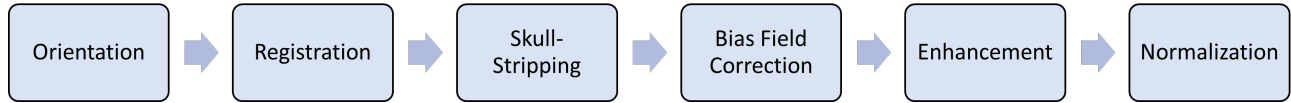
| Orientation | ➜ | Registration | ➜ | Skull-Stripping | ➜ | Bias Field Correction | ➜ | Enhancement | ➜ | Normalization |

**Figure 1**. The overall workflow of preprocessing.

### 3.2.1. Generation of 2D slices from 3D MRIs

Each 3D MRI image was sliced to generate slices from three perspectives, i.e. axial, coronal, and sagittal. After preprocessing, the middlemost slices tend to be the most informative slices of the image and contain the most significant information entropy. When choosing slices for classification, the slices in the middle should be used. First, select the middlemost slice of the nonzero part of the image. Then take two more slices from a few steps away before and after to compensate for missing 3D data. This step number is a hyperparameter, and was taken as five in our research. Consider the three slices mentioned above as the three channels of a three-channel image and stack them together to compose a fake RGB image. As a result, each MRI image will generate an axial, a coronal, and a sagittal slice.

### 3.2.2. Data augmentation

Data augmentation is a technique that, without generating new data, significantly broadens the range of data that are easily available for training models. To artificially extend the training set, data augmentation is the process of altering existing data to produce changed copies of datasets. To produce reliable predictions, deep learning models usually require an adequate quantity of training data, which is not always available. As a result, additional data are added to the original data to create a more broadly applicable model. The two categories of data enhancement techniques are position augmentation and color augmentation. A picture's pixel positions are altered through position augmentations. Position augmentation includes scaling, flipping, cropping, rotation, padding, affine transformation, translation, etc. Color augmentation is an approach to changing the color properties of an image by modifying its pixel values. Color augmentation consists of brightness, contrast, saturation, etc. Specifically, contrast brightness, contrast, random flipping, random affine, random blur, and random noise were used in the present study. Each data augmentation approach is implemented dynamically, which means when loading an image from a disk an augmented image will be generated. The augmented image will be resized to $224 \times 224$ before being fed into a model.

### 3.3. Network architecture

The performance of multiple cutting-edge models for diagnosing AD using 2D MRI slices was evaluated and compared in the present work. Three categories of models, i.e. CNN-based, ViT-based, and hybrid, were chosen for analysis. Specifically, the models considered for analysis include ResNet, ConvNeXt, CaiT, Swin Transformer (Swin-T), and CVT. These models were selected due to their success in the computer vision field and their potential for assisting in a precise diagnosis of AD. The architectures of these models are shown in the following sections. Some models, like vanilla ViT, were excluded due to the limited computational resources.

### 3.3.1. CNNs

CNNs are deep learning models designed specifically for processing and analyzing data like images. As CNNs can automatically extract and learn complex hierarchical features from images, they have drastically changed computer vision. Convolutional, pooling, and FC layers are basic components of CNNs. Convolutional layers analyze the input image for local patterns and features using filters and convolutions. Pooling layers allow extraction of the most valuable information while increasing computing efficiency and reducing the spatial dimensions of the feature maps. Finally, FC layers integrate the extracted features and make the final prediction or classification.

**ResNet** When the depth of CNNs reaches a particular threshold, the gradient disappears, which causes the accuracy to drop rather than rise. ResNet solves this issue by introducing residual connections. By skipping some intermediary levels and connecting the layer to succeeding layers, the residual connection forms a residual block. ResNet is constructed by stacking these residual building blocks. This type of skip connection, or identity mapping, has the advantage that regularization will not include any layer that impairs architecture effectiveness. As a result, vanishing or exploding gradient problems are not encountered while training very deep neural networks.

In the present research, the pretrained ResNet on ImageNet, ResNet18, ResNet34, and ResNet50, were adopted. Since the latest PyTorch version provides two pretrained weights for ResNet50, the newer one was employed. ResNet34 and ResNet50 share a similar architecture with ResNet18 but contain different numbers of residual blocks. Initially, the last dense layer of ResNet has an output dimension of 1000, but, in our study, we modified the output dimension to 3, aligning it with the specific task of AD diagnosis. As an example, the architecture of ResNet18 is depicted in Figure 2.
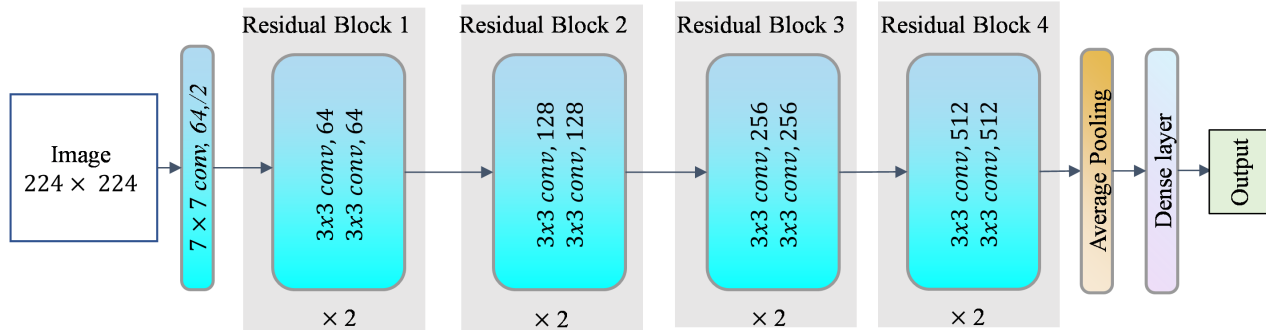


**Figure 2**. Architecture of ResNet18. ResNet34 and ResNet50 share a similar architecture with ResNet18 but contain different numbers of residual blocks. The original output dimension of the last dense layer is 1000 and we changed it to 3 in the present study.

**ConvNeXt** Considering that the ViT has outperformed CNNs in numerous tasks of computer vision, the author modified a conventional ResNet by incorporating the design of the Swin-T [26]. Through this process, the author identified notable performance differences. The ConvNeXt model mimics the patching approach of the Swin-T and substitutes the ResNet-style stem cell with a patchy layer. Specifically, a large kernel with a correspondingly large stride was utilized to ensure that there was no overlap among the sliding windows.

These sliding windows exhibit comparable behavior to the patches in ViT. ConvNeXt also modifies the number of blocks within every stage following Swin. The utilization of depthwise convolution in ConvNeXt bears a resemblance to the weighted summation process observed in self-attention. Last, moving up the depthwise convolutional layer and utilizing larger convolutional kernel sizes were performed to enhance the global receptive field. In the present study, ConvNeXt-tiny and ConvNeXt-small were utilized.

### 3.3.2. ViT

In [27], Vaswani et al. proposed transformer architecture to solve issues in the field of natural language processing (NLP). The transformer is introduced and explained with an encoder–decoder architecture and becomes the foundation for many state-of-the-art NLP models. The transformer now holds a dominant position in the NLP field, and more and more research is being done to try to apply it in the realm of computer vision. One of the transformer's merits is that it excels at handling a wide variety of inputs. Additionally, the convolution operation mainly considers local neighbors, which leads to global information being missed. In contrast, the attention model is very adept at modeling lengthy periods as shown in Equation (1).

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V, \tag{1}$$

where query, key, and value are abbreviated as Q, K, and V.

The ViT model that Dosovitskiy proposed uses the transformer model framework from NLP to tackle all of the challenges in computer vision [28]. A transformer may be used in the image domain by inputting a series of tokens into the bottom layer of the transformer, using ViT's proprietary technology, which is analogous to NLP in image processing. Specifically, the image is separated into several parts, each of which is then squeezed and mapped into a 1D vector with a fixed dimension using a neural network. The converted 1D vector is subsequently fed into the transformer encoder. However, the drawback of ViT is splitting the image into patches, resulting in a lower-resolution output. In addition, the transformer model's computational cost grows with the sequence's length, and direct application of pixel-level prediction tasks can lead to a surge in computation and memory consumption.

**Swin Transformer** Liu et al. proposed the Swin-T and achieved a better speed–accuracy trade-off than with vanilla ViT [29]. Local attention is employed in the Swin-T to divide patches into windows, and interpatch attention is performed only within the windows to improve efficiency. However, there would be no information interaction between the patches of different windows. The Swin-T proposed a shifted window, borrowing from the sliding window approach, which used different window configurations in different layers to address this concern. The window positions are shifted horizontally and vertically by several patches, allowing the patches within different windows to interact with information from different layers. The multihead self-attention (MSA) block utilized in the ViT architecture is substituted with the Window and Shifted Window MSA block. In the present study, Swin-T, Swin Transformer Small (Swin-S), and Swin Transformer Big (Swin-B) were utilized. Similar to the method implemented on Resnet, the output dimension of the last dense layer was also revised to 3 and the rest of the layers were kept. An overview of a Swin-T's structure is shown in Figure 3.
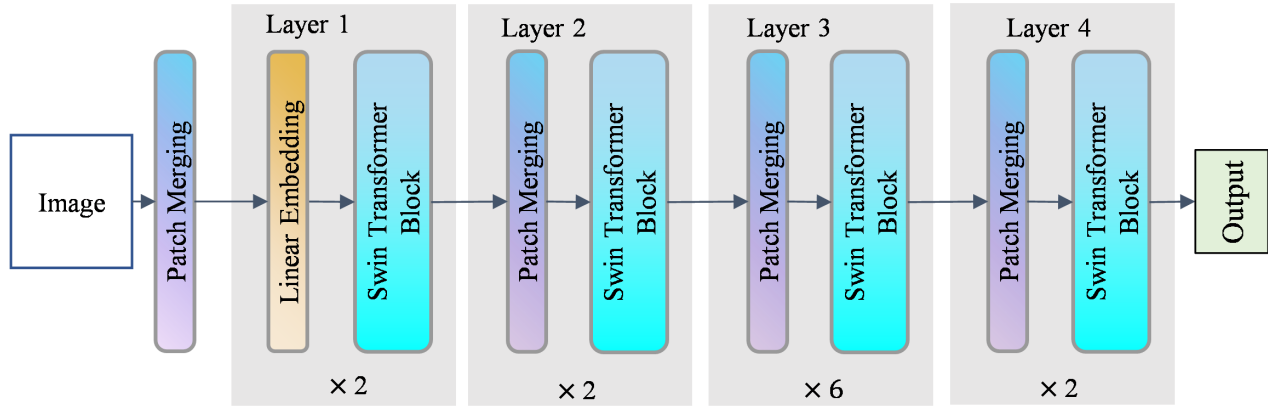
**Figure 3**. The architecture of a Swin Transformer (Swin-T).

**Class-attention in image transformers (CaiT)**  CaiT adds per-channel weighting (a diagonal learnable matrix) to each residual block's output [30] as shown in Equation (2).

$$x'_l = x_l + \text{diag}\left(\lambda_{l,1}, \ldots, \lambda_{l,d}\right) \times \text{SA}\left(LN\left(x_l\right)\right)$$

$$x_{l+1} = x'_l + \text{diag}\left(\lambda'_{l,1}, \ldots, \lambda'_{l,d}\right) \times \text{FFN}\left(LN\left(x'_l\right)\right), \tag{2}$$

where LN represents the LayerNorm operator, FFN stands for the feed-forward network, SA is for self-attention, and $\text{diag}\left(\lambda_{l,1}, \ldots, \lambda_{l,d}\right)$ stands for the learnable diagonal matrix to assign weights for each channel. The use of class embeddings is postponed compared with the ViT because, in the shallow layers, semantic information about classification is merely extracted. CaiT utilizes a separate set of attention layers called class attention (CA) to simulate the communication between the representations of the class token and the image patch. In the present study, CaiT-S36 was utilized. The architecture of CaiT is shown in Figure 4.
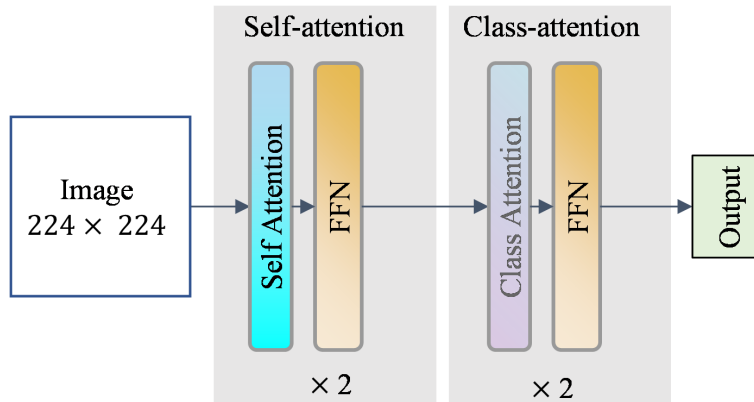


**Figure 4**. The architecture of CaiT.

### 3.3.3. Hybrid models

**Convolutional vision transformer (CVT)** CVT is a deep learning model that combines convolutional layers and transformers, providing a hybrid architecture for vision tasks [31]. In CVT, overlapping patches are initially created from the input image. Position encoding may not be required due to the presence of overlapping tokens. Rather than directly inputting the patches into a transformer encoder, CVT integrates the convolutional token embedding blocks to construct a model that captures the spatial context. Moreover, the linear projection utilized in the ViT is substituted with a convolutional projection to attain supplementary modeling of the local spatial context. The present study employed CvT-13 as one of the chosen models and the architecture of CvT is shown in Figure 5.
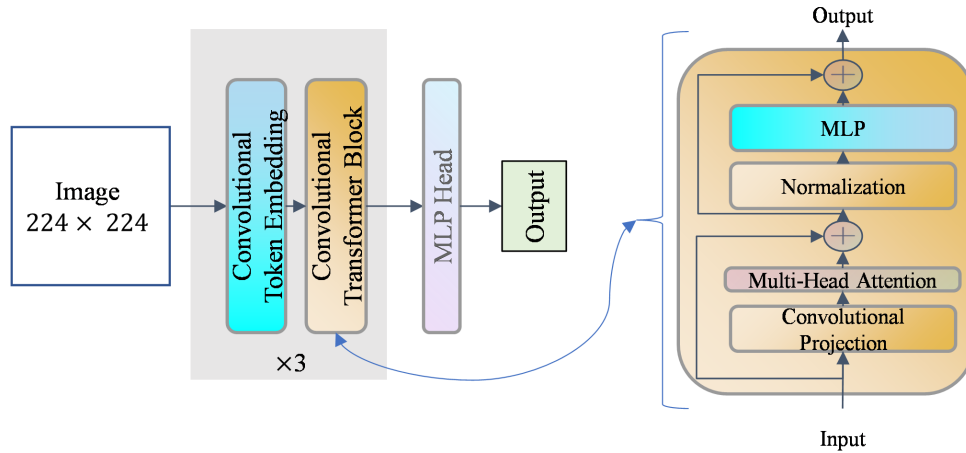


**Figure 5**. The architecture of CVT.

### 3.4. Network training

The current research entailed the evaluation and comparison of five distinct models, namely ResNet, Swin-T, ConvNeXt, CaiT, and CvT. Three experiments were conducted for each model, utilizing axial, sagittal, and coronal slices. Cross-validation was used to train all of the models using the ADNI dataset. Firstly, the whole dataset was randomly split into the test and nontest datasets with a ratio of $1:9$. Then the nontest dataset was further divided into ten folds of equal size. Nine of the ten folds were used for training, while the remaining one was used for validation. The AdamW optimizer was used for training, utilizing an initial learning rate of $5e-5$, and the batch size was configured to 32. Weight decay and momentum were set to $1e-4$ and 0.9. Since the dataset was unbalanced, the cross-entropy loss was applied along with manually adjusted weights assigned to each class according to Equation (3).

$$\text{weight } (x) = \frac{\text{training examples}}{\text{classes } \times \text{ training examples class } x} \tag{3}$$

All MRI images were processed and models were trained on a workstation equipped with an Intel Core i5 16-core 3.69 GHz CPU and a 12GB NVIDIA GeForce GTX 3080ti GPU. The operating system of this server was Ubuntu 20.04.3 LTS. Python 3.9.7 was used for preprocessing and model development. FMRIB Software Library v6.0 (FSL) was used for all phases of the MRI processing workflow.

## 4. Results

### 4.1. Evaluation metrics

Each model's accuracy, precision, sensitivity, and F1 score were computed to evaluate the performance of the classification. All models' performance measures were presented as a mean value across five cross-validation folds.

The evaluation of the classification performance was conducted using four metrics: classification accuracy, precision, sensitivity, and F1 score, as defined by Equations (4), (5), (6), and (7):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Sensitivity = Recall = \frac{TP}{TP + FN} \tag{6}$$

$$F1 = \frac{2 * TP}{2 * TP + FP + FN}, \tag{7}$$

where TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative, respectively.

### 4.2. Comparative analysis

The overall performance of the models is shown in Table 1. The deep learning models performed separate training on the axial, coronal, and sagittal views, and their performance was evaluated through the use of evaluation metrics consisting of accuracy, sensitivity, specificity, and F1 score. ConvNeXt-tiny demonstrated the best performance among the examined models in the comparative analysis that was conducted when evaluated with axial slices. Specifically, ConvNeXt-tiny achieved an average accuracy of 95.74%, precision of 96.71%, sensitivity of 95.74%, and F1 score of 96.14%. As shown in Table 2, the finding implied that, in terms of the given evaluation metrics, CNN-based models outperformed the other alternatives. The results in Table 3 indicated that the axial view exhibited superior accuracy compared to the other two perspectives.

## 5. DISCUSSION

The result of our comparative study shows that ViT-based models do not perform as well as CNN-based networks on small to medium-sized datasets[32]. When dealing with a medical dataset of the size of ADNI, it is recommended to use a CNN-based model rather than a ViT-based model. CNNs are developed to be capable of identifying regional patterns and spatial data, which is advantageous for image-based jobs. In contrast, ViTs rely heavily on self-attention mechanisms, making them more appropriate for larger datasets containing an abundance of data. In comparison to ViTs, CNNs often have fewer parameters, making them more parameter-efficient.

The axial view of MRI scans typically contains less nonbrain area and is easier to remove during brain extraction compared with the sagittal and coronal views. As a result, the axial view images tend to contain a reduced level of noise. Furthermore, axial slices effectively depict several prominent brain regions implicated with AD, including the hippocampus and entorhinal cortex. A noteworthy limitation of our study pertains to the comparatively small size of the dataset employed. The restricted size of the dataset utilized in the present study means that it may not comprehensively involve the diverse and intricate characteristics of AD cases, which

may restrict the generality of our results. Although the deep learning models employed in the present study have exhibited potential in diverse computer vision assignments, they might not entirely grasp the temporal or progressive characteristics of AD.

Furthermore, there currently exist general limitations within the realm of automated diagnosis of AD using deep learning. Several existing models for automatically diagnosing AD rely heavily on data from MRI and PET scans and other types of medical imaging. Automatic AD diagnosis using machine learning approaches can provide hard-to-understand and -interpret models, making it difficult for medical professionals to comprehend how the models yielded a specific diagnosis and to utilize this knowledge to guide treatment choices. Most models have yet to be tested in real-world situations, where data complexity and unpredictability might be significantly higher than in controlled laboratory conditions. Because of this, evaluating the precision and generalizability of these models in a clinical situation is challenging. Moreover, while the data used for training and testing models may contain sensitive information about individuals, using machine learning models for autonomous AD detection raises ethical and privacy issues. Researchers must ensure the data are gathered and used ethically and in accordance with applicable privacy laws. To summarize, the lack of diversity in datasets, reliance on imaging data, limited interpretability of models, poor validation in real-world settings, and ethical and privacy concerns all restrict current research on automatic AD detection.

The adoption of various datasets and an increase in the sample size are further options for improvement. Assemble models that take multiple slices may improve the performance further. Several variables relevant to AD detection are MRI, CT, PET, neurological examinations, cognitive or blood tests, sex, age, the pattern of speech, retinal abnormalities, $\alpha\beta$ protein, mini-mental state examination, Clinical Dementia Rating score, logical memory test, genes, etc. [33]. Since multimodality of input may provide complementary information, multimodal models that integrate more than one variable mentioned above may be helpful in future comprehensive diagnostics. In order to increase the accuracy and reliability of the diagnosis, research on automatic AD diagnosis is generally moving towards the development of more complex and accurate algorithms and the integration of multiple data sources.

A diffusion-based model proposed by Bedel and Çukur for fMRI interpretation also provides a new perspective for future research [34]. There is still more to be accomplished in this field and more research is required to create better systems for the early and precise identification of AD.

## 6. CONCLUSION

In the present study, using preprocessed MRI brain slices collected from the ADNI database, a series of experiments were conducted using CNN-based, ViT-based, and their hybrid architectures. ConvNeXt-tiny showed the best performance among the studied models in the comparative analysis carried out and evaluated with axial slices. CNN-based models performed better than the other models. Compared to the other two perspectives, the axial view demonstrated higher accuracy. These findings add substantial insight to the field and show that CNN-based models remain a solid method for establishing a precise and effective AD diagnosis. Additionally, axial slices emphasize how crucial it is to take into account the slice orientation when utilizing 2D MRI slices to diagnose AD.

**Table 1**. The number of parameters, testing accuracy, precision, recall, and F1 score for all models and slice types. The superscripts [a], [b], and [c] indicate the model was tested on ADNI, AIBL, and OASIS. - stands for not specified. Boldface indicates the highest-performing model in terms of each metric.

| Model | Params | Slice | Accuracy | Precision | Recall | F1 score | GFLOPS |
|---|---|---|---|---|---|---|---|
| ResNet[a,b,c][14] | 11.7M | Axial | 79 | - | - | - | 4.09 |
| ResNet-SVM[a][17] | - | Sagittal | 78.72 | 68.96 | 58.66 | 60.79 | - |
| ResNet-SVM[c][17] | - | Sagittal | 86.47 | 30.75 | 35.25 | 32.07 | - |
| VGG[a][18] | 138.4M | Axial | 78.54 | 78.74 | 78.56 | 78.49 | 15.47 |
| OViTAD[a][20] | 38.4M | Axial | 89.49 | 89.20 | 90.02 | 88.79 | - |
| ResNet18[a] | 11.7M | Axial | 91.87 | 91.70 | 91.87 | 91.72 | 1.81 |
| | | Coronal | 89.54 | 92.62 | 89.54 | 90.86 | |
| | | Sagittal | 86.47 | 88.86 | 86.47 | 87.51 | |
| ResNet34[a] | 21.8M | Axial | 90.89 | 92.26 | 90.89 | 91.51 | 3.66 |
| | | Coronal | 87.43 | 87.79 | 87.43 | 87.56 | |
| | | Sagittal | 85.53 | 85.23 | 85.53 | 85.36 | |
| ResNet50[a] | 25.6M | Axial | 93.08 | 94.64 | 93.08 | 93.73 | 4.09 |
| | | Coronal | 91.30 | 93.77 | 91.30 | 92.02 | |
| | | Sagittal | 86.74 | 90.85 | 86.74 | 88.06 | |
| Swin-T[a] | 28.3M | Axial | 88.93 | 88.66 | 88.93 | 88.18 | 4.49 |
| | | Coronal | 88.70 | 89.75 | 88.70 | 88.69 | |
| | | Sagittal | 85.61 | 89.46 | 85.61 | 86.43 | |
| Swin-S[a] | 49.6M | Axial | 89.20 | 90.19 | 89.20 | 89.50 | 8.74 |
| | | Coronal | 89.37 | 89.72 | 89.37 | 89.10 | |
| | | Sagittal | 87.50 | 89.57 | 87.50 | 87.96 | |
| Swin-B[a] | 87.8M | Axial | 88.24 | 89.40 | 88.24 | 88.59 | 15.43 |
| | | Coronal | 88.88 | 89.67 | 88.88 | 88.74 | |
| | | Axial | **95.74** | **96.71** | **95.74** | **96.14** | |
| ConvNeXt-tiny[a] | 28.6M | Coronal | 92.03 | 93.10 | 92.03 | 92.38 | 4.46 |
| | | Sagittal | 90.79 | 93.75 | 90.79 | 91.78 | |
| | | Axial | 92.85 | 92.19 | 92.85 | 92.38 | |
| ConvNeXt-small[a] | 50.2M | Coronal | 92.21 | 91.33 | 92.21 | 91.35 | 8.68 |
| | | Sagittal | 87.43 | 88.66 | 87.43 | 87.20 | |
| | | Axial | 82.14 | 84.12 | 82.14 | 82.37 | |
| CaiT-S36[a] | 68M | Coronal | 74.93 | 72.85 | 74.93 | 71.41 | 48 |
| | | Sagittal | 82.50 | 81.93 | 82.50 | 81.72 | |
| | | Axial | 86.28 | 84.87 | 86.28 | 84.49 | |
| CvT-13[a] | 20M | Coronal | 81.20 | 90.12 | 81.20 | 83.64 | 4.53 |
| | | Sagittal | 78.02 | 76.12 | 79.92 | 76.80 | |

**Table 2**. Performance metrics comparison of CNN-based, ViT-based, and hybrid models for Alzheimer's disease diagnosis. The measurements are presented as the mean ± standard deviation (std) value calculated for the same type of models.

| Model | Accuracy | Precision | Recall | F1 score |
|---|---|---|---|---|
| CNN-based | 90.26 ± 2.95 | 91.56 ± 2.93 | 90.26 ± 2.95 | 90.64 ± 2.89 |
| ViT-based | 86.01 ± 4.30 | 86.91 ± 5.12 | 86.01 ± 4.30 | 85.74 ± 5.19 |
| Hybrid | 81.83 ± 4.17 | 83.70 ± 7.07 | 82.47 ± 3.36 | 81.64 ± 4.22 |

**Table 3**. Performance metrics comparison of accuracy, precision, recall, and F1 score for axial, coronal, and sagittal slices in Alzheimer's disease diagnosis. The measurements are presented as the mean $\pm$ standard deviation (std) value calculated for the same type of models.

| Perspective | Accuracy | Precision | Recall | F1 score |
|---|---|---|---|---|
| Axial | $89.92 \pm 3.88$ | $90.47 \pm 3.95$ | $89.92 \pm 3.88$ | $89.86 \pm 4.17$ |
| Coronal | $87.56 \pm 5.43$ | $89.07 \pm 5.99$ | $87.56 \pm 5.43$ | $87.58 \pm 6.23$ |
| Sagittal | $85.67 \pm 3.39$ | $87.20 \pm 5.01$ | $85.86 \pm 2.93$ | $85.89 \pm 4.06$ |

# References

[1] Nichols E, Steinmetz J, Vollset S, Fukutaki K, Chalek J et al. Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease 2019. The Lancet Public Health 2022; 7 (2): e105-e125. https://doi.org/10.1016/S2468-2667(21)00249-8

[2] Livingston G, Huntley J, Sommerlad A, Ames D, Ballard C et al. Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. The Lancet 2020; 396 (10248): 413-446. https://doi.org/10.1016/S0140-6736(20)30367-6

[3] Yeoh P, Lai K, Goh S, Hasikin K, Hum Y et al. Emergence of deep learning in knee osteoarthritis diagnosis. Computational Intelligence in Image and Video Analysis 2021; 2021: 4931437. https://doi.org/10.1155/2021/4931437

[4] Anis S, Lai K, Chuah J, Ali S, Mohafez H et al. An overview of deep learning approaches in chest radiograph. IEEE Access 2020; 8: 182347-182354. https://doi.org/10.1109/ACCESS.2020.3028390

[5] Jahanzad Z, Liew Y, Bilgen M, McLaughlin R, Leong C et al. Regional assessment of LV wall in infarcted heart using tagged MRI and cardiac modelling. Physics in Medicine & Biology 2015; 60 (10): 4015-4031. https://doi.org/10.1088/0031-9155/60/10/4015

[6] Chai H, Wee L, Swee T, Salleh S, Chea L. An artifacts removal post-processing for epiphyseal region-of-interest (EROI) localization in automated bone age assessment (BAA). BioMedical Engineering OnLine 2011; 10: 87. https://doi.org/10.1186/1475-925X-10-87

[7] Zhao Z, Chuah J, Lai K, Chow C, Gochoo M et al. Conventional machine learning and deep learning in Alzheimer's disease diagnosis using neuroimaging: a review. Frontiers in Computational Neuroscience 2023; 17. https://doi.org/10.3389/fncom.2023.1038636

[8] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. ArXiv Preprint ArXiv:1409.1556 2014.

[9] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Los Alamitos, CA, USA; 2016. pp. 770-778. https://doi.org/10.1109/CVPR.2016.90

[10] Huang G, Liu Z, Maaten L, Weinberger K. Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Los Alamitos, CA, USA; 2017. pp. 2261-2269. https://doi.org/10.1109/CVPR.2017.243

[11] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S. Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Los Alamitos, CA, USA; 2015. pp. 1-9. https://doi.org/10.1109/CVPR.2015.7298594

[12] Valliani A, Soni A. Deep residual nets for improved Alzheimer's diagnosis. In: Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics; New York, NY, USA; 2017. p. 615. https://doi.org/10.1145/3107411.3108224

[13] Deng J, Dong W, Socher R, Li L, Li K et al. ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops); Los Alamitos, CA, USA; 2009. pp. 248-255. https://doi.org/10.1109/CVPR.2009.5206848

[14] Wen J, Thibeau-Sutre E, Diaz-Melo M, Samper-Gonzalez J, Routier A et al. Convolutional neural networks for classification of Alzheimer's disease: overview and reproducible evaluation. Medical Image Analysis 2020; 63: 101694. https://doi.org/10.1016/j.media.2020.101694

[15] Ellis K, Bush A, Darby D, De Fazio D, Foster J et al. The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. International Psychogeriatrics 2009; 21 (4): 672-687. https://doi.org/10.1017/S1041610209009405

[16] Marcus D, Wang T, Parker J, Csernansky J, Morris J et al. Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. Journal Of Cognitive Neuroscience 2007; 19 (9): 1498-1507. https://doi.org/10.1162/jocn.2007.19.9.1498

[17] Puente-Castro A, Fernandez-Blanco E, Pazos A, Munteanu C. Automatic assessment of Alzheimer's disease diagnosis based on deep learning techniques. Computers in Biology and Medicine 2020; 120: 103764. https://doi.org/10.1016/j.compbiomed.2020.103764

[18] Lim B, Lai K, Haiskin K, Kulathilake K, Ong Z et al. Deep learning model for prediction of progressive mild cognitive impairment to Alzheimer's disease using structural MRI. Frontiers in Aging Neuroscience 2022; 14: 876202. https://doi.org/10.3389/fnagi.2022.876202

[19] Bedel H, Sivgin I, Dalmaz O, Dar S, Cukur T. BolT: fused window transformers for fMRI time series analysis. Medical Image Analysis 2023; 88: 102841. https://doi.org/10.1016/j.media.2023.102841

[20] Sarraf S, Sarraf A, DeSouza D, Anderson J, Kabia M et al. OViTAD: optimized vision transformer to predict various stages of Alzheimer's disease using resting-state fMRI and structural MRI data. Brain Sciences 2023; 13 (2): 260. https://doi.org/10.3390/brainsci13020260

[21] Ge C, Qu Q, Gu I, Jakola A. Multiscale deep convolutional networks for characterization and detection of Alzheimer's disease using MR images. In: 2019 IEEE International Conference on Image Processing (ICIP); Taipei, Taiwan; 2019. pp. 789-793. https://doi.org/10.1109/ICIP.2019.8803731

[22] Qiu S, Miller M, Joshi P, Lee J, Xue C et al. Multimodal deep learning for Alzheimer's disease dementia assessment. Nature Communications 2022; 13: 3404. https://doi.org/10.1038/s41467-022-31037-5

[23] Jenkinson M, Beckmann C, Behrens T, Woolrich M, Smith S. FSL. NeuroImage 2012; 62 (2): 782-790. https://doi.org/10.1016/J.NEUROIMAGE.2011.09.015

[24] Smith S. Fast robust automated brain extraction. Human Brain Mapping 2002; 17 (3): 143-155. https://doi.org/10.1002/hbm.10062

[25] Tustison N, Avants B, Cook P, Zheng Y, Egan A et al. N4ITK: improved N3 bias correction. IEEE Transactions on Medical Imaging 2010; 29 (6): 1310-1320. https://doi.org/10.1109/TMI.2010.2046908

[26] Liu Z, Mao H, Wu C, Feichtenhofer C, Darrell T et al. A ConvNet for the 2020s. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); New Orleans, LA, USA; 2022. pp. 11966-11976. https://doi.org/10.1109/CVPR52688.2022.01167

[27] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L et al. Attention is all you need. Advances In Neural Information Processing Systems 2017; 30.

[28] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X et al. An image is worth 16x16 words: transformers for image recognition at scale. In: International Conference On Learning Representations 2021.

[29] Liu Z, Lin Y, Cao Y, Hu H, Wei Y et al. Swin Transformer: hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV); Montreal, QC, Canada; 2021. pp. 9992-10002. https://doi.org/10.1109/ICCV48922.2021.00986

[30] Touvron H, Cord M, Sablayrolles A, Synnaeve G, Jegou H. Going deeper with image transformers. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV); Montreal, QC, Canada; 2021. pp. 32-42. https://doi.org/10.1109/ICCV48922.2021.00010

[31] Wu H, Xiao B, Codella N, Liu M, Dai X et al. CvT: introducing convolutions to vision transformers. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV); Montreal, QC, Canada; 2021. pp. 22-31. https://doi.org/10.1109/ICCV48922.2021.00009

[32] Zhu H, Chen B, Yang C. Understanding why ViT trains badly on small datasets: an intuitive perspective. ArXiv Preprint ArXiv:2302.03751 2023.

[33] Cuingnet R, Gerardin E, Tessieras J, Auzias G, Lehericy S et al. Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. NeuroImage 2011; 56 (2): 766-781. https://doi.org/10.1016/j.neuroimage.2010.06.013

[34] Bedel H, Cukur T. DreaMR: diffusion-driven counterfactual explanation for functional MRI. ArXiv Preprint ArXiv:2307.09547 2023.