Pattern Analysis and Applications (2024) 27:76
https://doi.org/10.1007/s10044-024-01297-6

THEORETICAL ADVANCES

Computer-aided diagnosis of Alzheimer's disease and neurocognitive disorders with multimodal Bi-Vision Transformer (BiViT)

S. Muhammad Ahmed Hassan Shah[1,2] · Muhammad Qasim Khan[2] · Atif Rizwan[3] · Sana Ullah Jan[4] · Nagwan Abdel Samee[5] · Mona M. Jamjoom[6]

Received: 2 August 2023 / Accepted: 14 June 2024

Abstract

Cognitive disorders affect various cognitive functions that can have a substantial impact on individual's daily life. Alzheimer's disease (AD) is one of such well-known cognitive disorders. Early detection and treatment of cognitive diseases using artificial intelligence can help contain them. However, the complex spatial relationships and long-range dependencies found in medical imaging data present challenges in achieving the objective. Moreover, for a few years, the application of transformers in imaging has emerged as a promising area of research. A reason can be transformer's impressive capabilities of tackling spatial relationships and long-range dependency challenges in two ways, i.e., (1) using their self-attention mechanism to generate comprehensive features, and (2) capture complex patterns by incorporating global context and long-range dependencies. In this work, a Bi-Vision Transformer (BiViT) architecture is proposed for classifying different stages of AD, and multiple types of cognitive disorders from 2-dimensional MRI imaging data. More specifically, the transformer is composed of two novel modules, namely Mutual Latent Fusion (MLF) and Parallel Coupled Encoding Strategy (PCES), for effective feature learning. Two different datasets have been used to evaluate the performance of proposed BiViT-based architecture. The first dataset contain several classes such as mild or moderate demented stages of the AD. The other dataset is composed of samples from patients with AD and different cognitive disorders such as mild, early, or moderate impairments. For comprehensive comparison, a multiple transfer learning algorithm and a deep autoencoder have been each trained on both datasets. The results show that the proposed BiViT-based model achieves an accuracy of 96.38% on the AD dataset. However, when applied to cognitive disease data, the accuracy slightly decreases below 96% which can be resulted due to smaller amount of data and imbalance in data distribution. Nevertheless, given the results, it can be hypothesized that the proposed algorithm can perform better if the imbalanced distribution and limited availability problems in data can be addressed.

✉ Sana Ullah Jan
s.jan@napier.ac.uk

[1] Medical Imaging and Diagnostics Laboratory (MIDL), National Center of Artificial Intelligence (NCAI), COMSATS University Islamabad, Islamabad 44000, Pakistan

[2] Department of Computer Science, COMSATS University Islamabad, Attock Campus, Islamabad 43600, Pakistan

[3] Department of Computer Engineering, Jeju National University, Jejusi 63243, Jeju Special Self-Governing Province, Republic of Korea

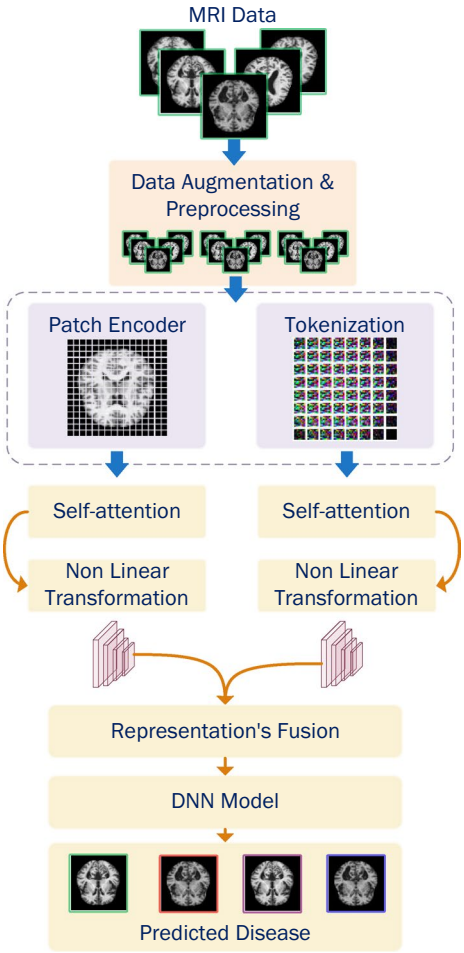
[4] School of Computing Engineering and the Built Environment, Edinburgh Napier University, Edinburgh EH10 5DT, UK

[5] Department of Information Technology, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, 11671 Riyadh, Saudi Arabia

[6] Department of Computer Sciences, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, 11671 Riyadh, Saudi Arabia

Published online: 01 July 2024

Graphical abstract



Keywords Vision transformers · Deep learning · Computer vision · Medical image processing · Alzheimer disease · Cognitive disorders

1 Introduction

Cognitive disorders have a significant impact on an individual's daily life, as they affect various cognitive functions, and Alzheimer's disease (AD) is one of the most commonly known cognitive disorders. AD and other cognitive disorders can be diagnosed and treated commonly through medical imaging. Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), and Computed Tomography (CT) allow medical professionals to identify neurological changes linked to these disorders by offering comprehensive visual depictions of brain structures and functions. In this section, cognitive disorders are discussed in general followed by analysis of deep learning methods for diagnosing them. Next, AD and different AI approaches used to diagnose it are presented. Then, the proposed methodology is summarized followed by the motivation behind the research, and finally, the significant contributions made in this study are highlighted.

1.1 Cognitive disorders

Cognitive impairments [32] refer to difficulties or limitations in cognitive function, which can include memory, attention, perception, language, or problem-solving abilities. These impairments can affect a persons daily life and activities, and can range from mild to severe. Common causes of cognitive impairments include brain injury, stroke, neurodegenerative disorders such as AD, and certain medical conditions such as HIV/AIDS or hypothyroidism [73]. There are various

types of cognitive impairments, including Mild Cognitive Impairment (MCI) [34], dementia, and AD [13]. MCI is a condition where a person experiences mild cognitive decline beyond that of what would normally be expected for their age. Dementia is a more severe form of cognitive decline that affects multiple cognitive domains and interferes with a person's ability to carry out daily activities [70].

Computer-Aided Diagnosis (CAD) techniques have been developed to help detect and diagnose cognitive impairments. These techniques involve the use of computer algorithms to analyze various types of data such as brain scans, medical records, and cognitive tests. For example, MRI can be used to identify structural changes in the brain that are indicative of cognitive impairment [22]. Deep learning algorithms can be trained on these images to help identify patterns and predict the likelihood of cognitive impairment [21, 31]. Other CAD techniques include cognitive screening tests such as the Montreal Cognitive Assessment (MoCA) [44, 45, 66] or the Mini-Mental State Examination (MMSE) [20, 55]. These tests are designed to assess various cognitive domains and can be administered in a clinical setting or remotely using computer-based assessments [94]. Ongoing research is being conducted on cognitive disorders such as AD, Parkinson's disease, schizophrenia, and depression to comprehend their underlying mechanisms and find effective treatments. Studies on AD are focused on detecting the disease early through biomarkers and developing potential therapies to slow or halt its progression [17, 72, 78, 88].

1.2 Alzheimer disease

One of the most prevalent cognitive disorders is AD, which is also the most common cause of dementia. In 1906, Dr. Alois Alzheimer was the first to discover AD [37, 63]. It is typified by a progressive loss of memory and other cognitive abilities [70]. With 70% of dementia cases, AD is the most prevalent type of dementia worldwide. AD is a cognitive disorder affecting cognitive function and memory, and is a leading cause of dementia in elderly individuals [48]. Over time, there is an irreversible decline in cognitive function associated with this progressive neurological disorder. The following are some of the symptoms and attributes of AD: memory loss, language difficulties, disorientation (forget where they are or how they got there, and have difficulty recognizing people they know), poor judgment, mood swings, loss of initiative, and changes in personality [7, 9–12, 16, 53, 75].

Proper classification of AD plays a crucial role in comprehending the disease, as it enables early diagnosis and prediction of patient outcomes, and facilitates informed decision-making regarding treatments. Deep learning has emerged as a promising approach for AD classification, particularly with regards to brain imaging data such as MRI or positron emission tomography (PET) scans [29, 57]. To diagnose the disease, a medical evaluation including patient history, mental state examination, physical and neurobiological tests, as well as non-invasive brain imaging techniques such as structural and functional magnetic resonance imaging are used [8, 92]. The process of diagnosing AD typically involves gathering a patient's medical history, assessing their clinical symptoms, and observing their behavior [33, 62, 77].

MRI scans, in particular, provide important information about AD through the use of deep learning and machine learning techniques. Using features taken from MRI images, machine learning algorithms are one such technique that uses to distinguish between people who are healthy and those who have AD [64, 93]. These characteristics include a range of parameters, including surface area, cortical thickness, and brain volume, and they serve as important markers of AD pathology [60]. Furthermore, deep learning algorithms provide an advanced method for interpreting brain networks seen in MRI scans [81]. These algorithms identify changes linked to AD by examining the patterns of connectivity between various brain regions [97]. Notably, research has shown that reduced connectivity between different parts of the brain is a hallmark of AD [59, 67]. Furthermore, AD detection through MRI analysis has shown notable success with computer vision techniques. The ability to identify structural anomalies indicative of AD is made possible by MRI's high-resolution imaging capabilities [3]. These developments highlight how important medical imaging is to improving our knowledge and ability to diagnose AD, especially MRI.

1.3 Introduction to proposed approach

Here, a discussion about Convolutional Neural Network (CNN), Vision Transfomer (ViT), and Compact Convolutional Transformer (CCT) models is presented. The methods employed, namely patch encodings and tokenization, will then be discussed. It is followed by an introduction and generic discussion about the suggested methodology. In Sect. 3, the complete working methodology of proposed architecture is presented in detail.

1.3.1 CNN vs ViT vs CCT

The ViT is a machine learning model for image classification that utilizes a transformer-based architecture on patches of the image. It was first introduced in a research paper titled "An Image is Worth 16×16 Words" presented at the ICLR 2021 conference by Neil Houlsby and colleagues [24]. The model is pre-trained on large image datasets such as ImageNet-21k and ImageNet [56], and employs a mechanism of attention seeking, which allows it to assign varying levels of importance to different parts of the input data. The ViT

model is composed of multiple self-attention layers, similar to those used in natural language processing, which hold great potential for use in various data modalities. Figure 1a illustrate the vision transformer.

The performance of ViT is superior to CNNs while using fewer resources. However, due to its weaker inductive bias, it needs more data augmentation or regularization while training on smaller datasets. ViT represents image inputs as a sequence of image patches and requires a significant amount of data to achieve optimal performance. The mathematical formulation of kernel convolution in CNN can be described by Eq. 1.

$$G[m,n] = (f * g)[m,n] = \sum_{j}\sum_{k} h[j,k]f[m-j,n-k] \tag{1}$$

where the f represent the input image and h denotes the kernel. The indices of columns and rows of the result matrix are marked with n and m, respectively. Unlike CNNs, which use pixel arrays, ViT splits images into visual patches during computation and employs a self-attention layer that embeds information globally in the overall image. The mathematical representation of self-attention is given in the Eq. 2. ViT can also learn to encode the relative location of the image, thereby reconstructing the image structure. Furthermore, ViT has a multi-head layer that concatenates all outputs in the appropriate dimensions, and most attention heads are used to train global and local dependencies in an image. The aspects which mainly differentiate ViT from CNN include patching and self-attention.

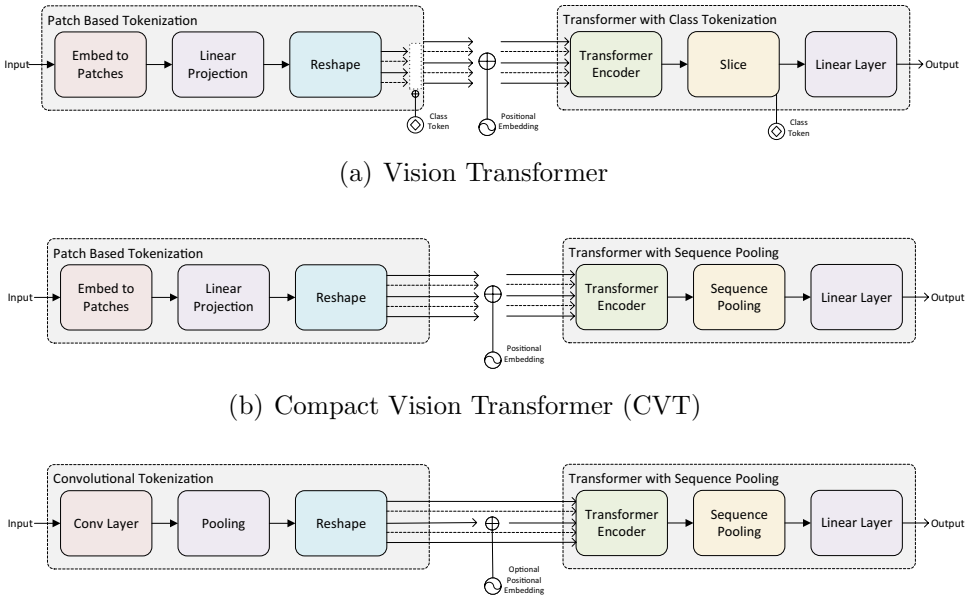
$$Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{D_h}}\right).V \tag{2}$$

On the other hand, CCT is a novel deep learning architecture that combines the strengths of both CNNs and transformers [38]. It aims to capture both local and long-range dependencies in input data efficiently. For instance, CNNs can be used to extract local features from input image followed by self-attention layers from transformers to model long-range dependencies between extracted features. Equations 3 and 4 present the mathematical formulation of the tokenization and encoding processes in CCTs, with the transformer encoder represented as f.

$$x_o = MaxPool(ReLU(Conv2d(x))) \tag{3}$$

$$x_L = f(x_o) \in R^{b\times n\times d} \tag{4}$$

The visual diagram of CCT is shown in Fig. 1c. The architecture of a CCT consists of two main components: a convolutional encoder and a transformer decoder. The convolutional encoder is responsible for extracting spatial features from the input image or video, while the transformer decoder processes the encoded features and generates the output. Self-attention layers have also been used in computer vision, but they are computationally expensive and require a large number of parameters, making them challenging to deploy on resource-constrained devices. CCTs address these limitations by using self-attention layers in a compact manner. Instead of applying self-attention to the entire input feature map, CCTs apply it to a smaller set of



(a) Vision Transformer

(b) Compact Vision Transformer (CVT)

(c) Compact Convolutional Transformer (CCT)

Fig. 1 ViT vs CVT vs CVT

features, reducing the computational cost. This is achieved by adding self-attention layers after every few convolutional layers. The self-attention layers enable the model to learn long-range dependencies between features, while the convolutional layers capture local spatial relationships.

In this work, a novel transformer-based model is developed that combines the beneficial features of CCT and ViT. The PCES is a novel method that combines the tokenization process from CCT with the patch encoding mechanism from ViT. Here, images are tokenized and patched simultaneously in different encoding modules, and their outputs are fed into transformer and self-attention layers. The two representations learned from transformer layers are combined in a single process known as MLF, which combines two different kinds of information. For classification, a multiple layer perceptron is employed. Section 3 offers a more detailed discussion of proposed methodology.

The aim of utilizing a new novel transformer for the classification of AD and cognitive disorders is to enhance the precision of diagnosis and promote the understanding of cognitive disorders. Medical imaging data presents challenges due to its complex spatial relationships and long-range dependencies [84]. Transformer applications in imaging have become a hot topic for research in recent years. By utilising their self-attention mechanism, transformers try to cope with these difficulties by producing detailed features and capturing intricate patterns by combining long-range dependencies and global context. Traditional diagnostic methods are not always reliable, and by employing sophisticated deep learning techniques like transformers, it is possible to detect patterns and characteristics in medical imaging data that may not be visible to the naked eye. This can potentially result in earlier and more accurate diagnoses, allowing for more efficient treatment and care of patients. In addition, as the world's population continues to age, there is a growing need for the development of more effective tools for diagnosing AD and cognitive disorders.

1.4 Contribution

This study involves the development of a deep transformer architecture for the classification of different stages or types of AD and other cognitive disorders in 2D MRI images data. The main contributions of this paper are listed as follows:

- A novel computer-aided diagnosis system is suggested for AD and cognitive impairments which can be used by medical professionals for decision making followed by quick and efficient treatment.
- A deep learning-based model called BiViT has been introduced to detect AD and cognitive disorders in 2D MRI imaging data. This system makes use of (PCES) and MLF, resulting in a significant improvement in the accuracy of the results.
- The study propose a new PCES technique that involves two types of encoding to process data leading to improved encoding that further enhances model's performance in terms of achieving accurate results.

The structure of this paper is as follows: Sect. 2 provides an overview of the relevant literature on the subject at hand. The methodologies employed in the current study are described in Sect. 3. Furthermore, Sects. 4 and 5 present the results and discussion, and Sect. 6 concludes the paper.

2 Literature review

This section presents a literature review on cognitive disorders, primarily related to AD, with a focus on the essential role that deep learning plays in the recognition and classification of cognitive disorders from imaging data. AD is a severe neurological condition that leads to progressive damage to brain cells, causing permanent memory loss and dementia [51]. Early detection of AD can help control its spread, and hence there is a need for an autonomous system that can classify medical condition into different stages. In recent times, machine learning and deep learning techniques have been successfully applied to many medical problems, including AD detection [30]. Deep learning has been used in many studies to classify cognitive disorders mainly AD, using imaging data such as MRI or PET scans, and clinical data [15]. Some studies have found that deep learning models can achieve high accuracy in classifying AD, particularly when using imaging data. In this regard, CNN, ViT and autoencoders have been used along with other deep learning architectures to determine the essential features of these MRI scans and categorizing them into healthy or disease groups. In this section, the different methods used for AD detection in relation to their pros and cons are analyzed.

Transfer learning is an important aspect when the training data is very low. Ghazal and Issa [36] aims to detect AD using brain MRI to classify images into four stages using transfer learning including healthy, mild demented, moderately demented, and severe demented. The work utilizes a transfer learning-based AlexNet model for characterizing the disease at an early stage with high accuracy. The proposed system's simulation results have demonstrated that it can achieve an accuracy of 91.70%, making it an effective tool for early detection of AD. Merits of the transfer learning-based model include fast training of the model, re-usability and reduced data requirement. However, fine-tuning of AlexNet trained on ImageNet dataset [23] on medical images dataset can be questionable. Moreover, there is a chance of overfitting with the new data, especially, if new

dataset is significantly different from training data. To sum up, the problem of domain discrepancy exists in this research and it can be considered as the main drawback of this study. In another study, transfer learning-based ResNet50 is used to achieve AD detection and determine its stage by applying brain images [99]. It means that, the approach is developed using hybrid Resnet50 with other CNN architectures like Alexnet, Densenet201, and Vgg16 [52, 91, 98]. The study demonstrated that the proposed hybrid model had an accuracy of 90%, which outperformed the individual CNN architectures. Hence, the hybrid model showed promising results in diagnosing AD and showed better performance than other CNN architectures reported in the literature. However, it has certain drawbacks such as small receptive fields, a lack of long-term dependencies, and a lack of attention mechanisms. As the proposed transfer learning-based CNN model is almost similar to [36] research, it has similar limitations ranging from domain mismatch to overfitting. The second main problem that comes with the use of CNN based models is that they do not use the attention mechanism unlike ViT. Attention-mechanism is one of the key components in image recognition which is used to locate regions of the image that are of importance. Some other shortcomings associated with CNN-based models are limited contextual knowledge, fixed-size input, lack of global attention and large number of parameters.

Mild Cognitive Impairment (MCI) is an intermediate condition between healthy individuals and AD. Taheri Gorji and Kaabouch [87] conducted a study on the significance of early detection of MCI using MRI. The study employed a CNN to classify MRIs of 600 individuals into healthy, Early MCI (EMCI), or Late MCI (LMCI) classes. The CNN, with an efficient architecture, discriminated between the healthy group and the two types of MCI groups, achieving an overall classification accuracy of 94.54%. The advantages include improved accuracy in MCI classification and the potential for early intervention, however, the model's performance may vary depending on the dataset used and may require further validation.

Ensemble learning is a machine learning technique where multiple models' predictions are combined to boost the overall performance. Unlike conventional models, it utilizes the heterogeneity among stand-alone models in order to achieve lower error and enhanced stability. One substantial advantage is its capability to accomplish better outcome of prediction by combining the powers of different models. Ensembling involves using several different simple models separately and then joining them together to obtain fitter and more generalizing results, especially if individual models are dissimilar. This approach also enhances the model's ability to handle noisy data and outliers. Kang et al. [47] presents a CNN-based ensemble learning approach for AD classification from the MRI data. They implemented the use of GAN's Discriminator, VGG19 and ResNet50 ensemble models, and majority voting is used to fuse the outcomes. The proposed model was able to achieve an excellent performance with an accuracy of 92%. The ensemble learning-based networks are usually more robust and have better generalization capabilities as compared to stand-alone models. However, ensemble learning is computationally expensive because of the need to train and handle multiple models. Moreover, if base models are not well-trained or they look the same, ensemble learning will not give much improvement to a single model and it will lead to overfitting. Furthermore, ensemble learning-based models are prone to overfitting if the base learners are too complex or if the ensemble has too many components. To summarize, this approach can achieve better results but it has issues including limited knowledge about context, lack of global attention, domain mismatch and large number of parameters, that would hinder the practical utilization of this model.

There are currently no biomarkers known to be extremely accurate in diagnosing AD in its early phases, making it a difficult task in medical practise to identify AD in its early stages. Moreover, AD is an incurable disease, and high failure rate was observed in clinical trials for AD treatments. To help slow down the progression of AD, researchers are striving to find ways for early detection. With a focus on neuroimaging and mostly academic articles released since 2016, [65] review the most recent state-of-the-art research on machine learning approaches used for the detection and classification of AD in this study. Various machine learning techniques, including Support Vector Machine (SVM), Random forest, CNN, K-means, and others, have been employed for the detection and classification of AD. The review indicates that there is no single best approach, but deep learning techniques, such as CNNs, appear to be promising for the diagnosis of AD. A similar research by [6] shows that among K-Nearest Neighbor (KNN), SVM, Decision Tree (DT), Linear Discrimination Analysis (LDA), Random Forest (RF) and CNN algorithms, the CNN performs the best for classifying AD using imaging data and some extra data from MRI such as the average cortical thickness, the standard deviation of cortical thickness, the volume of cortical parceling, white matter, and surface area. However, it has a limitation of lack of global attention and contextual understanding.

In contrast to supervised learning whose models are trained by using labeled data, unsupervised learning is essential because it can discover structures and patterns that are hidden in unlabeled data. One of the most widely used approaches of unsupervised learning for image data is the autoencoder. Encoder and decoder are the main components of autoencoders where encoded representation of inputs in a low-dimensional space occur with the help of the former and the latter reconstruct original images from this representation. Leveraging the benefit of autoencoders, [96] employed

a Stacked AutoEncoder (SAE) to extract features from MRI data, and a SVM was finally used to classify AD using those features. They found that the deep autoencoder was able to extract useful features that improved the accuracy of the SVM, resulting in an accuracy of 89%. SAEs offer several advantages, such as hierarchical representation learning and the ability to model non-linear transformations. However, they also come with disadvantages, including the potential for overfitting and the computational complexity of training deep models. SAEs also pose challenges in interpretation due to the complexity of their learned representations, making it hard to interpret model decisions. Additionally, they are data-hungry, needing substantial data for training meaningful representations, which can limit their effectiveness with small or unrepresentative datasets.

Recently, researchers have explored innovative approaches to improve the accuracy and efficiency of AD diagnosis using advanced technologies such as deep learning and neuroimaging. One such approach, proposed by [74] from Imperial College London, focuses on utilizing imaging data to differentiate AD from MCI and Normal Control (NC). Their method leverages an autoencoder and a 3D CNN architecture, achieving an impressive accuracy of 95.39% in distinguishing AD from NC individuals. Additionally, a 2D CNN design is developed which yields comparable accuracy results. Likewise, [61] designed a diagnostic approach for AD using multi-modal neuroimaging data. This method utilizes a novel zero-masking technique which preserves all the information contained within the data. SAE is used for extracting high-level features and subsequently feedint them into SVM for the purpose of multi-modal and multi-class MR/PET data classification. The study revealed a performance of 86.86% accuracy by the model, thereby presenting a possibility for the employed method in early detection of AD. Such cutting-edge advertising techniques reinforce the fact that the increased use of advanced technology is very significant to the progress in the diagnosis and treatment of AD. Unsupervised autoencoder approach is useful for nonlinear features learning, but one has to face issues like model generability and interpretability, as well as the model overfitting. The SAEs can be complex and thus hard to interpret, and this complexity may hamper their adoption in clinics where interpretability is very important.

Over the past decades, different deep-learning approaches in medical imaging have shown promising results and performance. Drewitt [25] explores ViT approach for classifying AD in MRI images. It is also compared with other deep learning-based networks and the article further points out limitations to present future prospects of this approach. The performance metrics include accuracy and F1-Score, with the model attaining an accuracy rate of 87.5% and a loss of 0.34 in AD classification. This indicates that the proposed model could help physicians to diagnose AD and give a remedial treatment to the patients accordingly that can ultimately decrease the mortality rate associated with the disease. In another study [43], a ViT is trained using natural images to maximize the large-scale data available in computer vision. The pre-trained ViT model is then mobilized to the brain imaging site where few public but relatively excellent samples are available to achieve an accuracy of 96.8%. This indicates the model's significant scalability performance which can be an improvement upon the traditional neural networks.

The growing importance of early AD diagnosis parallels the aging global population. A study by [100] introduced a novel approach using the SMIL-DeiT model for AD classification. It also used three categories of the disease including AD, MCI, and NC. The proposed model is inspired by ViT, preceded by data pre-training through DINO, a self-supervised task. The developed architecture is applied to the ADNI dataset and measured by several metrics such as precision, recall, accuracy, and F1-score. The proposed method recognized text with an accuracy of 93.2%, exceeding what was done by the transformer-based (90.1%) and CNN-based (90.8%) models. Self-supervised pre-training methods such as DINO typically necessitate substantial data volumes to achieve meaningful representation learning. This can be particularly challenging in medical imaging, where datasets are frequently constrained in size.

Several studies have proposed innovative approaches for the early diagnosis of AD using deep learning techniques applied to medical imaging data, particularly MRI scans. Sethi et al. [79] introduced a CNN-SVM model that combines the feature extraction capabilities of CNN with the classification abilities of SVM. This model achieved relative improvements in accuracy ranging from 0.85 to 3.4% on different datasets, with an impressive accuracy of 86.2% on the OASIS dataset. The model has shown its potential to be very accurate in terms of diagnosis of this specific condition and it also works very well with the complicated datasets. These advantages can be crucial for the AD diagnosis at the early stages and consideration of further researches in the particular field. On the other hand, the efficiency of the model may be affected by the longer training time and the dependence on big datasets, which could, theoretically, limit the practical applicability in certain contexts in real life. Similarly, [80] developed a CNN classifier named AlzheimerNet, which can identify all stages of AD and the NC class through MRI scans. This model achieved a remarkable test accuracy of 98.67% and outperformed five other pre-trained models. An ablation study demonstrated the model's superior performance and its ability to outperform traditional methods for classifying AD stages from MRI scans. The advantages of AlzheimerNet include its high accuracy and robustness to noise, but it may be computationally expensive and require

a large amount of data for training. Due to the fact that these algorithms are mainly based on CNNs, some challenges may occur such as capturing global context and long-range dependencies of data.

In conclusion, the literature highlights the effectiveness of DL techniques in classifying AD stages from 2D MRI images, with ongoing research focusing on improving accuracy and exploring new approaches such as ViT and transfer learning models. With no doubt, early detection of AD and MCI is essential for timely intervention and improved patient outcomes. However, the majority of methods employ CNN-based models that come with certain limitations such as the inability to capture long-term dependencies and the absence of an attention mechanism. Therefore, it is paramount to develop a more robust system that can cope with these issues.

3 Proposed Bi-Vision transformer (BiViT)

This section elaborates on the research methodology and the constituent components of the proposed BiViT algorithm. In this research, a novel BiViT architecture is developed incorporating parallel coupled encoding strategy (PCES) and mutual latent fusion (MLF). First, the proposed methodology is discussed below followed by each of the novel aspects including PCES and MLF. It is worht mentioning that the present study focuses on the classification of various categories of AD stages and cognitive disorders stages. Moreover, the methodology is comprised of five steps: data augmentation, preprocessing, patch encoding, CC tokenization, self-attention mechanism and MLF. The methodology is illustrated visually in the Fig. 2.

The BiViT model incorporates PCES and MLF, which enables it to capture local and global contextual information. Overall, the proposed methodology consisting of preprocessing, data augmentation, and the BiViT model provides a robust framework for computer-aided diagnosis of AD and other cognitive disorders. The model's performance is evaluated in terms of training, testing, and validation accuracy, demonstrating its efficacy in AD diagnosis and cognitive disorders classification. We'll get into the mathematical description of the Bi-Vision transformer in the following paragraphs. We start with the input images X_n and the associated labels, Y. We obtain enhanced versions of the data denoted as X_n' by using data augmentation techniques. Since we have limited computational resources, the method starts with normalizing and reducing the data. As a result, the images are downsized to 128×128. Following the augmentation phase, the minimum and maximum values in each instance are represented by the variables $X_{min}^{'}$ and $X_{max}^{'}$. After the data augmentation, resizing, and

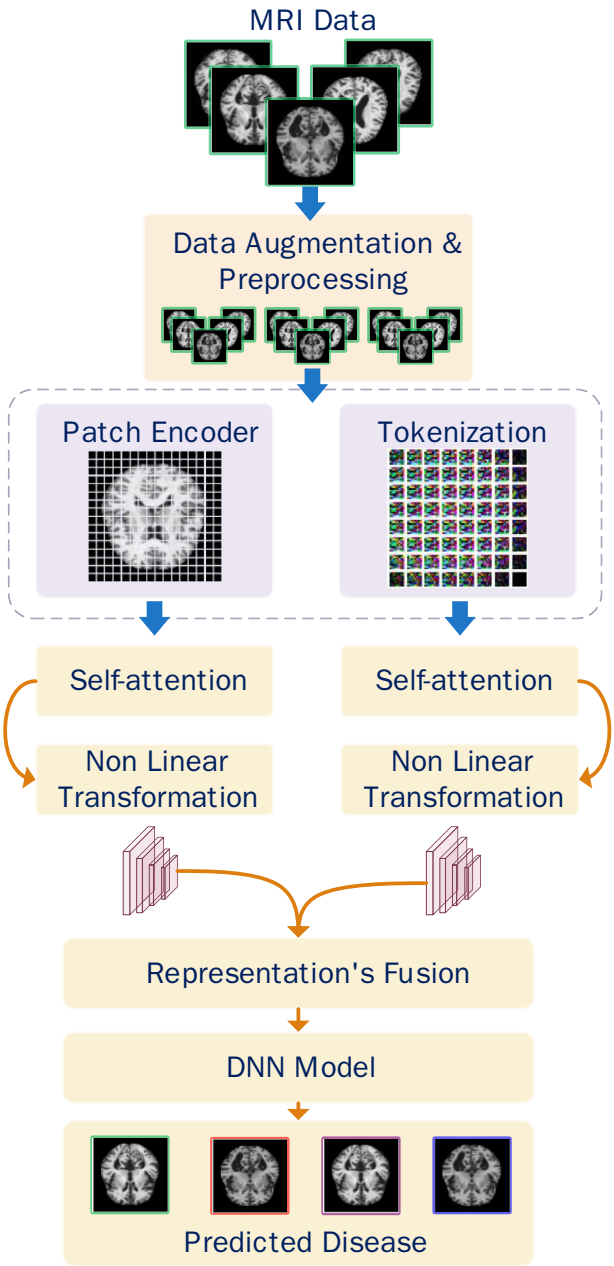


Fig. 2 Proposed methodology

scaling procedures are finished, the data instance that results is called X_n'', as Eq. 5 illustrates.

$$X_n'' = \frac{X_n' - X_{min}'}{X_{max}' - X_{min}'} \tag{5}$$

Once the initial preprocessing is completed, the subsequent steps involve parallel stages: patching and tokenization. As part of the patching process, the entire image is divided into 256 smaller patches, each measuring $8 \times 8 \times 3$. which are then projected into a lower-dimensional space.

Additionally, we incorporate positional embeddings into the projected patch embeddings. The mathematical formulation for patching on the data batch can be found in Eq. 6. Here, $x_n^{''N}$, where N is 1,2,3.. up to number of patches, represents the patches of a single instance in Eq. 6, and E is for the learnable embeddings.

$$Z_p = [X_{class};\mathrm{x}_n^{''1}E;\mathrm{x}_n^{''2}E....\mathrm{x}_n^{''N}E] + E_{pos} \tag{6}$$

In contrast to patching, tokenization approach involves utilizing convolutional and pooling transformations for operations. Convolutional layers are used to process the data, extracting spatial information through convolutions and downsampling operations. These local features are then presented as tokens and passed into the transformer for additional processing. After applying convolutional operations, tokenization process produces patches with a $16 \times 16 \times 3$ size, for a total of 64 patches. Tokenization process converts the images data into smaller tokens, as depicted in Eqs. 7 and 8. Tokenization involves first applying pooling and convolutional layers globally to the entire image, and then turning the resulting image into tokens.

$$F_n = \mathrm{MaxPool}(\mathrm{ReLU}(\mathrm{Conv2d}(X_n^{''}))) \tag{7}$$

$$Z_t = [x_{class};\mathrm{F}_n^{''1}E;\mathrm{F}_n^{''2}E....\mathrm{F}_n^{''N}E] + E_{pos} \tag{8}$$

The reason for using $8 \times 8 \times 3$ patches in patching module and $16 \times 16 \times 3$ in tokenization module is to effectively capture and represent the spatial and channel information present in the images. Different sizes are used to achieve both detailed and global context features as bigger token sizes aid in capturing a wider context, while smaller patch sizes enable the capture of more specific information.

Next, the self-attention is applied to Z_p and Z_t to capture dependencies between the patches and tokens on a local and global level. The Z_p attention block take (q, k, v) as (patches, patches, tokens), while Z_t takes (q, k, v) as (tokens, tokens, patches). This enhances the information provided to both attention mechanisms which is better for local and global subtle features learning. With the help of self-attention, the model is able to comprehend the spatial dependencies and contextual information present in the image by learning the relationships between patches and tokens. The mathematical formulations for the self-attention mechanism applied to Z_p and Z_t can be found in Eqs. 9 and 10.

$$Attention(Z_p, Z_p, Z_t) = Softmax\left(\frac{Z_p Z_p^T}{\sqrt{D_h}}\right) \cdot Z_t \tag{9}$$

The matrix product $Z_p Z_p^T$ is replaced with the covariance matrix, represented by ϕ. To normalize the dot product attention scores in self-attention, $\sqrt{D_h}$ is used as a scaling factor. This helps to sustain gradient stability during training.

$$Attention(Z_t, Z_t, Z_p) = Softmax\left(\frac{Z_t Z_t^T}{\sqrt{D_h}}\right) \cdot Z_p \tag{10}$$

The matrix product $Z_t Z_t^T$ is replaced with the covariance matrix, represented by Ψ.

Stochastic depth is another idea derived from CCT. Stochastic depth refers to an approach for randomly skipping or dropping network layers during training for the CCT. It is a regularisation technique designed to enhance the functionality and generalizability of deep neural networks. The use of stochastic depth to Eq. 10 can be observed in the Eq. 11. The drop probability ϑ is the main argument for stochastic depth, while the keep probability is represented as $1 - \xi$. The mathematical forms of probability (called keep probability) is described in Eq. 11.

$$\xi = 1 - \vartheta \tag{11}$$

The Eq. 12 represents the vector obtained from a random distribution, denoted as $\mathcal{U}_\delta$.

$$\mathcal{U}_\delta = \left(\xi + \mathcal{U}_\Theta\right) \in \mathcal{Z} \tag{12}$$

The vector $\mathcal{U}_\Theta$ in Eq. 12 represents a uniform vector drawn from a simple random distribution between 0 and 1. After adding ξ, we apply the floor function to the resulting vector values to convert them into integers within the domain $\mathcal{Z}$. The resulting output, obtained after applying stochastic depth, is illustrated in Eq. 13.

$$\mathcal{D}_{sd} = \frac{\mathcal{U}_\delta}{\xi}.Softmax\left(\frac{\Psi}{\sqrt{D_h}}\right) \cdot Z_p. \tag{13}$$

The dense transformation and concatenation operations, represented by λ_θ, are applied on the output of self-attention (in case of patching), as described in Eq. 14.

$$\lambda_\theta = \lambda_\theta\left(Softmax(\frac{\Phi}{\sqrt{D_h}}) \cdot Z_t\right) \tag{14}$$

For tokenization, the same transformations are applied to the output of self-attention, as depicted in Eq. 15.

$$\lambda_\Omega = \lambda_\Omega\left(\frac{\mathcal{U}_\delta}{\xi}.Softmax(\frac{\Psi}{\sqrt{D_h}}) \cdot Z_p.\right) \tag{15}$$

Finally, attention weights are computed, and the representations λ and Ω are multiplied by their respective weight matrices (as depicted in Eqs. 16 and 17).

$$\lambda_w = \sigma(\lambda_\theta.\omega_1)_i^T.\lambda_\theta \tag{16}$$

In Eq. 16, the softmax function σ is applied to the product of λ_θ and the trainable weight matrix ω_1, denoted as $(\sigma(\lambda_\theta.\omega_1)_i^T)$. The expression $\sigma(\lambda_\theta.\omega_1)_i^T$ represents the softmax function applied to the output (σ) multiplied by the weight matrix ω. Similarly, in the tokenization phase, the same process is performed as described in Eq. 17.

$$\Omega_w = \sigma(\Omega_\theta.\omega_2)_i^T.\Omega_\theta \tag{17}$$

Finally, we incorporate a fusion function, denoted as $\mathcal{F}_\partial$, to combine the weighted representations obtained. This fusion function combines the information from both representations. In our specific case, we use concatenation as the fusion method, which is illustrated in Eq. 18.

$$\mathcal{F}_\partial = \mathcal{F}_\partial\left(\lambda_w, \Omega_w\right) \tag{18}$$

The fusion function, denoted as $\mathcal{F}$fusion, takes two arguments, λw and Ω_w, as depicted in Eq. 19.

$$\mathcal{F}_\partial = \mathcal{F}_\partial\left(\sigma(\lambda_\theta.\omega_1)_i^T.\lambda_\theta,\ \sigma(\Omega_\theta.\omega_2)_i^T.\Omega_\theta\right) \tag{19}$$

The Eq. 19 represents the final representations that are input to the classifier in our case, which is a Softmax classifier due to the multi-class classification task. The process starting from self attention and continuing until Eq. 19 is repeated n times for efficient representation learning. This entire process is often referred to as the transformer encoding, which encodes patches/tokens into meaningful representations. The PCES and MLF processes are covered in the following sections, but first it's critical to comprehend the tokenization and patching processes. A detailed discussion about how tokenization and patching operate on images is given in Sect. 3.1.

3.1 Patch encoding and tokenization in transformers

Patch encoding and tokenization are both important techniques used in computer vision tasks because they allow input images to be divided into smaller, more manageable parts that can be more efficiently processed by neural networks. This is particularly crucial when dealing with large, high-resolution images, which can be computationally expensive to process using traditional methods. The patch encoding and tokenization processes are illustrated in Figs. 3 and 4, and are explained below.

3.1.1 Patch encoding

To use a transformer model to process image data, patch encoding is performed where the image is divided into fixed-size patches that are flattened into vectors [24]. These patch vectors are then sent to the transformer model which utilizes the self-attention mechanism to extract visual features and classify the image. Patch encoding is more flexible and efficient for processing high-resolution images as can be seen in Fig. 3.

3.1.2 Image tokenization

In CCT, the process of converting image patches into learnable representations called 'image tokens' using a trainable CNN is known as image tokenization [38]. These image tokens are subsequently fed into a transformer encoder to perform tasks like image classification or object detection. Unlike traditional CNN, this approach eliminates the need for fully connected layers and pooling layers, which enhances the efficiency and flexibility of

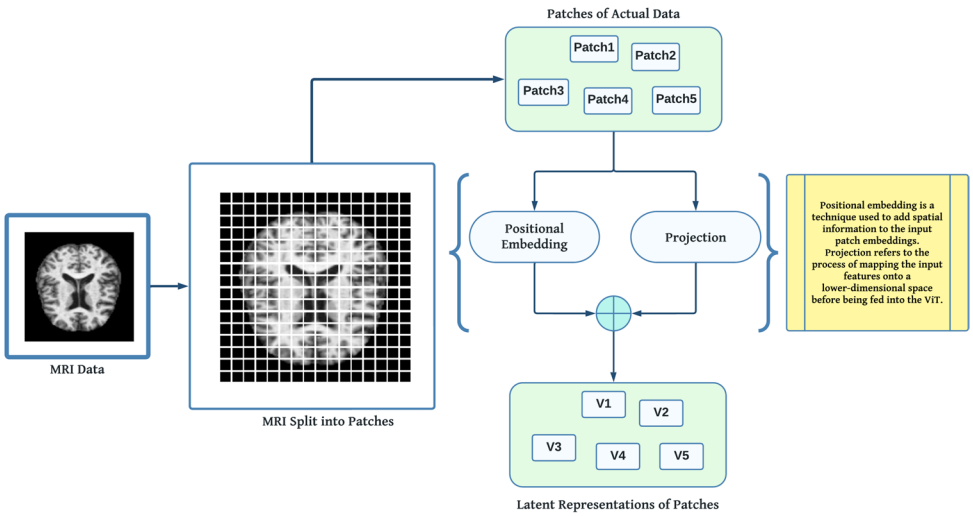


Fig. 3 Process of patch encoding in ViT

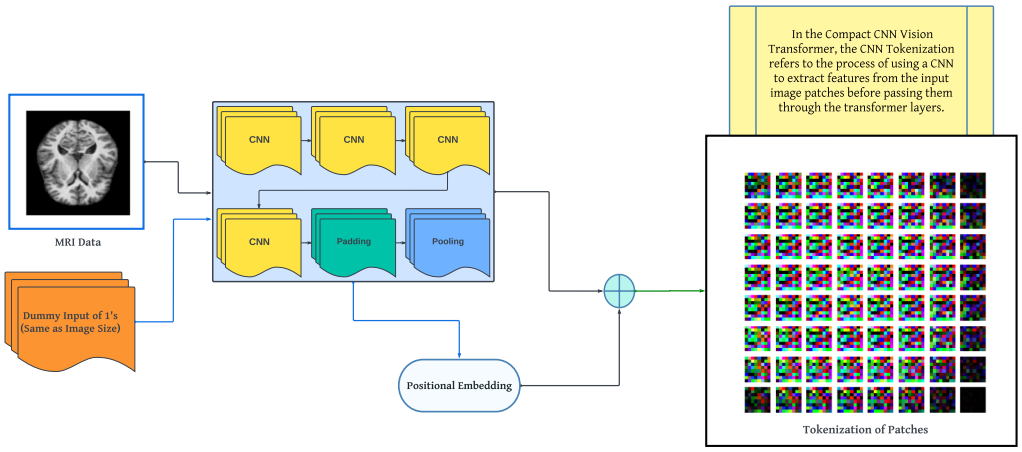


Fig. 4 Image tokenization process in CCT

image processing. The Fig. 4 shows the steps of tokenization visually. The visual representation of the tokenization process can be observed in Fig. 4.

3.2 Parallel coupled encoding strategy (PCES)

In transformer encoding phase, PCES is implemented to process images using both patch encoding, as shown in Fig. 3, and CC tokenizer, illsutrated in Fig. 4. In the context of a CCT, CC tokenizer refers to the process of breaking down an image or video into smaller, manageable pieces, or "tokens", that can be processed by the CCT's convolutional encoder. The CC tokenizer is responsible for dividing the input image or video into a fixed-size grid of non-overlapping patches, each of which is then represented as a set of pixels. These patches are then converted into tokens, which are then passed through the convolutional encoder to extract spatial features. These tokens are then passed to the transformer decoder to generate the output. The CC tokenizer is responsible for the pre-processing step that allows CCT to process images and videos more efficiently, as it reduces the dimensionality of the input data by breaking it down into smaller, more manageable pieces.

Patch encoding refers to the process of breaking down an image into smaller, manageable pieces, or "patches", that can be processed by the ViT's transformer-based architecture. The patch encoding is responsible for dividing the input image into a fixed-size grid of non-overlapping patches, each of which is then represented as a set of pixels. These patches are then flattened and passed through a linear layer to obtain a feature vector. These feature vectors are then used as the input to the transformer-based architecture, where the self-attention mechanism is applied to learn the relationships between the patches and generate the output. The patching process in ViT is mathematically described by Eqs. 6, 9, and 14.

The patch encoding step allows the ViT to process images more efficiently by reducing the dimensionality of the input data by breaking it down into smaller, more manageable pieces, and it also allows the model to learn the relationships between the patches, which is useful for image classification and other tasks. Here, the two concepts are combined together and interconnect in a transformer encoder. In simple transformers, a single encoding strategy is used, however, in proposed model, two strategies are applied and hence known as PCES. It means that two mechanisms run in parallel while information is

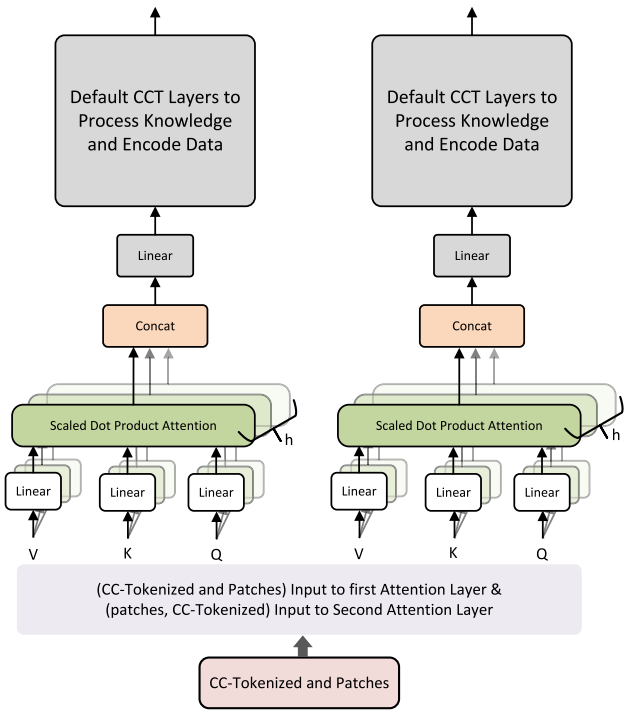


Fig. 5 PCES in transformer

exchanged continuously between two encoding as shown in Fig. 5.

3.3 Mutual latent fusion (MLF)

Following the patch encoding process in ViT, the self-attention mechanism is applied by the model to capture the connections between the patches and produce the ultimate output. The self-attention mechanism in ViT represents each patch with a key, query, and value vector and calculates the dot product between the query and key vectors for all patches [49]. The dot product values are then used to compute a weight for each patch, which determines the importance of the patch for the final output. These weights are used to compute a weighted representation of the patches by taking a linear combination of the value vectors.

After applying the self-attention mechanism in ViT, the weighted representation is produced and it includes the crucial information from the patches, which is further processed by a feedforward neural network to generate the final prediction; thus, the weighted representation in ViT is the result of the encoding and attention mechanism, consisting of the combined patch information weighted by the attention mechanism.

To combine the latent representations generated from both the CC tokenizer and patch encoding process, a fusion mechanism is used where the representations are merged to create a final output for the model. Equation 19 provide the mathematical form for the combined MLF process.

In neural network, feature fusion refers to the process of combining information from multiple layers or multiple channels of the network to form a more comprehensive and robust feature representation [85]. To combine the weighted representation of encoded information, the MLF concept is utilized followed by passing the features to the classifier to predict the instance's actual label, as depicted in the Fig. 6.

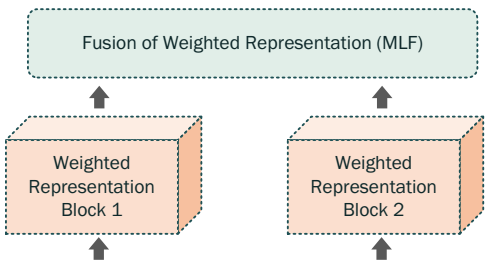


Fig. 6 MLF in transformer

4 Experiments and results

This section presents a comparison between the BiViT algorithm and various transfer learning algorithms including DenseNet121, ResNet50, VGG16, VGG19, Inception-ResNet V2, Inception V3, EfficientNet B0, EfficientNet B1, ResNet-101, Xception, MobileNet, and ResNet-152. Additionally, unsupervised deep learning techniques such as convolutional autoencoders, variational autoencoders, and sparse autoencoders were utilized for classifying abnormalities in instances with a primary focus on AD. The subsequent sections discuss the performance of different deep learning techniques in comparison to the proposed BiViT algorithm.

This section follows a structured outline beginning with Sects. 4.1.1, and 4.1.2, which describe the datasets used in the research. In Sect. 4.2, the preprocessing techniques and augmentation techniques utilized in this study for the classification task are outlined. Section 4.3 and 4.5.1 detail the use of various transfer learning algorithms, both with and without augmentation, for classifying AD, and then compare the performance of all algorithms. The results of this analysis are presented in Tables 3 and 4. Table 5 summarizes the class-wise performance of the top four models out of all trained models. Additionally, Sect. 4.5.2 describes the performance of different unsupervised deep learning techniques, such as autoencoders, to extract features from images and classify instances based on abnormalities. The results of unsupervised learning-based classification of AD stages are presented in Table 6. Section 4.6 focuses on the same transfer learning algorithms and deep autoencoders for the classification of cognitive disorders other than AD. The outcomes of this analysis are presented in Tables 7, 8, 9, and 10.

The proposed approach is applied to classify cognitive disorders and AD. The model is trained for up to 200 epochs using MRI images, with a total of 2,569,014

Table 1 Hyper parameter setting and information about the proposed BiViT

Coupled Bi-Vision Transformer (BiViT)		
Sr. No.	Detail	Quantitative values
	Parameters	2,569,014
2	Size of input images	$128 \times 128 \times 3$
3	Epoch	200
4	No of classes	4 and 5
5	Batch size	32
6	Learning rate	0.001
7	Optimizer	Adam
8	Verbose	False

trainable parameters. Table 1 provides a summary of the hyperparameters (learning rate, batch size, number of epochs etc.) used in this research work. Various metrics, including F1 score, accuracy, recall, precision, and AUC, are employed for evaluating the performance of the BiViT model.

4.1 Dataset

The research conducted uses two publicly available MRI datasets. The first dataset is utilized for AD stage classification and consists of four classes: mild demented, very mild demented, non-demented, and moderate demented. The second dataset is used for detecting various cognitive impairments and consists of five types of disorders: AD, CN, EMCI, LMCI, and MCI [35, 83]. The first dataset with four classes is used to classify the stage of AD, which is a progressive brain disorder affecting memory, thinking, and behavior. The classification process can help with early diagnosis and treatment of the disease. The four classes in the dataset represent different stages of the disease, with mild demented and very mild demented indicating early stages and non-demented and moderate demented indicating later stages. The stages of cognitive disorder range from mild to severe, with the severity of cognitive decline increasing as the disease progresses. This stage of cognitive decline is often referred to as moderate to severe dementia [27]. Figure 7 illustrates the hierarchical structure of both datasets.

The second dataset is used to detect various cognitive disorders, including AD, and is divided into five types of disorders. The CN class represents individuals without any cognitive impairment, while the other four classes represent different stages of cognitive decline. On the other hand, MCI is a condition in which an individual experiences mild cognitive decline beyond what is expected for their age but is still able to perform their daily activities. EMCI refers to the early stages of cognitive decline [42], and LMCI refers to a more advanced stage of cognitive decline [58]. Finally, AD is a progressive neurodegenerative disorder that affects memory, and behavior, and is the most common cause of dementia in older adults. Early detection of cognitive disorders can help with timely

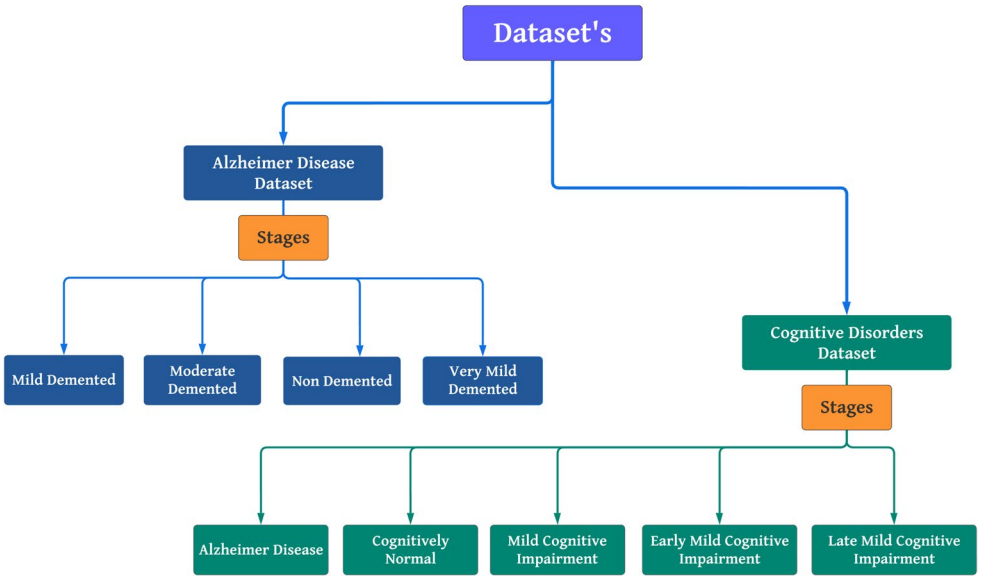


Fig. 7 Visual hierarchy of data

Table 2 Alzheimer disease and cognitive disorders data distribution

Sample distribution of cognitive disorders dataset					
Samples	CN	EMCI	LMCI	MCI	AZ
Training	401	172	47	166	114
Testing	179	68	25	67	57
Sample distribution of Alzheimer disease dataset					
Samples	Normal	Moderate	Mild	Very Mild	
Training	2797	58	781	1964	
Testing	403	6	115	276	

interventions to slow down the progression of the disease and improve the quality of life of affected individuals.

The Table 2 presents the details of training and testing samples utilized in the conducted experiments of this study. In the upcoming section, we provide a brief overview of each dataset.

4.1.1 Alzheimer disease data

The dataset used in this research was collected from several websites, hospitals, and public repositories [26]. The dataset consists of preprocessed MRI images that were resized into 128×128 pixels. The dataset is comprised of four different classes of images, which are classified based on the severity of dementia. In total, the dataset contains 6400 MRI images, after the augmentation process [95]. The Fig. 8 depicts the MRI samples from all the classes together.

The first class in the dataset is mild demented, which contains 896 images. This class represents patients who have mild symptoms of dementia. The second class is moderate demented, which contains 64 images. This class represents

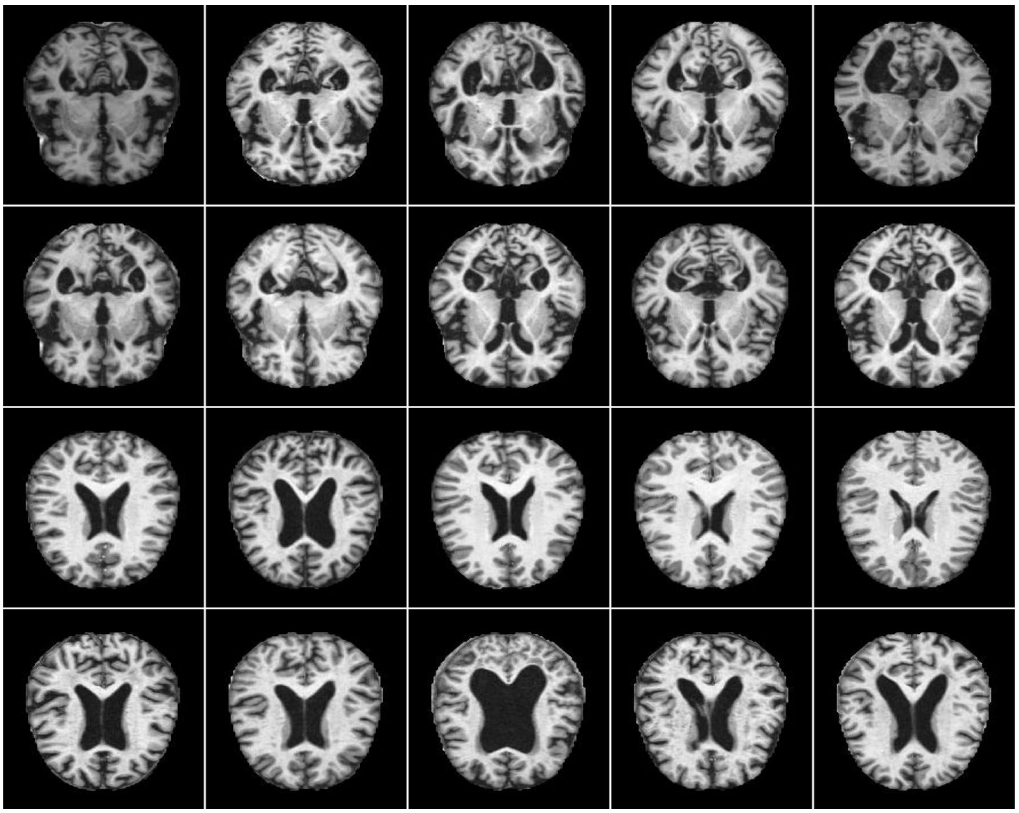


Fig. 8 Alzheimer MRI Data Samples

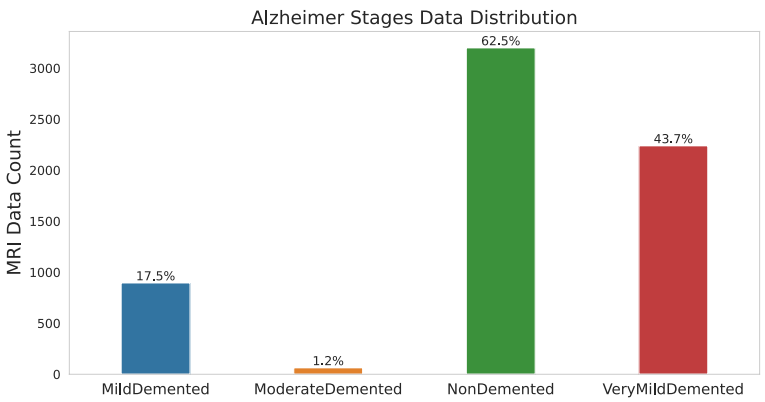


Fig. 9 Distribution of Alzheimer disease dataset

patients who have moderate symptoms of dementia. The third class is non demented, which contains 3200 images. This class represents patients who do not have dementia. The fourth and final class is very mild demented, which contains 2240 images. This class represents patients who have very mild symptoms of dementia. The distribution of MRI data for AD can be seen in Fig. 9.

4.1.2 Cognitive disorders data

The dataset used in this study comprises five different stages of cognitive disorders, which have been divided into two directories for the purposes of training and testing. The stages included in the dataset are EMCI, LMCI, MCI, AD, and CN. Specifically, there are 204 samples of EMCI, 61

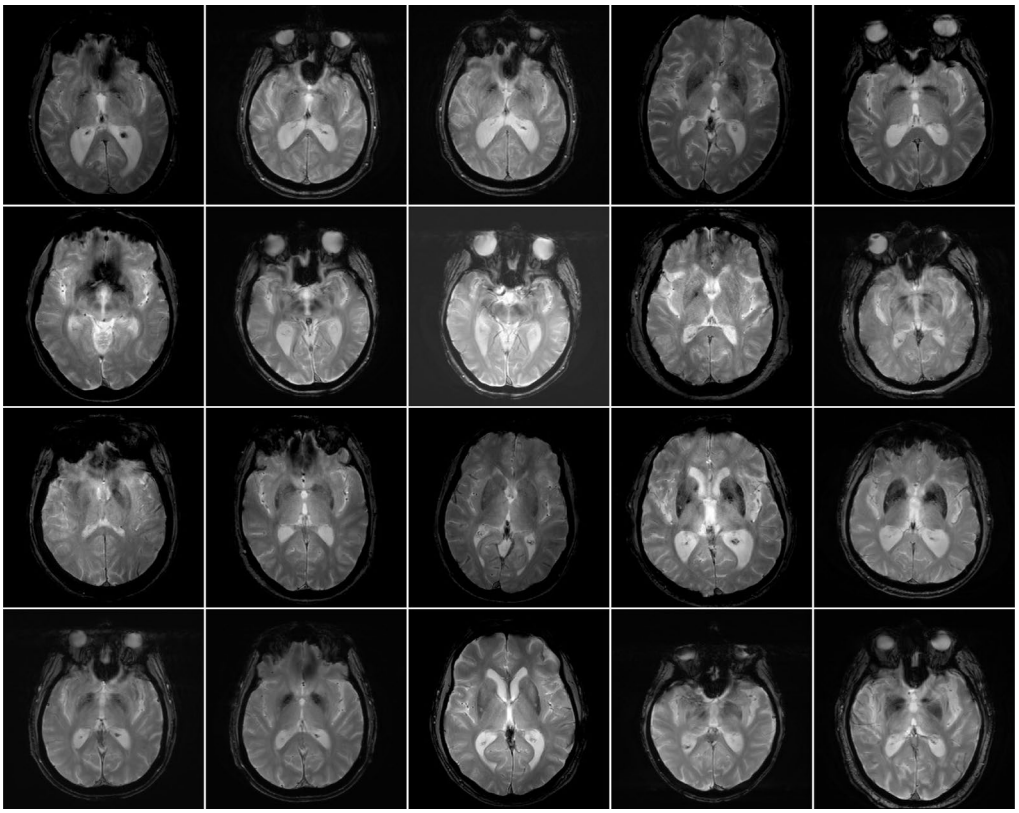


Fig. 10 Cognitive disorders MRI samples

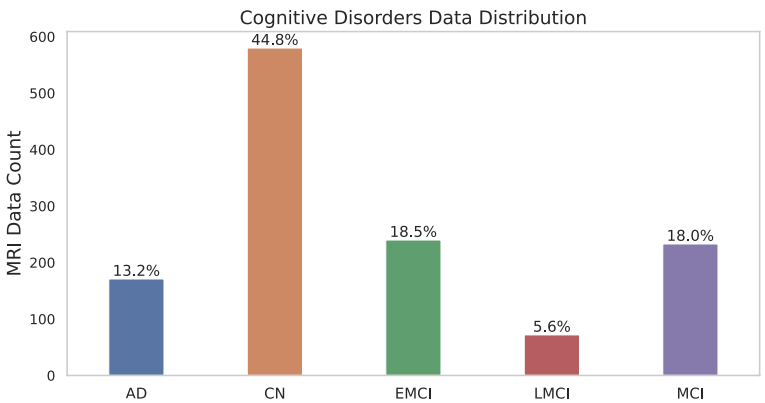


Fig. 11 Distribution of cognitive disorders dataset

samples of LMCI, 198 samples of MCI, 145 samples of AD, and 493 samples of CN, for a total of 1101 samples [18]. The dataset was sourced from the ADNI website and was created as a collaborative effort to accelerate Alzheimers research. The ADNI website is duly credited as the source of the dataset [2]. The MRI data samples for cognitive disorders are illustrated in Fig. 10.

Figure 11 illustrats the distribution of data across different stages of cognitive decline.

4.2 Data preprocessing

To make the images suitable for the model, the preprocessing step aims to enhance their quality thereby improving the performance of the model. The primary objective of preprocessing is to enhance important image features to improve the image data for further processing. One of the most important steps in improving the quality of images is preprocessing, which highlights important details. The scaling and resizing operation is performed to standardize the size and resolution of images. In contrast, scaling modifies pixel intensity values to enhance contrast and detail visibility, as a result, the pixels have values between 0 and 1. The images were resized to 128×128 dimensions, and subsequently, normalized using Eq. 20 such that the pixel values ranged between 0 and 1. Let I represent the input image having a size $(m \times n)$ and I_{norm} represent the normalized MRI, it can be given as follows.

$$I_{norm} = \frac{I - I_{min}}{I_{max} - I_{min}} \tag{20}$$

Data augmentation is incorporated into the methodology to further enhance the learning efficacy of deep learning models once the preprocessing phase is finished. These changes contribute to the training datasets diversification, which may result in a more resilient model with improved ability to generalize new data. Data augmentation helps to improve the generalization ability of the deep learning model by exposing it to a wider variety of images, thereby reducing overfitting and improving the model's ability to learn from limited data. For the augmentation process, two types of techniques are employed: random crop and random flip. Random crop involves selecting a random portion of the image while preserving its aspect ratio. This technique helps introduce variability in the training data by focusing on different regions of the image. On the other hand, random flip involves horizontally flipping the image. This augmentation technique helps the model to learn recognize objects and patterns from different orientations, further enhancing its ability to generalize and perform well on unseen data. By combining these two types of augmentations, a more diverse and robust training dataset is created, facilitating better learning and improved performance of the model.

4.3 Explanation of transfer learning architectures and unsupervised autoencoders

In this section, various transfer learning algorithms are explored to use in comparative analysis. The aim is to assess the performance of the proposed model in comparison to different transfer learning models including DenseNet-121, ResNet-50, VGG19, VGG16, Inception ResNet-v2, Inception-v3, EfficientNet-B0, EfficientNet-B1, Xception, MobileNet, and ResNet-152.

DenseNet-121 is a CNN-based architecture renowned for its dense connectivity pattern. Within this architecture, each layer is directly connected to every other layer in a feed-forward manner, making it highly parameter-efficient [41]. ResNet-50, on the other hand, is a variation of the Residual Neural Network (ResNet) architecture that was specifically developed to tackle the vanishing gradient problem in deep networks. This strategic addition ensures that gradient information is preserved during training, thereby enabling more effective learning as describe in Eq. 21.

$$x_u = x_u\left(\begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix} + \frac{\partial \mathcal{F}}{\partial x}\right) \tag{21}$$

ResNet-152, part of the ResNet family, is a CNN architecture introduced in 2015 by [39] for image recognition. By incorporating skip connections, ResNet-152 simplifies the training and optimization of deep neural networks.

VGG19 and VGG16 are part of a distinct category of transfer learning models, originating from the Visual Geometry Group (VGG) architecture, renowned for its simplicity and uniformity. These models consist of several layers of small-sized convolutional filters, followed by max-pooling layers. VGG19 has 19 layers, whereas VGG16 has 16 layers [71]. Inception ResNet-v2, another transfer learning model, combines the strengths of both Inception and ResNet architectures, resulting in a powerful hybrid model. The presence of residual connections ensures efficient gradient propagation during training [86].

The EfficientNet family comprises models aimed at achieving an optimal balance between model size and accuracy. EfficientNet-B0 serves as the baseline model within this family and leverages compound scaling to optimize depth, width, and resolution simultaneously. On the other hand, EfficientNet-B1 is designed to be slightly larger and more accurate than B0. It adopts the same compound scaling

technique but with greater model size and complexity when compared to B0 [90].

Xception, which stands for Extreme Inception, is a CNN architecture inspired by the Inception model. Xception employs depthwise separable convolutions, effectively reducing the number of parameters and computational burden when compared to conventional convolutions [19]. Finally, we have MobileNet, a family of lightweight CNN architectures, specially designed for mobile and embedded devices. These models leverage depthwise separable convolutions to decrease computational demands and model size without compromising on reasonable accuracy [40, 76].

Next, a discussion about various types of autoencoders employed in this study is presented for the purpose of conducting a comparative analysis in this paper. A Convolutional Autoencoder (CAE) is a specialized type of autoencoder that employs convolutional layers in its encoder and decoder, rather than fully connected layers. It is specifically designed for handling image data and excels in capturing spatial patterns and features from images. By using convolutional layers, the CAE can effectively reduce the dimensionality of the input data while retaining crucial features [101]. Variational Autoencoder (VAE) is a probabilistic variant of the conventional autoencoder. Unlike a standard autoencoder that merely learns an encoded representation of the input data, VAE also models the underlying probability distribution of the encoded data. This unique characteristic enables VAEs to generate new data samples by sampling from the learned probability distribution [50].

The sparse autoencoder (SPAE) is a type of autoencoder that enforces sparsity constraints during its training process. These constraints encourage the autoencoder to utilize only a limited number of neurons in its hidden layer to represent the input data. As a result, the autoencoder produces a more concise and efficient representation of the data, offering benefits such as reducing overfitting and enhancing generalization in certain situations [68].

Undercomplete and overcomplete autoencoders are two variations of the traditional autoencoder. An undercomplete autoencoder has a hidden layer with fewer neurons than the input layer, which results in a compressed representation of the input data. On the contrary, an overcomplete autoencoder has a hidden layer with more neurons than the input layer, leading to a redundant representation of the data. This allows the autoencoder to learn multiple representations of the same data, offering more flexibility but also increasing the risk of overfitting [14, 89].

4.4 Experiments setting and hyper-parameter configuration

In the following section, a comparative analysis to assess the performance of our proposed model is presented in comparison to state-of-the-art transfer learning models and unsupervised autoencoders. Here, the experiment setting and hyperparameters configuration utilized during the experiments and comparisons is discussed. Initially, details about the experiment setting is provided, followed by a comprehensive discussion of the hyperparameters configuration for all the other algorithms trained for comparative purposes. The hyperparameter configuration of proposed BiViT is presented at the beginning of Sect. 4.

The experiment setting remains consistent across all models, including the proposed BiViT model. A standardized process of loading, preprocessing, and splitting the data into training and testing sets is followed. Subsequently, all algorithms are trained for the same number of epochs (200), using a uniform learning rate (0.001), optimizer (Adam), and batch size of 32. For the transfer learning models, the weights are initialized with "imagenet" weights and utilized only for the classifier layer of multi-class classification after the flattened layer. For the learning rate, the value is set to default as it has shown satisfactory performance. The number of epochs is 200 to allow sufficient time for observing trends in accuracy and loss measures. This gives an opportunity to analyze how performance evolves over time. Additionally, since the dataset is extensive due to augmentation, a batch size of 32 is used to facilitate efficient and smooth execution of the experiments.

For unsupervised autoencoder, a 5-layer convolutional encoder and a 4-layer convolutional decoder is used. Throughout all layers, the filter size remains consistent at (3×3). The latent space has a dimension of 2, which we use for classifying the disease. The autoencoders are trained for 10 epochs with a learning rate of 0.001, a batch size of 32, and the Mean Squared Error (MSE) serving as the loss function.

4.5 Comparative analysis on Alzheimer disease diagnosis

In this section and the following subsections, a comparative analysis is performed on different transfer learning algorithms for AD classification. Transfer learning is a technique where a pre-trained neural network is used as a starting point for a new task. The pre-trained neural network has learned general features from a large dataset and these features can be reused for a new task, which can save time and computational resources. Two tables are created for the experiment, one with data augmentation experiments and the other without data augmentations. One table contains data that has been subjected to data augmentation techniques, which involves artificially generating new data by making changes to existing data samples. The other table contains data that has not undergone any data augmentation. Both tables are likely used to compare the performance of machine learning

or other data-driven models on the augmented versus non-augmented data.

Data augmentation is a technique where new training samples are generated by applying transformations such as rotations, flips, and scaling to the original images. This can increase the diversity of the training data and improve the performance of the model. The third table in the analysis described the class-wise prediction performance of the top three best-performing models. This table provided additional insights into how each algorithm performed for different classes of AD. Multiple algorithms are trained and the results are reported. Each algorithm has a different architecture and varying number of layers. The performance of these algorithms is compared in terms of accuracy, recall, precision, F1-score and AUC.

Section 4.5.2 discusses the usage of various deep unsupervised learning algorithms, also known as autoencoders, to classify instances with abnormality. These autoencoders included CAE, VAE, SPAE, and undercomplete and overcomplete CAEs. The performances of these autoencoders are evaluated based on their ability to extract relevant features from the input images and classify instances with abnormality.

4.5.1 Alzheimer detection with transfer learning algorithms

In this section, a comparative analysis is presented between different transfer learning algorithms for classifying AD and the results presented in Table 3 in terms of accuracy, recall, precision, F1-score, and AUC. It is important to note that the results in Table 3 are obtained by applying data augmentation techniques to the Alzheimer data, which means that the algorithms were trained on augmented data. Then, the models are tested using test data to evaluate their performance. The visual performance of the proposed BiViT algorithm (with data augmentation) after each epoch is illustrated in Fig. 12.

The performance of different algorithms in terms of various measures such as precision, recall, F1-score, accuracy, and AUC are compared in a Table 3. From these results, it can be concluded that the proposed BiViT model outperformed competing algorithms in terms of all performance measures. Furthermore, it can be observed that the precision is low for all the algorithms whereas the recall is high. This suggests that the algorithms are able to detect most of the positive cases, but at the cost of many false positives. However, the BiViT model can achieve good results in terms of both recall and F1 score.

In contrast, results illustrated in Table 4 reveals that all transfer learning algorithms perform remarkably well even without the use of data augmentation techniques. In addition, the results are significantly superior to those achieved by applying data augmentation in previous studies. A visual performance in terms of progress of the proposed BiViT algorithm after each epoch is depicted in Fig. 13, where data augmentation was not used. It can be observed that the proposed BiViT model outperforms all other transfer learning models with the highest accuracy and recall scores. Based on these results, it can be concluded that data augmentation techniques may not always be necessary to achieve good performance in image classification tasks, and that the proposed BiViT model shows promising results in this regard. Moreover, the Table 5 shows the classwise comparative analysis of the 4 best performing algorithms.

Table 3 The table provides a summary of the results obtained by various transfer learning algorithms for AD diagnosis with different data augmentation techniques

Transfer learning for Alzheimer diagnosis (with data augmentation)					
Algorithm	Accuracy	Recall	Precision	F1-Score	AUC-ROC
DenseNet-121	0.56 ± 0.057	1.00 ± 0.003	0.33 ± 0.016	0.50 ± 0.018	0.66 ± 0.026
Resnet-50	0.52 ± 0.033	0.96 ± 0.032	0.39 ± 0.023	0.55 ± 0.022	0.81 ± 0.028
VGG16	0.62 ± 0.034	0.93 ± 0.017	0.45 ± 0.016	0.60 ± 0.014	0.87 ± 0.016
Inception Resnet-V2	0.58 ± 0.039	1.00 ± 0.025	0.34 ± 0.021	0.51 ± 0.020	0.77 ± 0.024
Inception V3	0.53 ± 0.043	1.00 ± 0.005	0.25 ± 0.027	0.40 ± 0.032	0.53 ± 0.033
EfficientNet-B0	0.50 ± 0.70	1.00 ± 0.005	0.25 ± 0.028	0.40 ± 0.033	0.66 ± 0.048
ResNet-101	0.54 ± 0.016	0.93 ± 0.021	0.39 ± 0.018	0.55 ± 0.016	0.81 ± 0.007
VGG19	0.63 ± 0.030	0.77 ± 0.065	0.52 ± 0.026	0.63 ± 0.023	0.87 ± 0.013
EfficientNet-B1	0.51 ± 0.092	1.00 ± 0.004	0.25 ± 0.031	0.40 ± 0.037	0.66 ± 0.070
Xception	0.59 ± 0.040	1.00 ± 0.003	0.32 ± 0.019	0.49 ± 0.022	0.68 ± 0.031
MobileNet	0.60 ± 0.042	1.00 ± 0.011	0.34 ± 0.020	0.51 ± 0.021	0.77 ± 0.025
ResNet-152	0.54 ± 0.021	0.99 ± 0.012	0.36 ± 0.015	0.53 ± 0.015	0.82 ± 0.007
Proposed BiViT	0.93 ± 0.010	0.97 ± 0.021	0.75 ± 0.132	0.85 ± 0.026	0.98 ± 0.011

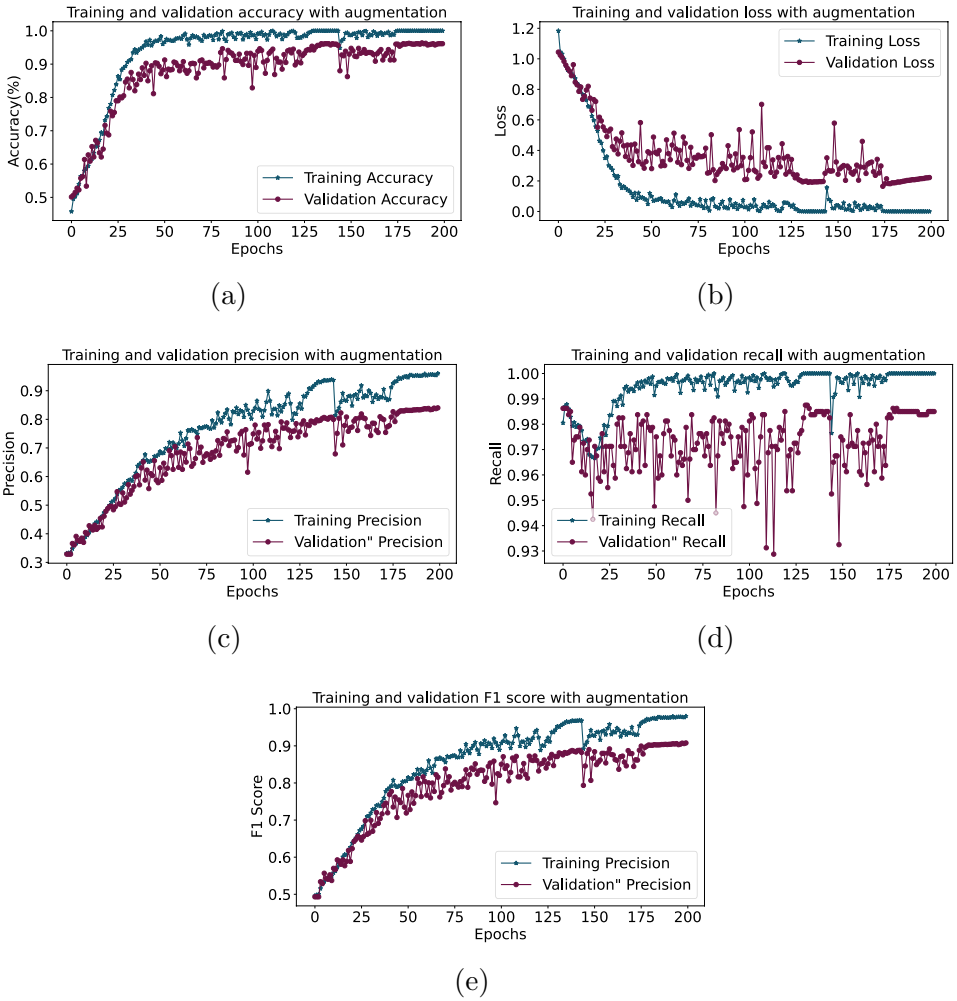


Fig. 12 Performance of proposed BiViT with data augmentation technique

Table 4 The summary table presents the results achieved by different transfer learning methods for diagnosing AD, without the use of any data augmentation techniques

Transfer learning for Alzheimer diagnosis (without data augmentation)					
Algorithm	Accuracy	Recall	Precision	F1-Score	AUC-ROC
DenseNet-121	0.59 ± 0.113	1.00 ± 0.000	0.32 ± 0.027	0.48 ± 0.033	0.67 ± 0.064
Resnet-50	0.71 ± 0.038	0.97 ± 0.015	0.46 ± 0.027	0.63 ± 0.026	0.91 ± 0.018
VGG16	0.88 ± 0.066	0.98 ± 0.005	0.63 ± 0.067	0.77 ± 0.057	0.97 ± 0.025
Inception Resnet-V2	0.83 ± 0.083	1.00 ± 0.003	40.35 ± 0.017	0.52 ± 0.020	0.77 ± 0.033
Inception V3	0.66 ± 0.049	1.00 ± 0.001	0.36 ± 0.020	0.53 ± 0.023	0.80 ± 0.035
EfficientNet-B0	0.56 ± 0.077	1.00 ± 0.003	0.25 ± 0.038	0.40 ± 0.046	0.79 ± 0.041
ResNet-101	0.63 ± 0.035	0.95 ± 0.024	0.43 ± 0.018	0.59 ± 0.017	0.88 ± 0.019
VGG19	0.89 ± 0.066	1.00 ± 0.007	0.52 ± 0.042	0.68 ± 0.041	0.97 ± 0.027
EfficientNet-B1	0.36 ± 0.095	1.00 ± 0.002	0.25 ± 0.038	0.40 ± 0.046	0.76 ± 0.040
Xception	0.79 ± 0.052	1.00 ± 0.001	0.39 ± 0.020	0.56 ± 0.022	0.84 ± 0.021
MobileNet	0.96 ± 0.035	1.00 ± 0.001	0.63 ± 0.063	0.78 ± 0.054	0.99 ± 0.013
ResNet-152	0.68 ± 0.043	0.92 ± 0.025	0.46 ± 0.023	0.61 ± 0.021	0.90 ± 0.019
Proposed BiViT	0.96 ± 0.102	0.98 ± 0.174	0.88 ± 0.010	0.93 ± 0.145	0.99 ± 0.038

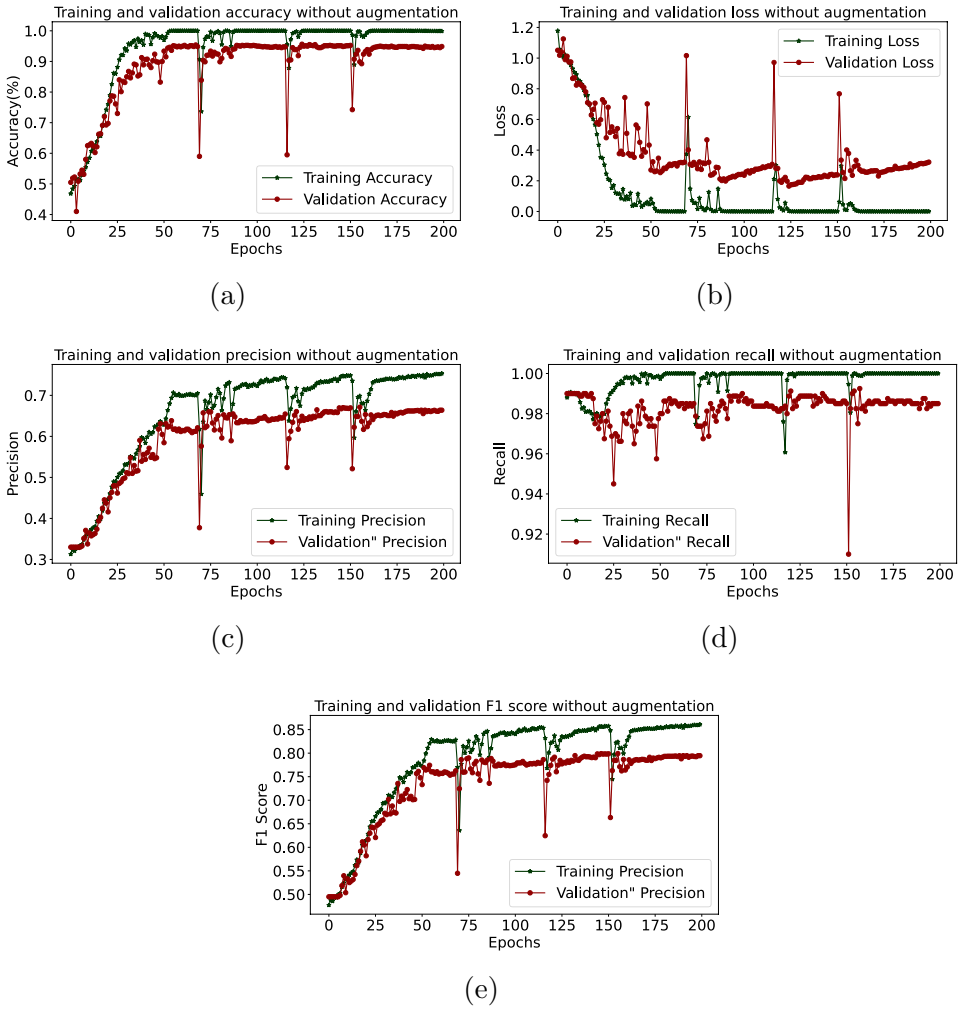


Fig. 13 Performance of proposed BiViT model without data augmentation

4.5.2 Alzheimer detection with deep unsupervised learning

Deep autoencoders are a type of unsupervised learning technique that can extract features by modeling latent manifold in data. These features can then be used to classify instances using a simple classifier. In this study, different autoencoders such as CAE, VAE, SPAE, undercomplete and overcomplete AE, are used for AD classification and the results are presented in Table 6. Each of these autoencoders were trained and tested on the dataset to classify instances with abnormality. The results show the performance of each autoencoder in terms of different evaluation metrics, such as accuracy, recall, and F1-score.

The results conclude that autoencoders perform well and can compete with transfer learning algorithms in terms of AD classification. In fact, some of the autoencoders even outperformed the transfer learning algorithms. Moreover, the SPAE performed the best out of all the autoencoders which means that the SPAE can extract the most informative features from the images and classify AD instances more accurately than the other autoencoders. The performance of SPAE, CAE, and VAE can be observed in Figs. 14, 15, and 16. Figures 17 and 18 display the performance of over and under autoencoders.

4.6 Comparative analysis on cognitive diseases diagnosis

In this section and its subsequent subsections (Sects. 4.6.1 and 4.6.2), a comparison is made among different transfer learning techniques for the purpose of classifying cognitive disorders. Two tables are presented, with data augmentation (Table 7) and without data augmentation (Table 8). The tables are used to compare the performance of machine learning models on augmented versus non-augmented data. Additionally, Table 9 provides the class-wise performance of the top performing four algorithms among all. These sections also present a comparison of various autoencoders for cognitive disorders classification illustrated in Table 10.

Table 5 Class-wise performance of best 4 models in AD classification case

Algorithms	Alzheimer stages	With augmentation			Without augmentation			
		Accuracy	Recall	Precision	Accuracy	Recall	Precision	F1-Score
BiViT	Mild-demented	0.93	0.90	0.93	0.96	0.93	0.98	0.95
	Moderate-demented	0.93	1.00	0.80	0.96	0.83	0.83	0.83
	Non-demented	0.93	0.96	0.93	0.96	0.98	0.96	0.97
	Very Mild-demented	0.93	0.90	0.93	0.96	0.95	0.97	0.96
Weighted average		0.93	0.90	0.93	0.96	0.96	0.96	0.96
VGG-16	Mild-demented	0.62	0.17	0.69	0.88	0.92	0.76	0.83
	Moderate-demented	0.62	0.58	1.00	0.88	1.00	1.00	1.00
	Non-demented	0.62	0.61	0.78	0.88	0.86	0.96	0.90
	Very Mild-demented	0.62	0.79	0.50	0.88	0.91	0.85	0.88
Weighted average		0.62	0.62	0.67	0.88	0.88	0.89	0.89
MobileNet	Mild-demented	0.60	0.72	0.35	0.96	0.91	0.97	0.94
	Moderate-demented	0.60	0.70	1.00	0.96	1.00	1.00	1.00
	Non-demented	0.60	0.60	0.77	0.96	0.95	0.97	0.96
	Very Mild-demented	0.60	0.54	0.60	0.96	0.98	0.94	0.96
Weighted average		0.60	0.54	0.60	0.96	0.96	0.96	0.96
VGG-19	Mild-demented	0.62	0.36	0.53	0.88	0.92	0.83	0.87
	Moderate-demented	0.62	0.27	0.75	0.88	1.00	1.00	1.00
	Non-demented	0.62	0.74	0.73	0.88	0.92	0.87	0.89
	Very Mild-demented	0.62	0.60	0.53	0.88	0.79	0.92	0.85
Weighted average		0.62	0.63	0.63	0.88	0.88	0.88	0.88

Table 6 The table presents a comparison between various autoencoders (an unsupervised learning technique) for diagnosing AD without using data augmentation

Comparison of unsupervised learning for Alzheimer diagnosis						
Algorithm	Accuracy	Recall	Precision	F1-Score	AUC	AE(mae)
Convolutional AE	0.9538	0.9825	0.5692	0.7231	0.9595	0.086
Variational AE	0.8462	0.8750	0.5952	0.7083	0.8569	0.090
Sparse AE	0.9825	0.9950	0.4500	0.6210	0.8794	0.087
UnderComplete AE	0.8500	0.9962	0.4377	0.6088	0.8277	0.079
OverComplete AE	0.9400	0.9912	0.5420	0.7016	0.9513	0.065

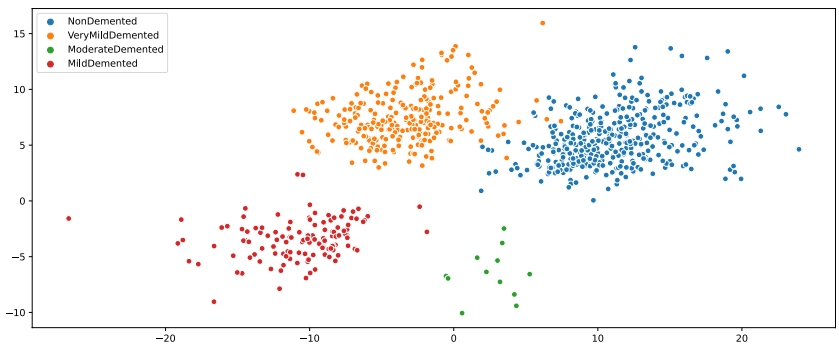


Fig. 14 Latent space of sparse autoencoder

4.6.1 Transfer learning for cognitive disorder diagnosis

Different transfer learning techniques, including the proposed BiViT model, were employed to classify cognitive disorders. However, due to limited and imbalanced dataset, the performance of not even a single algorithm is upto the mark even after utilizing data augmentation methods. Table 7 presents the results of various algorithms trained on the cognitive disease dataset. It can be seen that all the algorithms have an accuracy within the

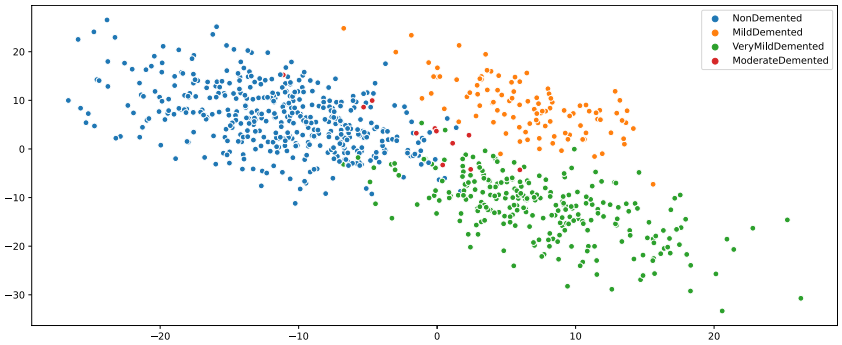


Fig. 15 Latent space of convolutional autoencoder

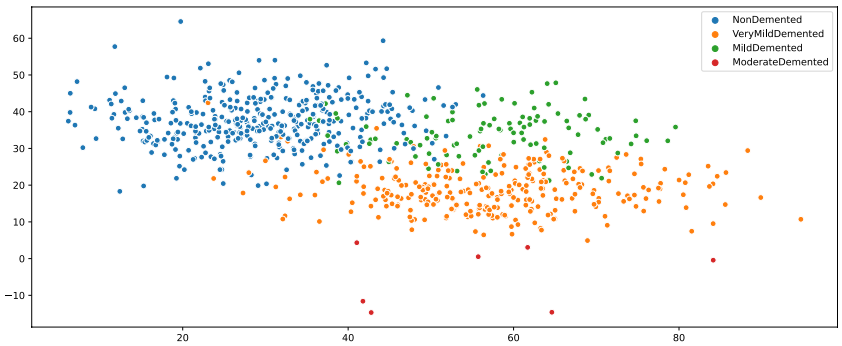


Fig. 16 Latent space of variational autoencoder

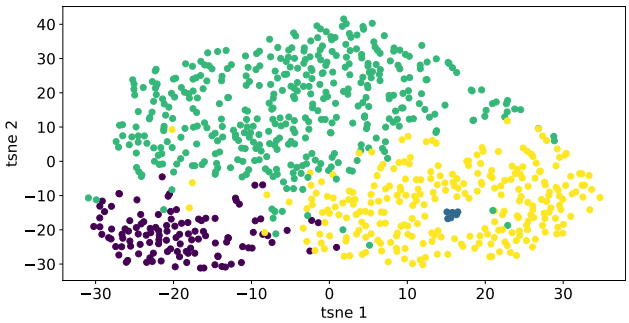


Fig. 17 Latent space of overcomplete autoencoder

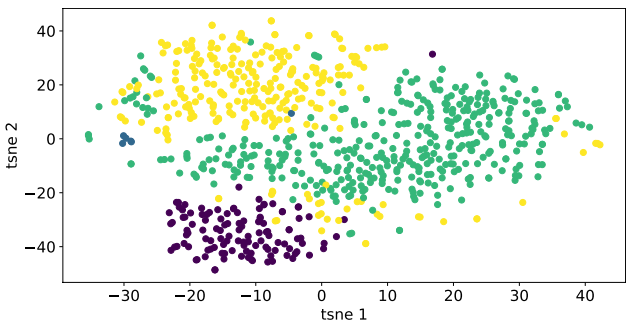


Fig. 18 Latent space of undercomplete autoencoder

range of 35–45%. Although the recall score is high, the precision and F1-score are low. However, the proposed BiViT model can perform very well in terms of accuracy and F1-score. Table 8 provides information on the performance of the same set of algorithms as in Table 7, but this time they are trained without performing data augmentation.

The Table 8 compares the performance of different transfer learning algorithms without data augmentation techniques. Although the range of accuracy remains similar, i.e., 40–45%, the precision and F1-score are consistently low. Even the proposed BiViT model performs poorly in this scenario, which is primarily attributed to the limited and imbalanced nature of the dataset. It is suggested that with an increase in data quantity, there may be an improvement in the performance of all algorithms in the future. The Table 9 demonstrates the class-specific performance of the five top performing algorithms for cognitive disorder classification.

4.6.2 Deep unsupervised learning for cognitive disorder diagnosis

Table 10 presents the performance of various autoencoders used for classification of different cognitive impairments. From the statistics, it can be concluded that the performance of autoencoder-based models for cognitive disorders classification is not satisfactory. This could be due to the limited nature of the dataset, which may not provide enough examples to help autoencoders generalize well. However, the SPAE performs relatively well and is competitive with transfer learning models.

Table 7 The table presents the results of various transfer learning methods employed for the detection of cognitive disorders (with the use of data augmentation)

Transfer learning for cognitive disorders detection					
Algorithm	Accuracy	Recall	Precision	F1-Score	AUC-ROC
DenseNet-121	0.35 ± 0.073	1.00 ± 0.003	0.20 ± 0.004	0.33 ± 0.005	0.55 ± 0.028
Resnet-50	0.41 ± 0.060	0.66 ± 0.078	0.29 ± 0.026	0.40 ± 0.016	0.69 ± 0.022
VGG16	0.42 ± 0.034	0.64 ± 0.055	0.33 ± 0.032	0.44 ± 0.024	0.73 ± 0.027
Inception Resnet-V2	0.36 ± 0.060	0.99 ± 0.020	0.21 ± 0.010	0.34 ± 0.011	0.66 ± 0.027
Inception V3	0.31 ± 0.072	1.00 ± 0.005	0.20 ± 0.004	0.35 ± 0.005	0.50 ± 0.028
EfficientNet-B0	0.42 ± 0.127	1.00 ± 0.024	0.20 ± 0.007	0.33 ± 0.008	0.68 ± 0.063
ResNet-101	0.44 ± 0.064	0.60 ± 0.097	0.35 ± 0.036	0.44 ± 0.016	0.70 ± 0.024
VGG19	0.35 ± 0.053	0.86 ± 0.043	0.27 ± 0.017	0.42 ± 0.018	0.71 ± 0.027
EfficientNet-B1	0.20 ± 0.123	1.00 ± 0.008	0.20 ± 0.05	0.33 ± 0.006	0.63 ± 0.049
Xception	0.43 ± 0.062	1.00 ± 0.009	0.21 ± 0.006	0.35 ± 0.007	0.64 ± 0.031
MobileNet	0.45 ± 0.062	0.92 ± 0.029	0.24 ± 0.010	0.38 ± 0.012	0.69 ± 0.032
ResNet-152	0.44 ± 0.033	0.61 ± 0.088	0.35 ± 0.036	0.44 ± 0.016	0.69 ± 0.013
Proposed BiViT	0.45 ± 0.014	0.69 ± 0.020	0.33 ± 0.018	0.46 ± 0.023	0.73 ± 0.011

Table 8 The table presents the results of various transfer learning methods employed for the detection of cognitive disorders (without data augmentation)

Transfer learning for cognitive disorders detection (without data augmentation)					
Algorithm	Accuracy	Recall	Precision	F1-Score	AUC-ROC
DenseNet-121	0.42 ± 0.047	0.90 ± 0.027	0.24 ± 0.010	0.38 ± 0.011	0.71 ± 0.031
Resnet-50	0.44 ± 0.030	0.86 ± 0.046	0.24 ± 0.014	0.37 ± 0.016	0.71 ± 0.018
VGG16	0.45 ± 0.040	0.82 ± 0.052	0.29 ± 0.022	0.43 ± 0.021	0.74 ± 0.018
Inception Resnet-V2	0.39 ± 0.039	0.93 ± 0.019	0.23 ± 0.006	0.37 ± 0.008	0.69 ± 0.018
Inception V3	0.40 ± 0.032	0.99 ± 0.003	0.20 ± 0.001	0.34 ± 0.001	0.60 ± 0.022
EfficientNet-B0	0.13 ± 0.135	1.00 ± 0.014	0.20 ± 0.010	0.33 ± 0.013	0.60 ± 0.069
ResNet-101	0.40 ± 0.021	0.69 ± 0.076	0.31 ± 0.032	0.43 ± 0.014	0.71 ± 0.014
VGG19	0.43 ± 0.036	0.41 ± 0.068	0.45 ± 0.039	0.43 ± 0.026	0.75 ± 0.023
EfficientNet-B1	0.43 ± 0.132	1.00 ± 0.007	0.20 ± 0.005	0.33 ± 0.007	0.62 ± 0.061
Xception	0.42 ± 0.031	0.82 ± 0.042	0.27 ± 0.016	0.41 ± 0.017	0.73 ± 0.025
MobileNet	0.50 ± 0.024	0.92 ± 0.015	0.24 ± 0.005	0.38 ± 0.006	0.75 ± 0.016
ResNet-152	0.44 ± 0.027	0.53 ± 0.061	0.40 ± 0.033	0.46 ± 0.019	0.72 ± 0.016
Proposed BiViT	0.41 ± 0.031	0.62 ± 0.064	0.32 ± 0.023	0.43 ± 0.019	0.69 ± 0.021

4.7 Comparison with state-of-the-art approaches

There are various models available in the literature for classifying AD and cognitive disorders. This subsection aims to compare the proposed BiViT model with other state-of-the-art models in the literature, presented in Table 11.

5 Discussion

This section discusses research findings as well as its limitations, applications, and future directions. The Sect. 4.5 presents the outcomes of the suggested BiViT, autoencoders, and transfer learning models when applied on the AD dataset. Table 3 shows that BiViT outperforms all other transfer learning models with 93% accuracy when data augmentation techniques are applied to AD data. However, the results of not applying augmentation are better than those of applying data augmentation 4.

It is true that when augmentation techniques are used on medical image datasets, they might not always produce significant improvements. The statement is made for multiple reasons. First of all, intricate anatomical structures and subtle characteristics are frequently seen in medical images, which make it difficult to enhance them without introducing distortions or unrealistic variations. Additionally, medical image datasets are often smaller and more specialized compared to general image datasets, making it challenging to find augmentation strategies that effectively capture the variability within the dataset without introducing biases or artifacts. Table 5 presents the class-wise performance and clearly shows that, when applied to the AD dataset, BiViT and MobileNet yield the best performance matrices.

Table 9 Class-wise performance of best 4 models for cognitive disorders detection

Algorithms	Alzheimer stages	With augmentation			Without augmentation			
		Accuracy	Recall	Precision	Accuracy	Recall	Precision	F1-Score
BiViT	Alzheimer disease	0.44	0.34	0.43	0.41	0.19	0.33	0.24
	Cognitive normal	0.44	0.67	0.46	0.41	0.66	0.46	0.54
	Early-mild cognitive impairment	0.44	0.28	0.39	0.41	0.19	0.36	0.25
	Late-mild cognitive impairment	0.44	0.04	0.50	0.41	0.20	0.19	0.20
	Mild cognitive impairment	0.44	0.29	0.41	0.41	0.27	0.34	0.30
Weighted average		0.44	0.44	0.43	0.41	0.41	0.39	0.38
MobileNet	Alzheimer disease	0.45	0.31	0.44	0.50	0.31	0.48	0.38
	Cognitive normal	0.45	0.76	0.49	0.50	0.76	0.51	0.61
	Early-mild cognitive impairment	0.45	0.40	0.37	0.50	0.35	0.51	0.41
	Late-mild cognitive impairment	0.45	0.00	0.00	0.50	0.15	0.75	0.25
	Mild cognitive impairment	0.45	0.06	0.40	0.50	0.27	0.40	0.32
Weighted average		0.45	0.45	0.41	0.50	0.50	0.50	0.47
VGG-16	Alzheimer disease	0.42	0.43	0.35	0.45	0.25	0.37	0.30
	Cognitive normal	0.42	0.82	0.46	0.45	0.58	0.59	0.58
	Early-mild cognitive impairment	0.42	0.01	0.50	0.45	0.51	0.32	0.39
	Late-mild cognitive impairment	0.42	0.04	0.50	0.45	0.30	0.29	0.29
	Mild cognitive impairment	0.42	0.06	0.19	0.45	0.26	0.43	0.32
Weighted average		0.42	0.42	0.41	0.45	0.45	0.46	0.44
DenseNet-121	Alzheimer disease	0.35	0.16	0.62	0.42	0.23	0.33	0.27
	Cognitive normal	0.35	0.90	0.93	0.42	0.66	0.48	0.56
	Early-mild cognitive impairment	0.35	0.89	0.26	0.42	0.25	0.38	0.30
	Late-mild cognitive impairment	0.35	0.04	1.00	0.42	0.10	0.17	0.12
	Mild cognitive impairment	0.35	0.08	0.33	0.42	0.27	0.32	0.29
Weighted average		0.35	0.35	0.50	0.42	0.42	0.39	0.40

Table 10 The comparison table presents various autoencoder models (an unsupervised learning technique) and their performance evaluation for the detection of cognitive disorders (without data augmentation)

Comparison of unsupervised learning for cognitive disorders detection						
Algorithm	Accuracy	Recall	Precision	F1-Score	AUC	AE(mae)
Convolutional AE	0.4242	0.4470	0.3978	0.4355	0.6577	0.086
Variational AE	0.4470	0.6616	0.2495	0.3701	0.6166	0.167
Sparse AE	0.4596	0.6818	0.2695	0.3938	0.7047	0.076
UnderComplete AE	0.4520	0.8131	0.2822	0.4200	0.7256	0.090
OverComplete AE	0.2601	0.6869	0.2477	0.3659	0.6030	0.1164

Although MobileNet has demonstrated its effectiveness in AD classification, our transformer-based BiViT model presents a strong substitute. Although CNNs are the foundation of MobileNet and are widely recognized for their effectiveness, they are naturally limited in their ability to capture long-range dependencies and maintain spatial relationships across images. Transformers, on the other hand, excel in these areas as well, using self-attention mechanisms to identify complex spatial patterns that are essential for precise diagnosis. Furthermore, because of their intricate hierarchical architectures, CNNs may be difficult to interpret, which makes it difficult to comprehend model decisions-a crucial component in medical applications. Transformers, on the other hand, provide improved interpretability and transparency in decision-making because of their attention mechanisms.

Following that, the results of autoencoder-based models are displayed in Table 6, which also demonstrates how well these models work when used with the AD dataset. Autoencoders attain high accuracy at the expense of precision and

Table 11 The presented table illustrates a comparison between the proposed BiViT model and other state-of-the-art models available in the existing literature

	Model-Name	Accuracy	Recall	Precision	F1-score	AUC
[36]	ADDTLA	91.70%	93.70%	91.50%	92.50%	–
[99]	Hybrid CNN	90.00%	90.00%	90.00%	90.00%	90%
[82]	DenseNet121	96.59%	97.25%	97.25%	97.50%	–
[46]	CNN	92.78%	90.78%	–	94.00%	–
[82]	ResNet50	93.52%	92%	95%	93.75%	–
[1]	Modified AlexNet	95.70	92.30%	91.90%	94.70%	–
[82]	VGG19	95.08%	93.00%	96.50%	94.75%	–
[28]	SqueezeNet+LSTM	87.50%	–	–	–	–
[82]	Xception	89.77%	88.25%	92.00%	89.50%	–
[5]	CNN	0.97%	–	–	–	–
[82]	EfficientNetB7	83.20%	68.25%	87%	73.5%	–
[4]	EfficientNetV2B1	90.37%	89.76%	–	90.06%	–
[82]	EfficientNetB7	83.20%	68.25%	87.00%	73.5%	–
[69]	PLF-VIT	81.25%	–	–	–	–
[100]	SMIL-DeiT	93.20%	–	–	–	–
[25]	ViT	87.5%	–	–	84.00%	–
Proposed	BiViT	96.38	97.87	88.28	92.84	99.47

F1-score. The same transfer learning algorithms, autoencoder, and suggested BiViT are then applied to cognitive disorder data with and without augmentation in Sect. 4.6. All models and the suggested BiviT-perform the worst in both augmentation and non-augmentation scenarios. Very little data and an uneven distribution of data are the causes of this, as was previously discussed in the sections above.

Finally, our suggested model shows up as a strong competitor with competitive performance metrics in AD classification when compared to state-of-the-art architectures, as evidenced by its effectiveness in the literature. Our model is positioned as an AD detection system that operates in real-time and can be easily integrated with the internet of things (IoT), making it adaptable to a range of healthcare environments. There are many uses for the BiViT model, especially in hospitals where accurate and timely diagnosis is critical. It's crucial to recognise some inherent limitations in the suggested BiViT model, though. Its training is primarily based on a single AD dataset, which means that it needs to be adjusted for use in various hospital settings. This emphasises how crucial it is to increase the size of the dataset in order to improve model performance, since larger datasets make it easier to extract more useful and subtle features. Although the model performs well on the current dataset, because it was trained on a small amount of data and requires domain-specific knowledge, its generalizability to other datasets may need careful fine-tuning. Given these challenges, scientists can now improve the functionality and applicability of this model in a range of clinical settings.

6 Conclusion

The early-stage detection of AD and different cognitive declines are crucial for the patient's health. To address this, a computer-aided diagnosis system is proposed that utilizes 2D-MRI images to detect different cognitive disorders including AD by incorporating PCES and MLF mechanisms. In contrast to the simple ViT, which utilizes only a single patching mechanism, the proposed BiViT employs two parallel coupled encoding mechanisms including simple patch encoding and image tokenization, to encode information more efficiently. A comparative analysis is presented using two different datasets of Alzheimer disease and cognitive disorders. Various state-of-the-art models from literature including transfer learning and autoencoder-based models are included in study to compete with the proposed algorithm. The accuracy, recall, precision, F1-score, and AUC for the AD classification task are reported as 96.38%, 97.87%, 88.28%, 92.84%, and 99.47%, respectively. For cognitive disorders, the proposed BiViT algorithm achieved an accuracy, recall, precision, F1-score, and AUC of 44.94%, 68.69%, 33.29%, 45.38%, and 72.62%, respectively. It means that, the proposed algorithm performed well in the case of AD classification but had lower performance for cognitive disorders due to limited and imbalanced data. Therfore, to improve the model's performance, incorporating diverse and representative datasets is recommended. Additionally, integrating techniques for explainability and interpretability would enhance transparency in the model's decision-making process.

Appendix A: Visualizing patching and tokenization

The conversion of images into a sequence of tokens in ViT is referred to as "patch encoding". Patch encoding includes the division of the input image into non-overlapping patches, which are then linearly projected into a lower-dimensional space to generate a set of embeddings for each patch. Next, positional embeddings are added to the resulting embeddings. The patch embeddings, along with the positional embeddings, are then fed into the transformer encoder. The division of images into patches is illustrated in Fig. 19.

Tokenization in CCT involves the conversion of image patches into a fixed-size sequence of tokens. To achieve this, the patch embeddings undergo several convolutional layers, producing feature maps that are then flattened into a token sequence. The transformed tokens are then fed into the transformer encoder for further processing. The visual representations in Fig. 20 depicts how images are transformed into tokens of a fixed size.

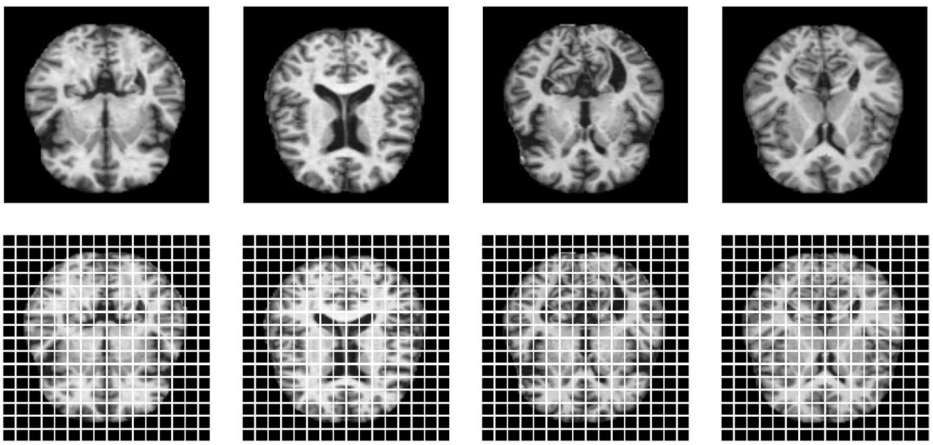


Fig. 19 Images into patches

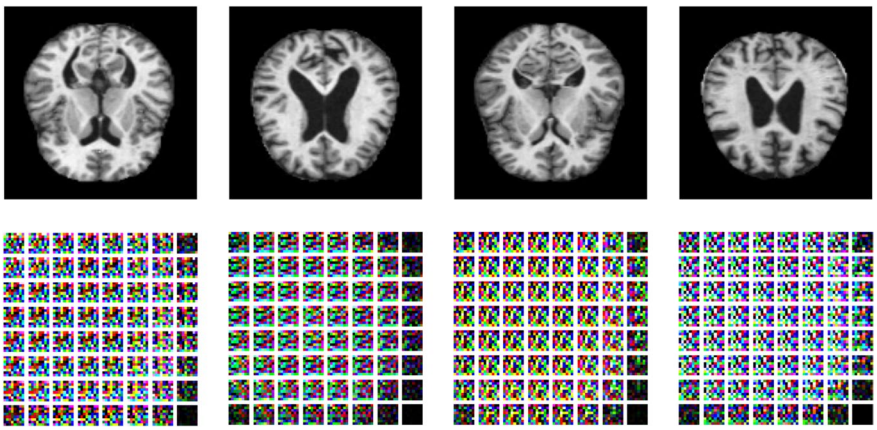


Fig. 20 Visualization of images converted into fixed-size tokens for better understanding

Table 12 Comparative analysis on binary classification in case of AD dataset (in this we take two classes at a time)

Binary class-wise comparative analysis							
No	Classes	Accuracy	Recall	Precision	F1-Score	Top-5	AUC
1	Non-Vs-Mild	0.9497 ± 0.078	0.9329 ± 0.074	0.9553 ± 0.095	0.9421 ± 0.077	1.000 ± 0.000	0.9871 ± 0.043
2	Mild-Vs-Moderate	0.9902 ± 0.012	0.9805 ± 0.014	0.9926 ± 0.009	0.9866 ± 0.011	1.000 ± 0.000	0.9996 ± 0.024
3	Mild-Vs-Very Mild	0.6981 ± 0.011	0.7689 ± 0.076	0.6626 ± 0.075	0.7128 ± 0.035	1.000 ± 0.000	0.7875 ± 0.021
4	Very Mild-Vs-Moderate	0.9967 ± 0.003	0.9976 ± 0.002	0.9967 ± 0.069	0.9967 ± 0.046	1.000 ± 0.000	0.9998 ± 0.002

Table 13 Comparative analysis on binary classification in case of cognitive disease dataset (in this we take two classes at a time)

Binary class-wise comparative analysis							
No	Classes	Accuracy	Recall	Precision	F1-Score	Top-5	AUC
1	Alzheimer vs CN	0.7530 ± 0.046	0.0080 ± 0.000	1.00 ± 0.000	0.0152 ± 0.000	1.000 ± 0.000	0.8001 ± 0.047
2	Alzheimer vs EMCI	0.5714 ± 0.047	0.7329 ± 0.089	0.5463 ± 0.025	0.6336 ± 0.029	1.000 ± 0.000	0.6015 ± 0.049
3	Alzheimer vs LMCI	0.7460 ± 0.134	0.8413 ± 0.062	0.5955 ± 0.062	0.6961 ± 0.029	1.000 ± 0.000	0.7567 ± 0.142
4	Alzheimer vs MCI	0.5161 ± 0.046	0.0806 ± 0.019	1.000 ± 0.3847	0.1435 ± 0.036	1.000 ± 0.000	0.5857 ± 0.043

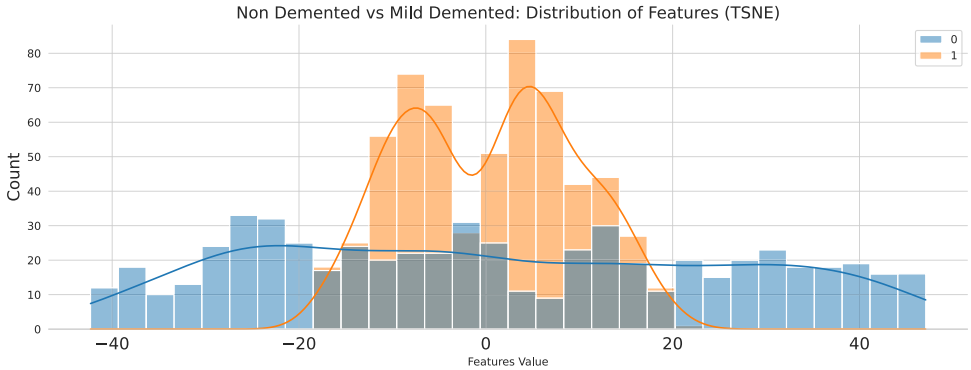


Fig. 21 Feature distribution visualization after feature reduction using TSNE algorithm (Non vs Mild)

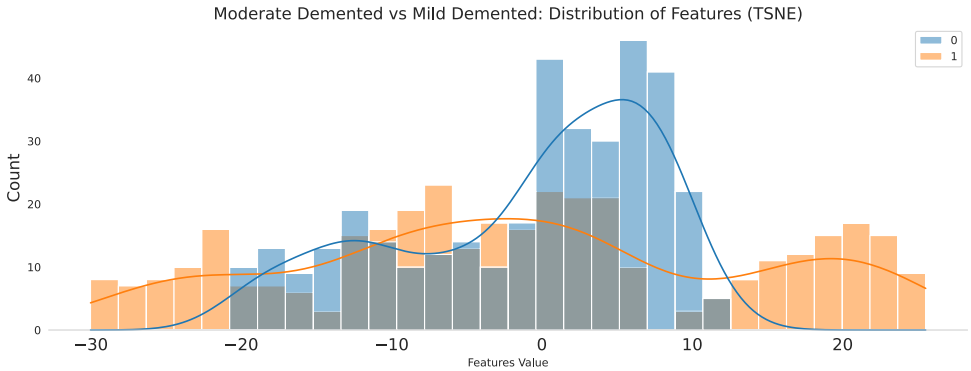


Fig. 22 Feature distribution visualization after feature reduction using TSNE algorithm (Moderate vs Mild)

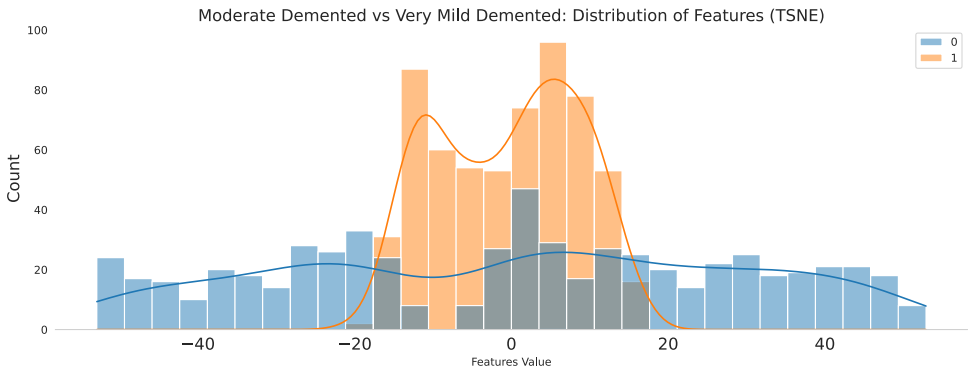


Fig. 23 Feature distribution visualization after feature reduction using TSNE algorithm (Moderate vs Very-Mild)

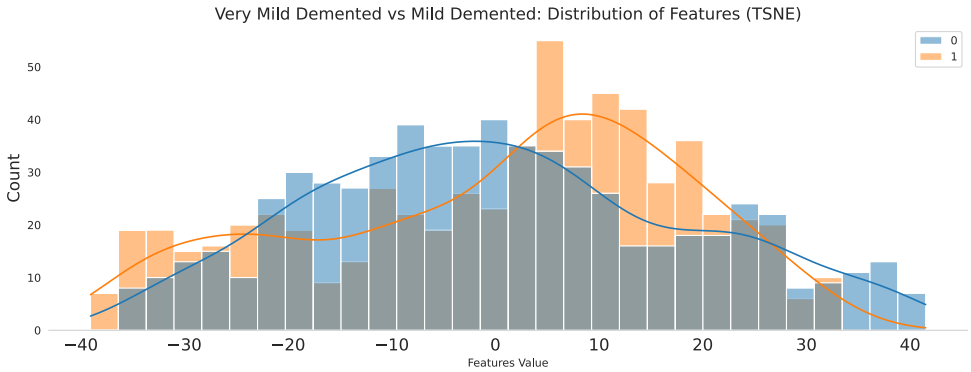


Fig. 24 Feature distribution visualization after feature reduction using TSNE algorithm (Very-Mild vs Mild)

Appendix B: Ablation studies

We conducted an ablation study to assess the effectiveness of our proposed model, BiViT. The study was performed on two datasets: the AD dataset and the Cognitive disorders dataset. In addition, we conducted a comparative analysis by training several transfer learning algorithms and autoencoders on the same datasets. Our results indicate that the proposed BiViT model achieved a very satisfying performance.

To gain further insights into the performance of our model, we conducted experiments in which we tested the BiViT algorithm on two classes at a time and reported the results in the "Appendix C". We found that in some cases of binary classification, it was relatively easy to draw a decision line, while in other cases, it was more challenging. These findings suggest that while our model can be successfully applied to a range of image classification tasks, the complexity of the task may impact its performance.

Appendix C: Comparative analysis (binary classification)

In the following "Appendix" section, we provide a comparative analysis of the BiViT algorithm using the AD and cognitive disorders datasets. Specifically, we evaluate the algorithm's performance in a binary class-wise manner, by selecting two classes at a time, training the BiViT model and comparing the results. The binary class-wise classification results for both datasets are presented in Tables 12 and 13, respectively.

In the following paragraphs we describe about the TSNE-transformed features distributions and the relation scatter plot just after projecting the features from n-dimension to 2-dimesional space. Apart from this, We apply LASSO regression to determine some of the important features only and afterwards evaluate them. Firstly, the visual representations of the TSNE-transformed feature distribution for Alzheimer stages classification are depicted in Figs. 21, 23, 22, and 24. These features are derived from BiViT after training on Alzheimer's stage data. Subsequently, we employ TSNE transformation to map these high-dimensional features into a two-dimensional space. Analyzing these features provides insight into how the model makes decisions internally. The histogram plot of t-SNE transformed BiViT features provides insights into the distribution of data points in the

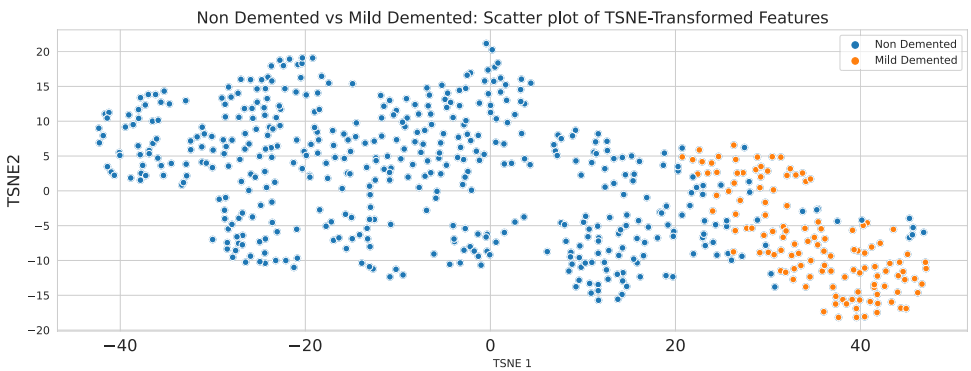


Fig. 25 Visualizing relation between TSNE transformed features after feature reduction (Non vs Mild)

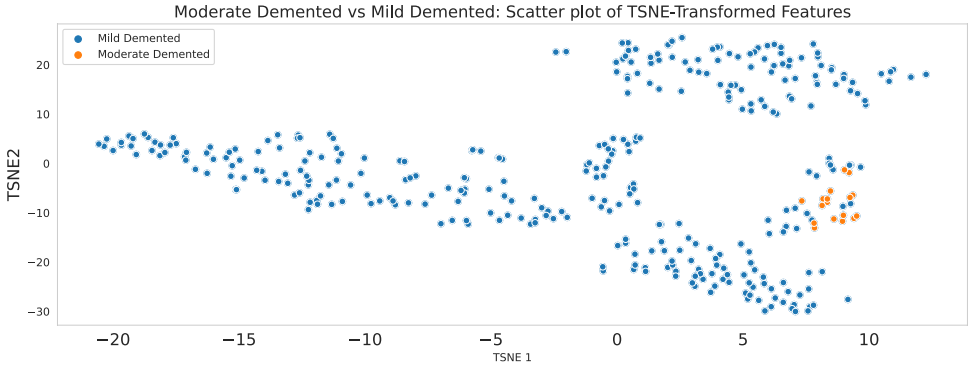


Fig. 26 Visualizing relation between TSNE transformed features after feature reduction (Moderate vs Mild)

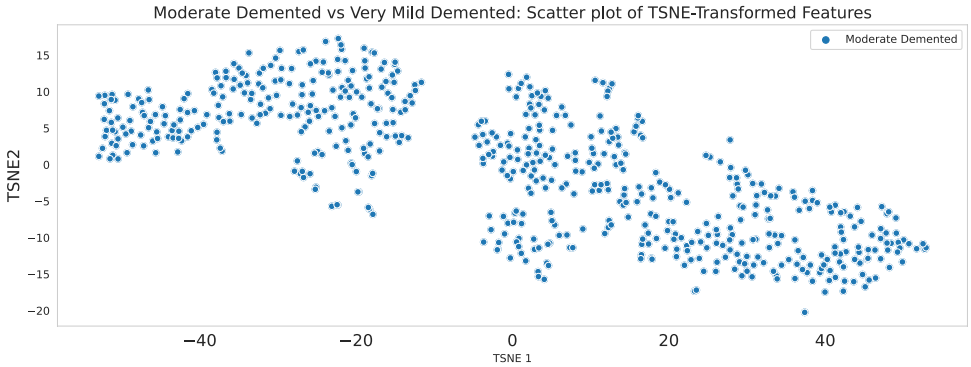


Fig. 27 Visualizing relation between TSNE transformed features after feature reduction (Moderate vs Very-Mild)

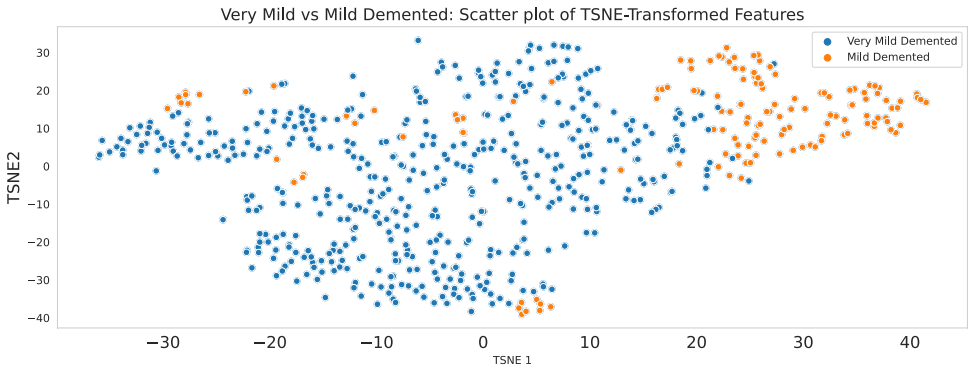


Fig. 28 Visualizing relation between TSNE transformed features after feature reduction (Very-Mild vs Mild)

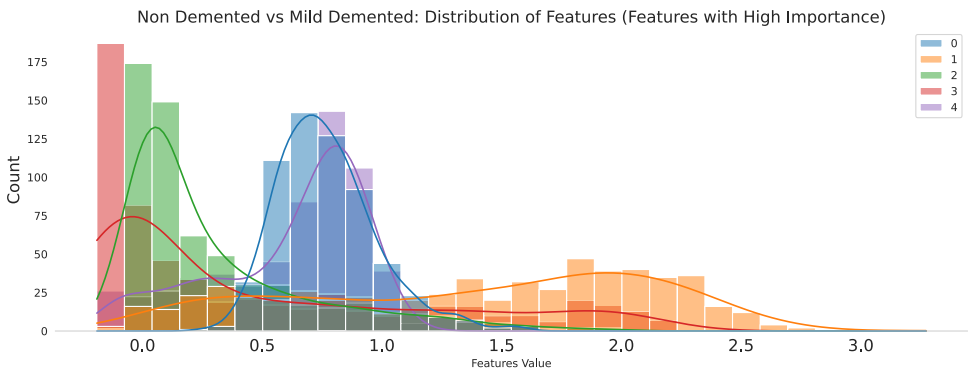


Fig. 29 Feature distribution visualization (Non vs Mild)

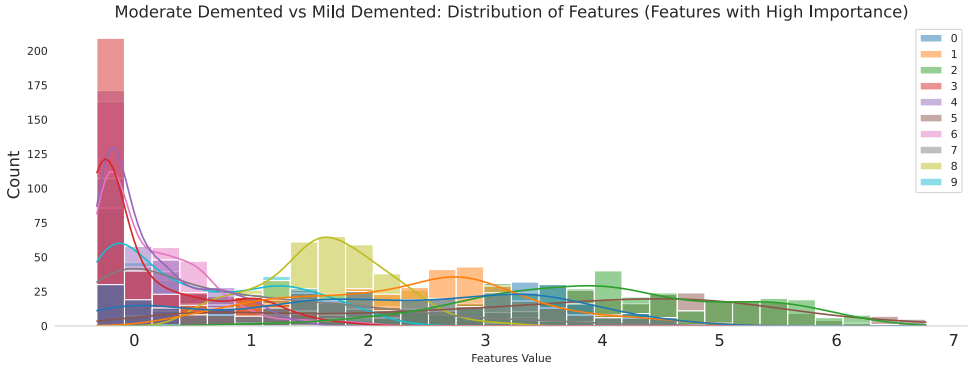


Fig. 30 Feature distribution visualization of only highly important features (Moderate vs Mild)

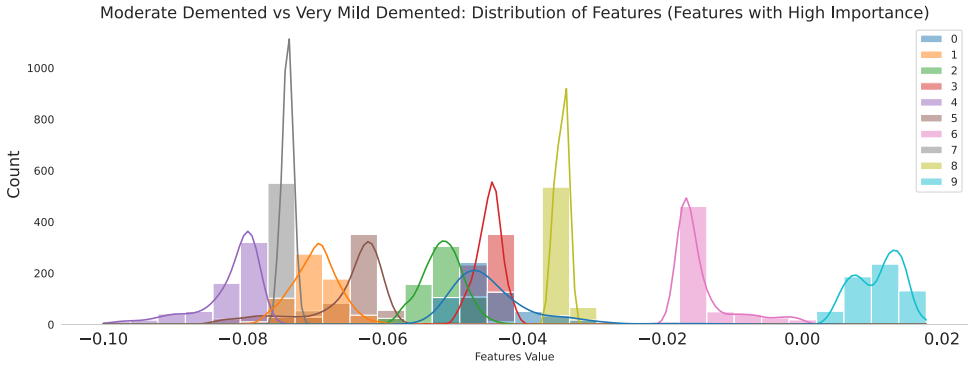


Fig. 31 Feature distribution visualization of only highly important features (Moderate vs Very-Mild)

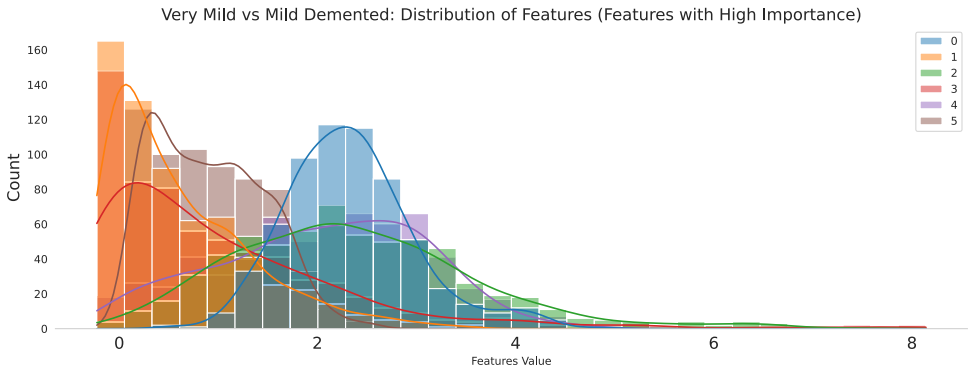


Fig. 32 Feature distribution visualization of only highly important features (Very-Mild vs Mild)

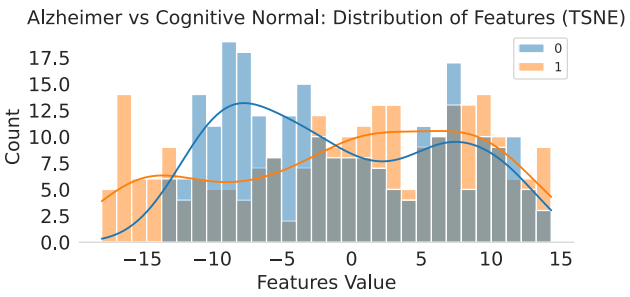


Fig. 33 Feature distribution visualization after feature reduction using TSNE algorithm (Alzheimer vs Normal)

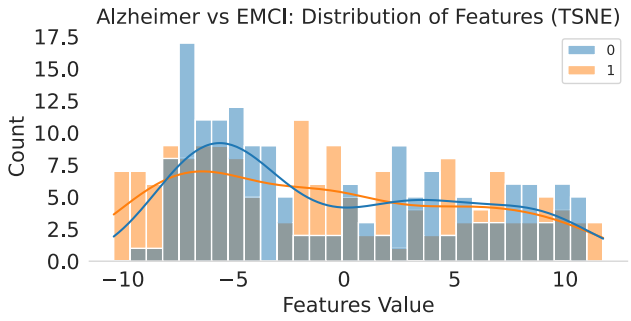


Fig. 34 Feature distribution visualization after feature reduction using TSNE algorithm (Alzheimer vs EMCI)

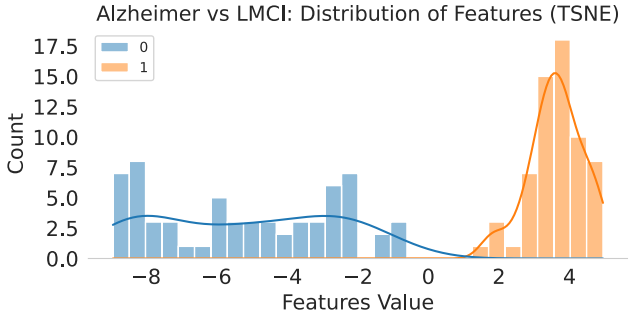


Fig. 35 Feature distribution visualization after feature reduction using TSNE algorithm (Alzheimer vs LMCI)

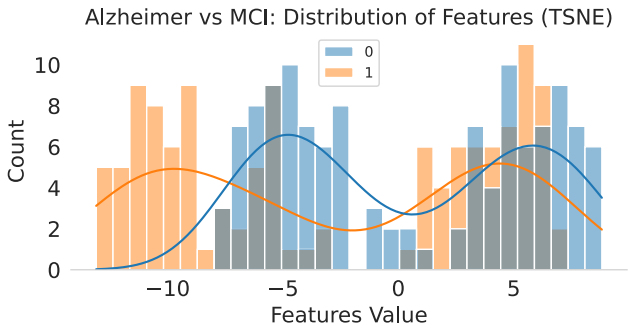


Fig. 36 Feature distribution visualization after feature reduction using TSNE algorithm (Alzheimer vs MCI)

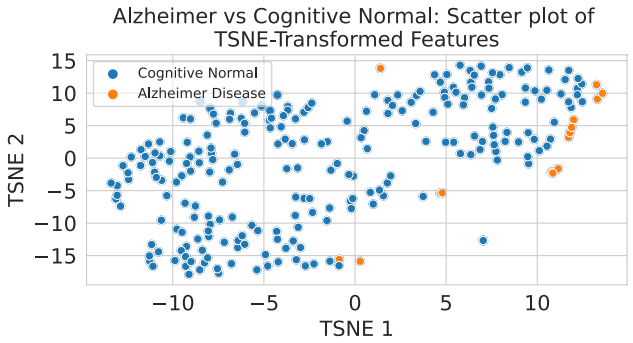


Fig. 37 Visualizing relation between TSNE transformed features after feature reduction (Alzheimer vs Normal)

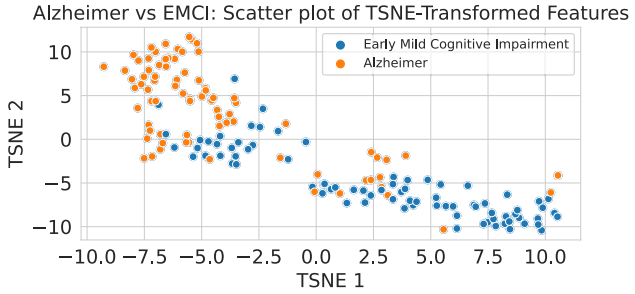


Fig. 38 Visualizing relation between TSNE transformed features after feature reduction (Alzheimer vs EMCI)

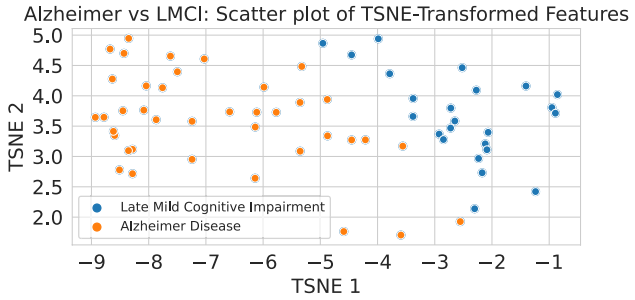


Fig. 39 Visualizing relation between TSNE transformed features after feature reduction (Alzheimer vs LMCI)

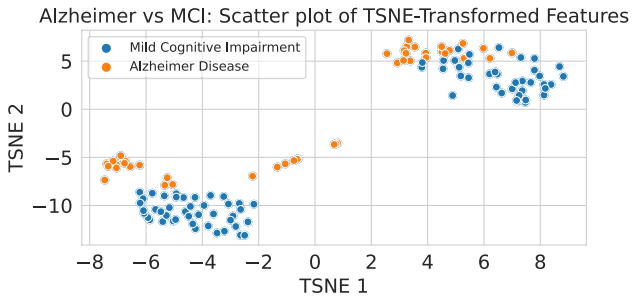


Fig. 40 Visualizing relation between TSNE transformed features after feature reduction (Alzheimer vs MCI)

reduced-dimensional space. Each bin in the histogram represents the number of data points that fall within a specific range of values along the axes of the t-SNE plot. In this case, a peak of the histogram highlights the high density of data points, which is clustered together in the original high-dimensional space. Whereas, peaks or high points of the histogram denote areas with higher data concentration, low points indicate the low density zones. Whether it's the distribution of data points, or the correlation with their original features, the analysis of the tangled features involves the process of examining how each point correlates with the other. Data points that have close relationships or are highly distingushing are bound to be clustered based on their location on the t-SNE plot. Through the histogram comparison with the standard features, you can get an understanding of which ones are considered as the most ordered or significant for the data distributing or classifying task.

Following that, the generate of a scatter plot of the TSNE-transformed features helps to understand the relationships and gain the insight into the presence of clusters within the data that tells how the data is grouped in general. The relationship between the features after undergoing feature reduction with TSNE, specifically in the context of Alzheimer stages, is illustrated in Figs. 25, 26, 27, and 28. We

further analyze feature importance using Lasso regression, a method that helps identify the most influential features. By plotting the histogram of the top N important features, we gain insights into their distribution. Figures 29, 30, 31, and 32 illustrate the distribution of the most significant features in AD as determined by Lasso Regression.

Lastly, we display the TSNE-transformed features using histograms and the relationship plots within the context of the cognitive disorder. Figures 33, 34, 35 and 36 demonstrate that TSNE features are being used in the classification of cognitive disorders. Further, as depicted in the Figs. 37, 38, 39, and 40, the relationship between these features is plotted after TSNE reduces dimensionality. TSNE transformation is crucial for visualizing feature relationships in scatter plots, particularly when dealing with high-dimensional data. It projects features into two dimensions, enhancing interpretability and facilitating a clearer understanding of feature interactions.

Acknowledgements The authors would like to express their grateful to Edinburgh Napier University and the Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2024R104), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia

Funding This research was funded by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2024R104), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Availability of data and materials In this study, two datasets were used for classification of AD. The first dataset was obtained from [https://www.kaggle.com/datasets/tourist55/alzheimers-dataset-4-class-of-images] and consists of AD images. The second dataset includes multiple cognitive disorders, including AD, and was obtained from [https://www.kaggle.com/datasets/madhucharan/alzheimersdisease5classdatasetadni]. The sources cited as [54] and [18] offer different information about the two datasets being discussed. According to [54], the main objective of creating the Alzheimer's dataset was to design a precise framework or architecture for AD classification. On the other hand, [18] obtained the cognitive disorders dataset from the ADNI website and acknowledged the ADNI for its creation. The dataset consists of patients data that aims to accelerate Alzheimer's research.

Code availability The source code for the experiments conducted in this research is available at the following GitHub repository: [https://github.com/Hassanshah531/Computer-Aided-Diagnosis-of-Alzheimers-Disease-cognitive-Disorders].

Declarations

Conflict of interest The authors declare no Conflict of interest associated with this work.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

1. Acharya H, Mehta R, Singh DK (2021) Alzheimer disease classification using transfer learning. In: 2021 5th international conference on computing methodologies and communication (ICCMC). IEEE, pp 1503–1508
2. ADNI (2017) Adni | alzheimer's disease neuroimaging initiative. https://adni.loni.usc.edu/
3. Ahmad MF, Akbar S, Hassan SAE, Rehman A, Ayesha N (2021) Deep learning approach to diagnose alzheimer's disease through magnetic resonance images. In: 2021 international conference on innovative computing (ICIC). IEEE, pp 1–6
4. Almufareh MF, Tehsin S, Humayun M, Kausar S (2023) Artificial cognition for detection of mental disability: a vision transformer approach for Alzheimer's disease. In: Healthcare, MDPI, p 2763
5. Alshammari M, Mezher M (2021) A modified convolutional neural networks for mri-based images for detection and stage classification of alzheimer disease. In: 2021 National computing colleges conference (NCCC). IEEE, pp 1–7
6. Amini M, Pedram M, Moradi A, Ouchani M (2021) Diagnosis of Alzheimer's disease severity with FMRI images using robust multitask feature extraction method and convolutional neural network (CNN). Comput Math Methods Med 2021:1–15
7. An N, Ding H, Yang J, Au R, Ang TF (2020) Deep ensemble learning for Alzheimer's disease classification. J Biomed Inform 105:103411
8. Arevalo-Rodriguez I, Smailagic N, i Figuls MR, Ciapponi A, Sanchez-Perez E, Giannakou A, Pedraza OL, Cosp XB, Cullum S (2015) Mini-mental state examination (mmse) for the detection of alzheimer's disease and other dementias in people with mild cognitive impairment (MCI). Cochrane database of systematic reviews
9. Association A et al (2009) 2009 Alzheimer's disease facts and figures. Alzheimer's Dementia 5:234–270
10. Association A et al (2010) 2010 Alzheimer's disease facts and figures. Alzheimer's Dementia 6:158–194
11. Association A et al (2013) 2013 Alzheimer's disease facts and figures. Alzheimer's Dementia 9:208–245
12. Association A et al (2014) 2014 Alzheimer's disease facts and figures. Alzheimer's Dementia 10:e47–e92
13. Ballard C, Gauthier S, Corbett A, Brayne C, Aarsland D, Jones E (2011) Alzheimer's disease. The Lancet 377:1019–1031
14. Bank D, Koenigstein N, Giryes R (2020) Autoencoders. arXiv preprint arXiv:2003.05991
15. Baydargil HB, Park J, Ince IF (2024) Anomaly-based alzheimer's disease detection using entropy-based probability positron emission tomography images. ETRI J 46(3):513–525

16. Beheshti I, Demirel H, Matsuda H, Initiative ADN et al (2017) Classification of Alzheimer's disease and prediction of mild cognitive impairment-to-Alzheimer's conversion from structural magnetic resource imaging using feature ranking and a genetic algorithm. Comput Biol Med 83:109–119
17. Blauwendraat C, Nalls MA, Singleton AB (2020) The genetic architecture of Parkinson's disease. Lancet Neurol 19:170–178
18. Charan M (2021) Alzheimers-disease-5-class-dataset-adni. https://www.kaggle.com/datasets/madhucharan/alzheimersdisease5classdatasetadni
19. Chollet F (2017) Xception: deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1251–1258
20. Cockrell JR, Folstein MF (2002) Mini-mental state examination. Principles and practice of geriatric psychiatry, pp 140–141
21. Daliri MR (2012) Automated diagnosis of Alzheimer disease using the scale-invariant feature transforms in magnetic resonance images. J Med Syst 36:995–1000
22. Davatzikos C, Fan Y, Wu X, Shen D, Resnick SM (2008) Detection of prodromal Alzheimer's disease via pattern classification of magnetic resonance imaging. Neurobiol Aging 29:514–523
23. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. IEEE, pp 248–255
24. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al (2020) An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929
25. Drewitt A (2023) An approach to classify Alzheimer's disease using vision transformers. Ph.D. thesis. Dublin, National College of Ireland
26. Dubey S (2020) Augmented Alzheimer MRI dataset. https://www.kaggle.com/datasets/tourist55/alzheimers-dataset-4-class-of-images
27. Duyckaerts C, Delatour B, Potier MC (2009) Classification and basic pathology of Alzheimer disease. Acta Neuropathol 118:5–36
28. Ebrahimi-Ghahnavieh A, Luo S, Chiong R (2019) Transfer learning for Alzheimer's disease detection on mri images. In: 2019 IEEE international conference on industry 4.0, artificial intelligence, and communications technology (IAICT). IEEE, pp 133–138
29. Ebrahimighahnavieh MA, Luo S, Chiong R (2020) Deep learning to detect Alzheimer's disease from neuroimaging: a systematic literature review. Comput Methods Programs Biomed 187:105242
30. Ebrahimighahnavieh MA, Luo S, Chiong R (2020) Deep learning to detect Alzheimer's disease from neuroimaging: a systematic literature review. Comput Methods Programs Biomed 187:105242
31. Ferrarini L, Frisoni GB, Pievani M, Reiber JH, Ganzola R, Milles J (2009) Morphological hippocampal markers for automated detection of Alzheimer's disease and mild cognitive impairment converters in magnetic resonance images. J Alzheimers Dis 17:643–659
32. Folstein M, Anthony JC, Parhad I, Duffy B, Gruenberg EM (1985) The meaning of cognitive impairment in the elderly. J Am Geriatr Soc 33:228–235
33. Galasko D, Klauber MR, Hofstetter CR, Salmon DP, Lasker B, Thal LJ (1990) The mini-mental state examination in the early diagnosis of Alzheimer's disease. Arch Neurol 47:49–52
34. Gauthier S, Reisberg B, Zaudig M, Petersen RC, Ritchie K, Broich K, Belleville S, Brodaty H, Bennett D, Chertkow H et al (2006) Mild cognitive impairment. The Lancet 367:1262–1270
35. Gauthier S, Reisberg B, Zaudig M, Petersen RC, Ritchie K, Broich K, Belleville S, Brodaty H, Bennett D, Chertkow H et al (2006) Mild cognitive impairment. The Lancet 367:1262–1270
36. Ghazal TM, Issa G (2022) Alzheimer disease detection empowered with transfer learning. Comput Mater Continua 70:5005–5019
37. Gooblar J, Roe CM, Selsor NJ, Gabel MJ, Morris JC (2015) Attitudes of research participants and the general public regarding disclosure of Alzheimer disease research results. JAMA Neurol 72:1484–1490
38. Hassani A, Walton S, Shah N, Abuduweili A, Li J, Shi H (2021) Escaping the big data paradigm with compact transformers. arXiv preprint arXiv:2104.05704
39. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
40. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H (2017) Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861
41. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708
42. Jessen F (2014) Subjective and objective cognitive decline at the pre-dementia stage of Alzheimer's disease. Eur Arch Psychiatry Clin Neurosci 264:3–7
43. Jo T, Nho K, Saykin AJ (2019) Deep learning in Alzheimer's disease: diagnostic classification and prognostic prediction using neuroimaging data. Front Aging Neurosci 11:220
44. Julayanont P, Nasreddine ZS (2017) Montreal cognitive assessment (moca): concept and clinical review. A practical approach, Cognitive screening instruments, pp 139–195
45. Julayanont P, Nasreddine ZS (2017) Montreal cognitive assessment (MoCA): concept and clinical review. A practical approach, Cognitive screening instruments, pp 139–195
46. Kabir A, Kabir F, Mahmud MAH, Sinthia SA, Azam SR, Hussain E, Parvez MZ (2021) Multi-classification based Alzheimer's disease detection with comparative analysis from brain MRI scans using deep learning. In: TENCON 2021-2021 IEEE region 10 conference (TENCON). IEEE, pp 905–910
47. Kang W, Lin L, Zhang B, Shen X, Wu S, Initiative ADN et al (2021) Multi-model and multi-slice ensemble learning architecture based on 2d convolutional neural networks for alzheimer's disease diagnosis. Comput Biol Med 136:104678
48. Katzman R (1989) Alzheimer's disease is a degenerative disorder. Neurobiol Aging 10:581–582
49. Kim K, Wu B, Dai X, Zhang P, Yan Z, Vajda P, Kim SJ (2021) Rethinking the self-attention in vision transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3071–3075
50. Kingma DP, Welling M et al (2019) An introduction to variational autoencoders. Found Trends Mach Learn 12:307–392
51. Knopman DS, Amieva H, Petersen RC, Chételat G, Holtzman DM, Hyman BT, Nixon RA, Jones DT (2021) Alzheimer disease. Nat Rev Dis Primers 7:33
52. Kora P, Ooi CP, Faust O, Raghavendra U, Gudigar A, Chan WY, Meenakshi K, Swaraja K, Plawiak P, Acharya UR (2022) Transfer learning techniques for medical image analysis: a review. Biocybern Biomed Eng 42:79–107
53. Korolev IO (2014) Alzheimer's disease: a clinical and basic science review. Med Student Res J 4:24–33
54. Kumar S, Shastri S (2022) Alzheimer mri preprocessed dataset. https://www.kaggle.com/dsv/3364939, https://doi.org/10.34740/KAGGLE/DSV/3364939

55. Kurlowicz L, Wallace M (1999) The mini-mental state examination (MMSE). J Gerontol Nurs 25(5):8–9
56. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z et al (2015) Imagenet large scale visual recognition challenge. Int J Comput Vision 115:211–252
57. Lazli L (2022) Machine learning classifiers based on dimensionality reduction techniques for the early diagnosis of Alzheimer's disease using magnetic resonance imaging and positron emission tomography brain data. Computational intelligence methods for bioinformatics and biostatistics: 17th international meeting, CIBB 2021, Virtual Event, November 15–17, 2021. Springer, Revised Selected Papers, pp 117–131
58. Lee ES, Yoo K, Lee YB, Chung J, Lim JE, Yoon B, Jeong Y (2016) Default mode network functional connectivity in early and late mild cognitive impairment. Alzheimer Disease Assoc Disord 30:289–296
59. Liu M, Li F, Yan H, Wang K, Ma Y, Shen L, Xu M, Initiative ADN et al (2020) A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in alzheimer's disease. Neuroimage 208:116459
60. Liu W, Wang Z, Liu X, Zeng N, Liu Y, Alsaadi FE (2017) A survey of deep neural network architectures and their applications. Neurocomputing 234:11–26
61. Liu Z, Lu H, Pan X, Xu M, Lan R, Luo X (2022) Diagnosis of Alzheimer's disease via an attention-based multi-scale convolutional neural network. Knowl-Based Syst 238:107942
62. Martin-Khan M, Flicker L, Wootton R, Loh PK, Edwards H, Varghese P, Byrne GJ, Klein K, Gray LC (2012) The diagnostic accuracy of telegeriatrics for the diagnosis of dementia via video conferencing. J Am Med Dir Assoc 13:487-e19
63. McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR Jr, Kawas CH, Klunk WE, Koroshetz WJ, Manly JJ, Mayeux R et al (2011) The diagnosis of dementia due to Alzheimer's disease: recommendations from the national institute on aging-Alzheimer's association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's Dementia 7:263–269
64. Mirzaei G, Adeli H (2022) Machine learning techniques for diagnosis of Alzheimer disease, mild cognitive disorder, and other types of dementia. Biomed Signal Process Control 72:103293
65. Mirzaei G, Adeli H (2022) Machine learning techniques for diagnosis of Alzheimer disease, mild cognitive disorder, and other types of dementia. Biomed Signal Process Control 72:103293
66. Nasreddine ZS, Phillips NA, Bédirian V, Charbonneau S, Whitehead V, Collin I, Cummings JL, Chertkow H (2005) The montreal cognitive assessment, MoCA: a brief screening tool for mild cognitive impairment. J Am Geriatr Soc 53:695–699
67. Newcombe EA, Camats-Perna J, Silva ML, Valmas N, Huat TJ, Medeiros R (2018) Inflammation: the link between comorbidities, genetics, and Alzheimer's disease. J Neuroinflammation 15:1–26
68. Ng A, et al (2011) Sparse autoencoder. CS294A Lecture Notes 72:1–19
69. Odusami M, Maskeliūnas R, Damaševičius R (2023) Pixel-level fusion approach with vision transformer for early detection of Alzheimer's disease. Electronics 12:1218
70. Organization WH (2023) Dementia. https://www.who.int/news-room/fact-sheets/detail/dementia
71. Pak M, Kim S (2017) A review of deep learning in image recognition. In: 2017 4th international conference on computer applications and information processing technology (CAIPT). IEEE, pp 1–3
72. Papazacharias A, Nardini M (2012) He relationship between depression and cognitive deficits. Psychiatr Danub 24:179–182
73. Petersen RC, Roberts RO, Knopman DS, Boeve BF, Geda YE, Ivnik RJ, Smith GE, Jack CR (2009) Mild cognitive impairment: ten years later. Arch Neurol 66:1447–1455
74. Raghavaiah P, Varadarajan S (2021) Novel deep learning convolution technique for recognition of Alzheimer's disease. Mater Today Proc 46:4095–4098
75. Rose VL (1998) Alzheimer's disease genetic fact sheet. Am Fam Physician 58:578
76. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC (2018) Mobilenetv2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4510–4520
77. Sarraf S, DeSouza DD, Anderson J, Tofighi G, Initiativ ADN (2016) Deepad: Alzheimer's disease classification via deep convolutional neural networks using mri and fmri. BioRxiv, 070441
78. Schnakers C, Monti MM (2020) Towards improving care for disorders of consciousness. Nat Rev Neurol 16:405–406
79. Sethi M, Rani S, Singh A, Mazón JLV (2022) A cad system for Alzheimer's disease classification using neuroimaging MRI 2D slices. Comput Math Methods Med 2022(1):8680737
80. Shamrat FJM, Akter S, Azam S, Karim A, Ghosh P, Tasnim Z, Hasib KM, De Boer F, Ahmed K (2023) Alzheimernet: An effective deep learning based proposition for Alzheimer's disease stages classification from functional brain changes in magnetic resonance images. IEEE Access 11:16376–16395
81. Shen X, Finn ES, Scheinost D, Rosenberg MD, Chun MM, Papademetris X, Constable RT (2017) Using connectome-based predictive modeling to predict individual behavior from brain connectivity. Nature Protocols 12:506–518
82. Sisodia PS, Ameta GK, Kumar Y, Chaplot N (2023) A review of deep transfer learning approaches for class-wise prediction of Alzheimer's disease using MRI images. Arch Comput Methods Eng, pp. 1–21
83. Sona A, Ellis KA, Ames D (2013) Rapid cognitive decline in Alzheimer's disease: a literature review. Int Rev Psychiatry 25:650–658
84. Sorour SE, Abd El-Mageed AA, Albarrak KM, Alnaim AK, Wafa AA, El-Shafeiy E (2024) Classification of alzheimer's disease using MRI data based on deep learning techniques. J King Saud Univ-Comput Inf Sci 101940
85. Sun QS, Zeng SG, Liu Y, Heng PA, Xia DS (2005) A new method of feature fusion and its application in image recognition. Pattern Recogn 38:2437–2448
86. Szegedy C, Ioffe S, Vanhoucke V, Alemi A (2017) Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proceedings of the AAAI conference on artificial intelligence
87. Taheri Gorji H, Kaabouch N (2019) A deep learning approach for diagnosis of mild cognitive impairment based on MRI images. Brain Sci 9:217
88. Takemori Y, Sasayama D, Toida Y, Kotagiri M, Sugiyama N, Yamaguchi M, Washizuka S, Honda H (2021) Possible utilization of salivary ifn-γ/il-4 ratio as a marker of chronic stress in healthy individuals. Neuropsychopharmacol Rep 41:65–72
89. Tan CC, Eswaran C (2008) Performance comparison of three types of autoencoder neural networks. In: 2008 second asia international conference on modelling & simulation (AMS). IEEE, pp 213–218
90. Tan M, Le Q (2019) Efficientnet: rethinking model scaling for convolutional neural networks. In: International conference on machine learning. PMLR, pp 6105–6114
91. Tang W, Sun J, Wang S, Zhang Y (2023) Review of alexnet for medical image classification. arXiv preprint arXiv:2311.08655
92. Tang-Wai DF, Knopman DS, Geda YE, Edland SD, Smith GE, Ivnik RJ, Tangalos EG, Boeve BF, Petersen RC (2003) Comparison of the short test of mental status and the mini-mental state examination in mild cognitive impairment. Arch Neurol 60:1777–1781

93. Tanveer M, Richhariya B, Khan RU, Rashid AH, Khanna P, Prasad M, Lin C (2020) Machine learning techniques for the diagnosis of Alzheimer's disease: a review. ACM Trans Multimed Comput Commun Appl 16:1–35
94. Tsoi KK, Chan JY, Hirai HW, Wong SY, Kwok TC (2015) Cognitive tests to detect dementia: a systematic review and meta-analysis. JAMA Intern Med 175:1450–1458
95. Uraninjo (2022) Alzheimer's dataset (4 class of images). https://www.kaggle.com/datasets/uraninjo/augmented-alzheimer-mri-dataset
96. Venugopalan J, Tong L, Hassanzadeh HR, Wang MD (2021) Multimodal deep learning models for early detection of Alzheimer's disease stage. Sci Rep 11:1–13
97. de Vico Fallani F, Richiardi J, Chavez M, Achard S (2014) Graph analysis of functional brain networks: practical issues in translational neuroscience. Philos Trans R Soc B Biol Sci 369:20130521
98. Wang J, Zhu H, Wang SH, Zhang YD (2021) A review of deep learning on medical image analysis. Mob Netw Appl 26:351–380
99. Yildirim M, Cinar A (2020) Classification of Alzheimer's disease MRI images with CNN based hybrid method. Ingénierie des Systèmes d Inf. 25:413–418
100. Yin Y, Jin W, Bai J, Liu R, Zhen H (2022) Smil-deit: multiple instance learning and self-supervised vision transformer network for early Alzheimer's disease classification. In: 2022 international joint conference on neural networks (IJCNN). IEEE, pp 1–6
101. Zhang Y (2018) A better autoencoder for image: convolutional autoencoder. In: ICONIP17-DCEC. Available online: http://users.cecs.anu.edu.au/Tom.Gedeon/conf/ABCs2018/paper/ABCs2018_paper_58. Accessed 23 Mar 2017