Computers in Biology and Medicine 167 (2023) 107667

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Do it the transformer way: A comprehensive review of brain and vision transformers for autism spectrum disorder diagnosis and classification

Asrar G. Alharthi *, Salha M. Alzahrani

Department of Computer Science, College of Computers and Information Technology, Taif University, Saudi Arabia

ARTICLE INFO

INDEX TERMS:
Autism spectrum disorder
ASD
MRI
fMRI
sMRI
Neuroimaging
Classification
Deep learning
Transfer learning
Vision transformers
Brain transformers

ABSTRACT

Autism spectrum disorder (ASD) is a condition observed in children who display abnormal patterns of interaction, behavior, and communication with others. Despite extensive research efforts, the underlying causes of this neurodevelopmental disorder and its biomarkers remain unknown. However, advancements in artificial intelligence and machine learning have improved clinicians' ability to diagnose ASD. This review paper investigates various MRI modalities to identify distinct features that characterize individuals with ASD compared to typical control subjects. The review then moves on to explore deep learning models for ASD diagnosis, including convolutional neural networks (CNNs), autoencoders, graph convolutions, attention networks, and other models. CNNs and their variations are particularly effective due to their capacity to learn structured image representations and identify reliable biomarkers for brain disorders. Computer vision transformers often employ CNN architectures with transfer learning techniques like fine-tuning and layer freezing to enhance image classification performance, surpassing traditional machine learning models. This review paper contributes in three main ways. Firstly, it provides a comprehensive overview of a recommended architecture for using vision transformers in the systematic ASD diagnostic process. To this end, the paper investigates various pre-trained vision architectures such as VGG, ResNet, Inception, InceptionResNet, DenseNet, and Swin models that were fine-tuned for ASD diagnosis and classification. Secondly, it discusses the vision transformers of 2020th like BiT, ViT, MobileViT, and ConvNeXt, and applying transfer learning methods in relation to their prospective practicality in ASD classification. Thirdly, it explores brain transformers that are pre-trained on medically rich data and MRI neuroimaging datasets. The paper recommends a systematic architecture for ASD diagnosis using brain transformers. It also reviews recently developed brain transformer-based models, such as METAFormer, Com-BrainTF, Brain Network, ST-Transformer, STCAL, BolT, and BrainFormer, discussing their deep transfer learning architectures and results in ASD detection. Additionally, the paper summarizes and discusses brain-related transformers for various brain disorders, such as MSGTN, STAGIN, and MedTransformer, in relation to their potential usefulness in ASD. The study suggests that developing specialized transformer-based models, following the success of natural language processing (NLP), can offer new directions for image classification problems in ASD brain biomarkers learning and classification. By incorporating the attention mechanism, treating MRI modalities as sequence prediction tasks trained on brain disorder classification problems, and fine-tuned on ASD datasets, brain transformers can show a great promise in ASD diagnosis.

1. Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental disorder which is clinically diagnosed after several tests based on factors such as interaction, communication and behavior patterns [1]. Some common characteristics of ASD include difficulty with eye contact or understanding social cues, difficulty with verbal or nonverbal communication, difficulty with attention and focus, and engaging in repetitive behaviors or routines such as flapping, rocking, or spinning [2–4]. Approximately, one in 100 children worldwide has been diagnosed with ASD in 2022 according to the World Health Organization (WHO).[1] Early diagnosis of ASD leads to effective interventions and can significantly improve the social skills of people having ASD [3,5]. At present, three commonly used dependable diagnostic methods for identifying autism are Autism

* Corresponding author.
E-mail address: asrar.g.alharthi@gmail.com (A.G. Alharthi).
[1] https://www.who.int/.

https://doi.org/10.1016/j.compbiomed.2023.107667
Received 17 June 2023; Received in revised form 25 October 2023; Accepted 31 October 2023
Available online 3 November 2023

Diagnostic Observation Schedule (ADOS), Autism Diagnostic Interview-Revised (ADI-R), and Diagnostic and Statistical Manual of Mental Disorders 5 (DSM-5) [4]. The DSM-5 distinguishes between three levels of autism severity, categorized according to deficiencies in social communication and repetitive behavior [3]. Other diagnostic tools typically involve a parental interview, growth history, and some medical evaluation to rule out any underlying medical conditions that may contribute to the ASD's symptoms. It is worth mentioning that not all individuals with ASD should have entirely these symptoms as the severity of symptoms can vary widely. Other conditions like attention deficit hyperactivity disorder (ADHD), anxiety, and intellectual disability may coexist in some individuals with ASD [2,3]. Due to this variability in symptoms, it is difficult to establish clear diagnostic criteria that apply to all individuals having ASD. Additionally, the criteria for diagnosing ASD have changed over the years, and there are still some disagreements among experts about the best way to diagnose the condition. Furthermore, these diagnostic tools involve long questionnaires which are not only time-consuming and expensive but also have the limitation when children cannot communicate [6]. Therefore, to ensure patients with ASD receive immediate intervention, it is urgent to develop automated diagnosis methods for earlier identification. More importantly, automated methods can lead to identifying precise brain biomarkers to assist physicians in diagnosis.

The availability of non-invasive brain imaging techniques has enabled a better understanding of the neural circuitry underlying neurological disorders. Quantitative analysis of brain imaging data has the potential to provide valuable biomarkers that can improve the accuracy of brain disorder diagnoses. Currently, magnetic resonance imaging (MRI) can assist to understand neurological disorders and their associated symptoms, such as schizophrenia, ASD, and Alzheimer's diseases [7]. Although MRI modalities provide high-resolution imaging capabilities for detecting smaller anatomical structures and abnormalities, the growing number of images per patient presents challenges. The time required for clinicians to interpret these images can be significant, resulting in increased workload and potentially diminishing their diagnostic capabilities [8]. Moreover, MRI scans utilize various protocols and generate multiple slices, necessitating clinicians to meticulously analyze each individual slice to ensure a comprehensive examination [9]. The fatigue and lack of experience among many clinicians when it comes to interpreting MRI slices can contribute to difficulties in diagnosing ASD in its early stages.

Research has demonstrated that combining artificial intelligence (AI) with MRI neuroimaging modalities can improve the accuracy of ASD diagnosis. At present, many studies have utilized machine learning (ML) models to identify patterns and abnormalities from MRI modalities for ASD diagnosis [10–12]. After identifying ASD patterns, models are inferred from the training set and applied to unseen data. Researchers have found differences in the brain's activity of ASD subjects compared to control (i.e., healthy) subjects. Modern advances in this research area have led subjects with ASD to receive immediate intervention requiring less diagnostic time than standard clinicians' diagnostics [13]. However, traditional ML models usually require handcrafted feature extraction and engineering. The success of ML models is dependent on choosing the appropriate features which usually require extensive domain knowledge and experience. In addition, analyzing MRI neuroimaging modalities for brain disorders is a difficult research task due to several factors such as the complex structure, non-linear separability, high dimensionality of data, and the sequential changes of traceable signals in every voxel [14].

Recent advances in deep learning (DL) research have shown impressive abilities and performance that outperform traditional ML models on medical classification tasks. Numerous researchers within the ASD community have been motivated to utilize DL models for the purpose of diagnosing and classifying ASD [15–19]. DL models have the advantage of automatically learning hidden representations from raw datasets, making them adaptable to new data and less prone to overfitting. However, the practicality of DL models in the field of ASD diagnosis is limited by the difficulty and expense of obtaining large neuroimaging datasets. Despite this, DL advancements have primarily focused on CNNs, which excel at extracting features but struggle to encode the relative positions of these features, resulting in a loss of global context. To address this limitation, researchers have proposed various architectural changes, including the integration of attention mechanisms [20]. These attention mechanisms allow models to focus on specific regions of an image, leading to improved performance. Inspired by the success of transformer networks in natural language processing (NLP), researchers have developed vision transformers such as VGG [21,22], ResNet [23,24], Xception [25], Inception [26,27], DenseNet [28], MobileNet [29,30], NASNet [31], EfficientNet [32,33], and large transformer-based models such as ViT [34], Swin [35], MobileViT [36], BiT [37], and ConvNeXt [38] transformers. These vision transformers leverage attention mechanisms to capture global context and extract detailed features from image patches. As a result, vision transformers offer a promising alternative to CNNs, overcoming their limitations in encoding feature positions.

This comprehensive review paper primarily focuses on the recent advancements in utilizing vision and brain transformers for transfer learning in the field of ASD diagnosis. The main objective is to emphasize the potential of these models in harnessing pre-trained knowledge from related tasks in order to enhance ASD diagnosis, as opposed to starting model training from scratch [39]. This approach has the potential to reduce the data requirements for training and expedite the training time for ASD models. Traditional research has primarily focused on extracting features from MRI modalities and utilizing traditional ML techniques for ASD diagnosis. However, there has been limited exploration of using raw MRI images as input for deep learning models, particularly CNNs. Moreover, the studies that have investigated this approach have typically been conducted on small datasets [19,40–42]. Few research works have utilized transfer learning using recent pre-trained vision and brain transformers for ASD diagnosis. Recent advances in research have initiated this direction by using VGG [1,19,43] and ResNet [42,44] for ASD diagnosis. By examining the existing research on diagnosing ASD, this review paper offers valuable insights into the advantages and obstacles related to vision and brain transformers. It presents new avenues for research and provides guidelines, emphasizing the potential of vision and brain transformers in diagnosing ASD using MRI. This review serves as a valuable resource for researchers and practitioners, shaping the future trends in MRI-based ASD diagnosis.

1.1. Objective of our review

The purpose of this review paper is to analyze the progress, methods, and challenges associated with using MRI modalities data for diagnosing Autism Spectrum Disorder (ASD), and how artificial intelligence (AI) techniques, specifically vision and brain transformers, can improve the accuracy and efficiency of the diagnostic process. We also conducted a thorough examination of ASD diagnosis using MRI techniques and explored the potential of using these images as input for vision and brain transformers. To achieve this goal, we accomplished an extensive search across various conferences and journals spanning from 2020 to 2023. The chosen papers were carefully reviewed to understand their methodologies and findings in relation to ASD diagnosis, detection, and classification. This comprehensive review aims to bridge the gap between the computer vision community and medical experts, encouraging collaboration and facilitating future research and advancements in the field of medical computer vision.

1.2. Comparison with other review papers

Although previous reviews have extensively covered the topic of using MRI modalities for ASD diagnosis, we believe there is still room for improvement. Specifically, there is a noticeable absence of reviews that specifically examine the applications of vision transformers and

pretrained transformer-based models in enhancing the efficiency of ASD diagnosis. To bridge this gap, we have conducted a survey that specifically focuses on analyzing the use of transfer learning using vision transformers and transformer-based architectures in ASD diagnosis. To elaborate, this survey discusses a thorough review of recent articles that have applied AI models for diagnosing ASD. To provide a comprehensive understanding of the field, we first provided a brief overview of studies that utilized both machine learning (ML) and deep learning (DL) models for ASD diagnosis. We then delved into a more detailed analysis, specifically focusing on studies that employed vision transformers and brain transformers for ASD diagnosis. Our analysis primarily focused on the usage of MRI modalities as input for ASD diagnosis, and we explored emerging trends in this area. Additionally, we summarized the various methods used by researchers to represent MRI data when incorporating them into different AI models. Our goal was to gain insights into how these modalities can be effectively utilized in vision and brain transformers for ASD diagnosis. Furthermore, we presented a concise overview of available MRI datasets for ASD and non-ASD subjects. To facilitate comparison, we organized the state-of-the-art methods in a tabular format, highlighting the performance metrics and outcomes from the available datasets. Lastly, we identified several challenges and proposed potential future directions for further advancements in the field of ASD diagnosis using MRI modalities. The comparison between our review and other published review articles dedicated to ASD diagnosis using MRI modalities and AI techniques is presented in Table 1.

In their studies [9], conducted a review of 233 articles that explored the use of DL and ML models along with MRI modalities for ASD diagnosis. The primary focus of their research was on crucial preprocessing techniques, feature extraction, feature selection, and ML models employed in ASD diagnosis. The authors highlighted the significant computational challenges associated with 3D DL models, which have limited extensive research in this field. Study [45] conducted a thorough review of 35 articles to provide a comprehensive survey of the utilization of AI techniques in ASD diagnosis using structural MRI (sMRI)/-functional MRI (fMRI) scans. Their review covers various topics including the application of general radiomic features and classifier models for diagnosing ASD. The authors also proposed future directions for research, emphasizing the need to develop models that quantify deep texture from CNN models in order to accurately classify subjects with ASD and those without ASD. Study [46] conducted a review of 59 articles that examine the use of DL techniques and neuroimaging data for ASD diagnosis. Their review paper thoroughly explores the different types of DL networks employed in ASD diagnosis and emphasizes the challenges associated with implementing these models. To enhance future research, the authors suggested incorporating handcrafted features alongside raw data in DL networks. In a separate study [47], 47 articles were reviewed that utilized AI models to analyze fMRI modalities for ASD diagnosis. Their review covers various aspects such as constructing features from raw fMRI data, selecting discriminant brain regions, identifying networks, and examining factors contributing to high classification accuracy. Study [48] examined 46 articles to investigate ML research conducted on ASD diagnosis using three different imaging techniques: sMRI, fMRI, and hybrid imaging. Their findings indicate that by combining behavioral data obtained through tracking systems with brain MRI data, a more dependable and precise diagnosis of ASD can be attained. Another work [49] conducted a review of 45 articles that employed AI models for ASD diagnosis using sMRI modalities. The authors' main objective was to explore the sMRI-based biomarkers associated with ASD and the learning models used for extracting and analyzing these data. They highlighted the advantage of using sMRI images, as they require less time and effort from both patients and clinicians compared to other MRI methods like fMRI. However, they also noted that expanding AI model architectures from 2D to 3D/4D can lead to increased parameters and runtime, which poses limitations in recognizing psychological indicators of ASD. To end, study [50] conducted a review of 119 articles that focused on using AI models for diagnosing ASD based on neuroimaging. The researchers started by employing feature extraction and feature selection methods specifically for MRI modalities. They then provided a detailed analysis of the training and evaluation of classification models. The study highlighted the importance of exploring the capabilities of deep learning in ASD diagnosis. However, the researchers also acknowledged significant limitations that need to be addressed, such as small sample sizes, difficulties in interpreting results, and concerns about the quality of the data. This review paper, on the other hand, not only covers more recent gaps of years but also presents new research directions of using vision transformers (VT) and brain transformers (BT).

Table 1
Comparison between our review paper and the published review papers that discuss AI models with MRI modalities for ASD diagnosis and classification.

Review Paper	#Papers Reviewed	Years Covered	Data	Reviewed AI Techniques
[9]	233	2016–2022	sMRI/fMRI modalities	ML + DL
[45]	35	2018–2021	sMRI/fMRI modalities	ML + DL
[46]	59	2016–2020	sMRI/fMRI modalities	DL Only
[47]	47	2011–2021	fMRI Only	ML + DL
[48]	46	2009–2020	sMRI/fMRI modalities	ML + DL
[49]	45	2017–2022	sMRI Only	ML + DL
[50]	119	2010–2020	sMRI/fMRI modalities	ML + DL
Ours	78	2020–2023	sMRI/fMRI modalities	ML + DL + VT + BT

1.3. Organization of the paper

The rest of this paper is structured as follows: Section 2 describes the systematic approach used to select relevant papers for this review. Section 3 provides a detailed review of MRI modalities and various methods used prior to ASD diagnosis, including MRI acquisition, preprocessing, and data representation. Section 4 briefly discusses the transition from machine learning-based approaches in ASD diagnosis into deep learning-based methods, including Multi-Layer Perceptron (MLP), Convolutional Neural Network (CNN), Autoencoder (AE), Graph Convolutional Networks (GCN), Graph Attention Network (GAT), and other models. Section 5 focuses on transformer-based architectures, specifically vision transformers (VT), such as VGG, ResNet, Inception, Xception, DenseNet, MobileNet, NASNet, EfficientNet, Vision Transformer (ViT), Swin Transformer, MobileViT, BiT, and ConvNeXt. Section 6 explores Brain Transformers (BT), which are transfer learning models pre-trained on neuroimaging data for ASD diagnosis. BT architectures discussed in this review paper include METAFormer, Com-BrainTF, Brain Network, ST-Transformer, STCAL, BolT, and BrainFormer, along with their results in ASD detection. The paper also summarizes brain-related transformers for other brain disorders, such as MSGTN, STAGIN, and MedTransformer, and their potential usefulness in ASD. Sections 7 and 8 provide further discussions on traditional methods compared to VT and BT transformers, as well as challenges in this research area. Section 9 concludes the paper and outlines future directions for research on using brain and vision transformers in ASD diagnosis and classification.

2. Method

This section outlines the systematic approach utilized to choose appropriate papers for inclusion in this review. It establishes a strong basis for the subsequent analysis and discussion by ensuring the inclusion of pertinent studies and transparently presenting the methodology employed. By employing this systematic approach, the review aims to

offer a thorough and insightful evaluation of the present state and future prospects of AI-based ASD diagnosis.

2.1. Search strategy

To conduct a comprehensive literature search pertaining to ASD diagnosis using MRI modalities, a range of citation databases and search engines were employed. The search procedure involved the utilization of relevant keywords associated with ASD diagnosis, MRI modalities, and AI techniques. The databases that were explored encompassed IEEE, Wiley, Frontiers, ScienceDirect, SpringerLink, ArXiv, and Google Scholar. The search terms employed included "ASD classification", "ASD diagnosis", "ASD detection", "autism spectrum disorder", "fMRI", "sMRI", "vision transformer", "brain transformer", "MRI transformer", "pretrained ASD", "machine learning ASD", "deep learning ASD", "transfer learning ASD", "transformers ASD", and "ABIDE".

2.2. Inclusion and exclusion criteria

The papers selected for this review specifically focused on the diagnosis, detection, or classification of ASD using MRI modalities as the primary data source. These studies employed various approaches such as ML, DL, transfer learning, vision/brain transformers (VT/BT) for ASD diagnosis. It was also a requirement for the papers to be available in English. Papers that did not utilize MRI datasets or AI techniques for ASD diagnosis were excluded from this review. The initial search involved using specific keywords and databases, and the identified papers were screened based on their titles and abstracts to determine their relevance. The inclusion and exclusion criteria for ASD diagnosis are summarized in Table 2. The full texts of the remaining papers were then thoroughly examined to assess their relevance to the topic.

2.3. Search results

The total number of initial articles identified through the search strategy is 134 papers, along with the breakdown of articles screened at the title and abstract level. After applying the inclusion and exclusion criteria, 78 papers were finally selected and used in this review. These selected papers cover the period from 2020 to 2023, ensuring the inclusion of recent research. We summarize the findings of these papers, focusing on MRI data acquisitions, MRI data representation methods, ML, DL, VT, and BT models for ASD diagnosis. Fig. 1 presents a comprehensive overview of the reviewed articles, showcasing both the chronological distribution across the years and illustrating the percentage distribution of the different types of AI models used in the selected studies.

3. MRI modalities

Nowadays, numerous studies have applied AI models to brain disorder research such as schizophrenia [51], depression [52], Alzheimer's [41,53], and ASD [3,4,10,12,15,54]. In recent years, the prevalence of ASD has increased among neuro-developmental disorders, posing challenges for clinicians in making accurate diagnoses. The utilization of MRI imaging methods has proven valuable in understanding the neural correlates of ASD and aiding in comprehensive brain examination. The application of AI models has further facilitated the automatic generation of ASD biomarkers, allowing for estimation of diagnosis and prognosis without the need for time-consuming manual annotation of MRI data. This paper focuses on two commonly employed MRI techniques, namely structural MRI (sMRI) and functional MRI (fMRI), in ASD diagnosis research. Fig. 2 shows the differences between sMRI and fMRI imaging techniques. These modalities capture different aspects of brain activity and structure. To analyze irregular neuroanatomy, sMRI scans employ volumetric and morphometric investigations across three acquisition planes: axial, coronal, and sagittal [4]. These scans are widely used in clinical studies due to their ability to detect precise changes in brain structure and produce high-contrast, spatially-resolved images [4]. In the realm of ASD research, numerous studies have been conducted to explore variations in anatomical structure associated with ASD. The primary objective of these studies is to identify image features that effectively distinguish ASD cohorts from other groups. In fMRI, the brain is represented as voxels, which are small cubes. Each voxel's activity is recorded over time as time series data. These brain scans provide insights into how different brain regions function by measuring blood flow to those areas [55]. This measurement is based on the brain's response to specific tasks or stimuli. When the brain is active, there is an increase in blood flow to supply glucose and oxygen to the relevant regions [56]. By measuring this increased metabolic demand of active neurons, fMRI enables researchers to identify the brain regions involved in specific tasks, including those related to ASD diagnosis.

The availability of online neuroimaging MRI datasets and the development of new learning algorithms have led to a significant amount of ASD research. The research area for ASD diagnosis can be divided into two directions. The first focuses on using MRI modalities to classify patients with ASD and Typical Control (TC) subjects [4,10,12,15,54]. Another area of focus in ASD research is to differentiate between subtypes of ASD such as autistic disorder, Asperger's disorder, and pervasive developmental disorder compared to TC subjects [3,57]. The application of AI models in ASD neuroimaging ranges from analysis of the brain's connectivity patterns, identifying ASD biomarkers, measurement of brain volume, and automatic segmentation of brain structure. Despite being an active area of research, these applications have not yet been implemented in clinical practice. Since many ASD research works have proved the ability of ASD classifiers using MRI modalities, the movement from research into practice is expected to be successful in the next few years. However, there is room for research works to investigate sMRI and fMRI neuroimaging data with advanced AI trends such as transfer learning and transformers from the medical domain.

In this review paper, we will examine the various steps involved in preparing MRI data for researchers in the ASD diagnosis field to use as inputs for AI models. These steps are outlined in Fig. 3. Firstly, the large MRI modalities are digitized. Next, different preprocessing techniques are applied to each MRI modality, including intensity normalization, brain extraction, brain segmentation, and atlas registration. The preprocessed MRI data is then transformed into appropriate representations such as connectivity matrix, graph-based and region-based representations, 2D/3D slices from MRI across sagittal, coronal, and axial planes, and times series data representations. Each representation and a combination of them can capture relevant brain characteristics specific to each class. Finally, classification models, such as ML, DL, VT or BT models, are utilized to distinguish between ASD and other subjects. The following subsections explain each stage in detail.

Table 2
The exclusion and inclusion criteria for ASD diagnosis.

Criteria for Inclusion	Criteria for Exclusion
Articles in English language.	Articles unrelated to ASD.
Articles published in 2020 to 2023.	Articles which are not using MRI.
Articles related to ASD datasets.	Articles that are not related to ASD dataset.
Articles related to sMRI and fMRI neuroimaging.	Articles do not employ AI models for ASD diagnosis.
Articles related to ASD with AI models.	

3.1. MRI acquisition

MRI data are used to study valuable biomarkers to understand the neural underpinnings of ASD. The most commonly open-access datasets

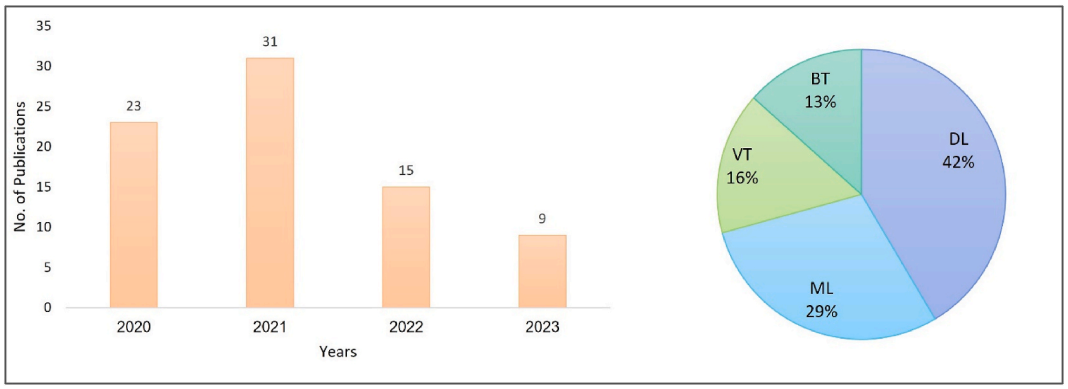
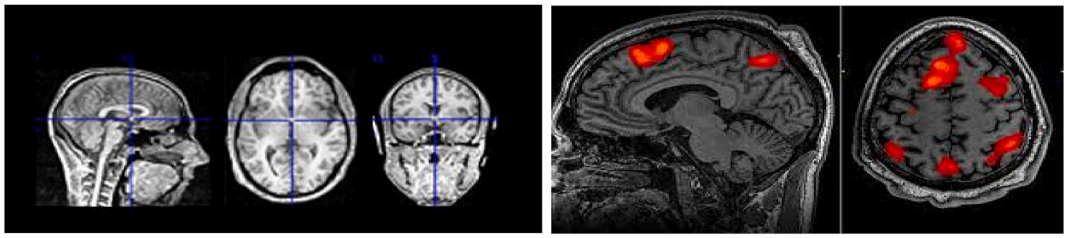


Fig. 1. Chronological distribution of the reviewed articles and the percentage distribution of AI-based techniques used.



(a) Structural MR (sMRI) *(b) Functional MR (fMRI)*

Fig. 2. Structural magnetic resonance imaging (sMRI) as a non-invasive technique for examining the anatomy and pathology of the brain (a), as opposed to using functional magnetic resonance imaging (fMRI) which examine brain activity (b) (photos by The University of Edinburgh) [2]
2 https://www.ed.ac.uk/clinical-sciences/edinburgh-imaging/research/themes-and-topics/medical-physics/imaging-techniques.

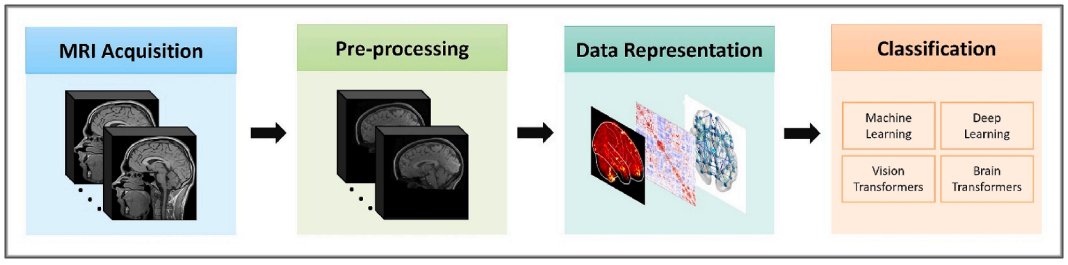


Fig. 3. MRI data preparation workflow for ASD diagnosis.

for ASD diagnosis are Autism Brain Imaging Data Exchange (ABIDE)[2] and the National Database for Autism Research (NDAR).[3] The ABIDE datasets are freely available datasets that consist of both sMRI and fMRI associated with phenotypic data collected from different laboratories in North America and Europe. The repository of ABIDE consists of two large repositories: ABIDE I and ABIDE II. The heterogeneity of the dataset has been pointed out by many researchers as a limitation in performing ML on neuroimaging data. Some researchers conducted their studies based on a single site [12], while others tried to aggregate the data across multiple sites [19,54]. It appears to be more difficult to classify a large database from several sites than a small database from one site. Moreover, scanning parameters and procedures vary between different sites while ensuring data quality. The ABIDE platform offers several advantages, including access to preprocessed data generated using different pipelines and the ability to treat images from a single site as a unified dataset. ABIDE I contains 1112 fMRI, including 539 ASD and 573 others, collected from 17 international sites. Different research studies gave employed datasets from ABIDE I for ASD diagnosis [1,2,5,6,12,14,18,40–44,58–84]. ABIDE II has 521 ASD patients and 593 healthy controls obtained from 19 sites. Each site uses different scanners, data collection protocols, and participant populations that cause heterogeneity in the data [17]. Some works have utilized ABIDE II for ASD research [4,40,68,84].

Another database available for ASD diagnosis is NDAR, established in 2016 by the National Institutes of Health to accelerate autism research. It was created to support the advancement of ASD diagnosis research by facilitating collaboration among researchers and providing access to a large and diverse dataset. NDAR collects, archives, and distributes data from a variety of sources, including demographic information, behavioral assessments, genetic data, neuroimaging data, physiological data, sensor data, survey data, and biospecimen data. The goal of NDAR is to enhance the speed and efficiency of ASD research and ultimately lead to better understanding, treatment, and outcomes for ASD individuals [85,86]. Both resources of datasets have achieved satisfactory results based on the results of the literature studies. Table 3 summarizes these datasets repositories used in various literature works in ASD diagnosis.

[2] http://fcon_1000.projects.nitrc.org/indi/abide/.
[3] https://nda.nih.gov

Table 3
List of ASD data resources with the dataset description, number of participants and their age range.

Repository	Dataset	No. of participants	Age Range
ABIDE I	fMRI, sMRI, Phenotypic	539 ASD 573 TC	7–64 years
ABIDE II	fMRI, sMRI, Phenotypic	521 ASD 593 TC	5–64 years
NDAR	Neuroimaging (including sMRI, fMRI, EEG, and Eye tracking), Phenotypic, Genetic, Omics	Over 80,203 participants (for both ASD and TC)	From toddlers to adults

3.2. Preprocessing techniques

Neuroimaging datasets are large and complex and require preprocessing before they can be utilized effectively. The choice of preprocessing pipeline has a significant impact on the variability of neuroimaging analysis and the reproducibility of findings. This highlights the importance of accurate and efficient data preprocessing to reduce bias and ensure reliable diagnoses [46]. Variations in tasks such as image normalization, registration, and segmentation can greatly influence the features extracted from MRI data. It is crucial to understand and address these preprocessing biases in order to accurately differentiate between ASD-related biological effects and methodological variations [87]. To tackle these challenges, the ABIDE repository provides researchers with access to a dataset that has undergone standard preprocessing. This solution aims to maintain consistency across studies by preventing differences in preprocessing methods and promoting reliable comparisons and analyses. Popular pipelines utilized for analyzing connectomes from ABIDE dataset include the configurable pipeline for the analysis of connectomes (CPAC) [10], data processing assistant for fMRI (DPARSF) [10], connectome computation system (CCS) [19], and neuroimaging analysis kit (NIAK) [19]. In addition, different studies have utilized software tools for preprocessing MRI modalities, including brain extraction tools (BET) [74], the deformable medical image registration toolbox (DRAMMS) [42], FMRIB software libraries (FSL) [68], FreeSurfer [4], and statistical parametric mapping (SPM) [3]. MATLAB was also used to preprocess fMRI data [81,85].

Preprocessing steps for fMRI and sMRI can vary. In the case of fMRI, it entails several techniques to identify brain areas that may be affected by ASD. These techniques include brain extraction, which eliminates non-brain tissue from fMRI images, spatial smoothing that uses a low-pass filter to reduce noise and enhance the signal-to-noise ratio, and temporal filtering to eliminate unwanted noise or enhance specific signal features. Motion correction is performed to address any slight head movements during the scan and prevent image misalignment. Slice timing correction is used to rectify temporal differences in fMRI image acquisition between different slices. Intensity normalization ensures consistency across scans, and registration to a standard atlas is conducted. The registration process aligns an individual's image to a standard atlas, which serves as a reference image segmented into various structures like gray matter, white matter, and cerebrospinal fluid. This alignment allows for the transformation of spatial information into a common space, facilitating the comparison of brain activity across subjects and studies. In ABIDE repository, different atlases can be applied to fMRI including Craddock 200 (CC200) [54], Craddock 400 (CC400) [5], Harvard-Oxford (HO) [63], Automated Anatomical Labeling (AAL) [61], and Dosenbach 160 [58]. On the other hand, the preprocessing step that applied to sMRI involve brain segmentation, skull stripping, and tessellation of the gray-white matter boundary [66]. Brain segmentation is the process of labeling the tissue type or anatomical structure of each voxel. Skull tripping can be part of tissue segmentation but is mostly done by the removal of non-cerebral tissue in sMRI. In the tessellation of the gray-white matter boundary, this step corrects the spherical topology of the surface by constructing a mapping between the original surface onto a sphere and then detecting the topological defects as the minimal non-homeomorphic regions. Intensity normalization and atlas registration can also be applied to sMRI. An example by Ref. [86], of typical preprocessing steps for sMRI including intensity normalization, brain extraction, brain segmentation, and atlas registration is shown in Fig. 4.

3.3. Data representation methods for ASD diagnosis using MRI modalities

While a qualified radiologist can visually inspect medical images and identify signs of various diseases, the sheer volume of MRI data makes it difficult to manually describe and classify the data in a timely manner. Raw MRI modalities are characterized by their large size and high dimensionality, mainly due to the inclusion of spatial information across multiple slices or volumes. It is important to note that only a limited number of studies have utilized raw 3D/4D MRI images as input for AI models in the context of ASD diagnosis. The lack of localized features in raw MRI data can hinder accurate diagnosis and classification tasks. Instead, most research in this field focuses on extracting representative data from the raw images, aiming to capture the essential information relevant to ASD diagnosis. Additionally, using raw 3D/4D MRI modalities directly as input for DL and pretrained VT models may not be feasible due to specific requirements such as input shape, channels, or 2D images. To overcome this limitation, researchers have explored alternative data representations to enable the input of MRI modalities into these models, improve the identification of ASD biomarkers, and enhance classification accuracy. This section provides an explanation of various MRI data representation methods used in ASD diagnosis, including connectivity matrices, 2D slice images, 3D brain volume, graph-based, region-based, and time series analysis. The purpose of this section is to enhance researchers' understanding of these methods and provide guidance in the field.

3.3.1. Connectivity matrices

Connectivity matrices, derived from MRI scans, are commonly used to identify biomarkers for ASD [2,3,5,10,12,14,15,17,44,54,60,70,73,77,79–81,85,88–90]. These matrices provide insights into the structure and activity of the brain. They are widely employed in MRI analysis because they can capture and quantify the interactions between different brain regions. To generate these matrices, researchers typically use a brain atlas to divide the brain into regions with similar properties, reducing the dimensionality. Statistical measures, such as correlation coefficients, are then calculated to assess the level of co-activation between pairs of regions of interest (ROIs). These measures are computed based on the selected brain atlas, resulting in weighted connectivity matrices that indicate the strength of connections between corresponding ROIs. These matrices serve as valuable input for classification models that require 2-dimensional data. Fig. 5 provides a detailed visual representation of the process of generating a 2D connectivity matrix from MRI modalities. It is important to note that connectivity matrices derived from fMRI are referred to as functional connectivity (FC), while those obtained from sMRI are known as structural connectivity (SC). Various statistical measures, such as correlation [10,15,63], covariance [3], and standard deviation [62], are used to generate the FC and SC matrices. In the case of FC, several studies have used Pearson's correlation to construct the matrix [2,5,14,44,60,70,73,79–81,85,88–90], while others have explored FC using different measures [3,10,12,15,17,54,77]. A recent advancement in neuroimaging data fusion has examined the integration of multiple datasets from the same subject to investigate the functional significance of brain regions and changes in activation in neurological disorders. By combining data processed from different atlases, more informative features for ASD diagnosis can be obtained [7,59]. Some studies have merged multiple FC based on different brain atlases using DL models. Additionally, the combination and evaluation of spatial information of brain connections with Pearson correlation FC have also been explored [58,63]. The SC can be estimated

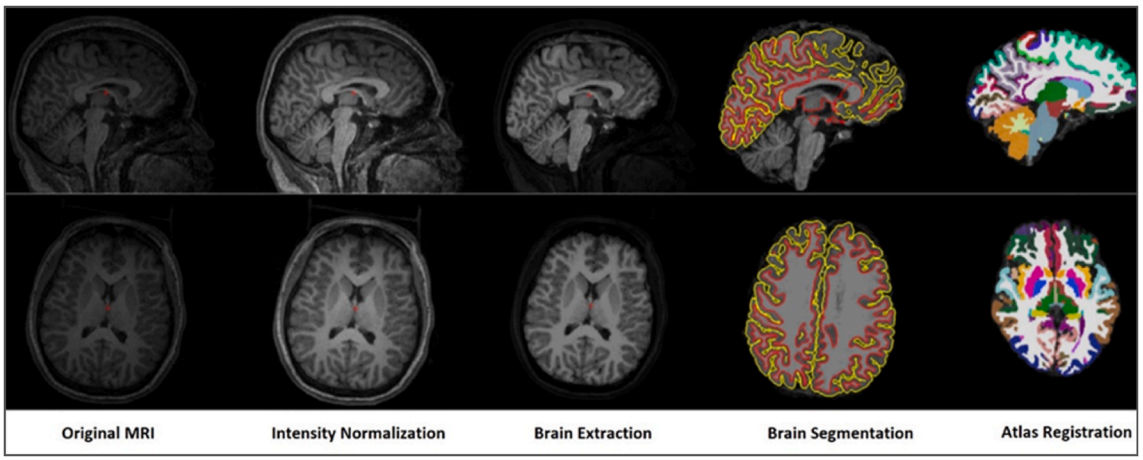


Fig. 4. Example of general preprocessing steps for sMRI [86].

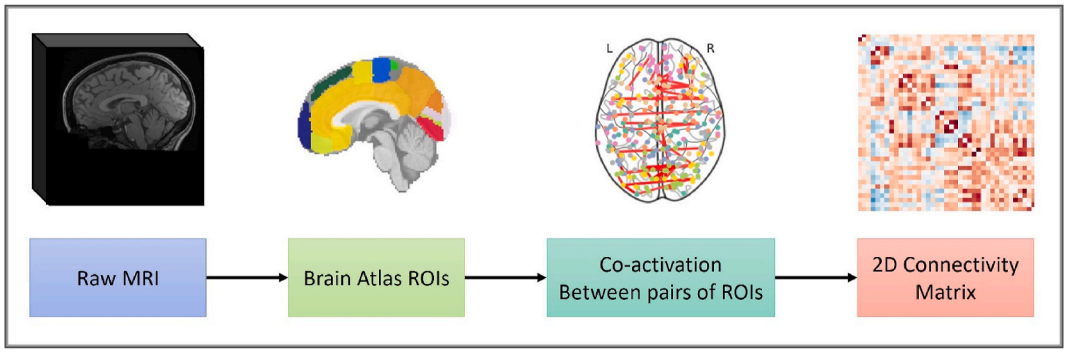


Fig. 5. Generation of connectivity matrix from MRI modalities.

using the individual-level morphological covariance brain network approach, which utilizes sMRI data to capture interregional structural connectivity and variations. Previous studies, such as [42,83], have employed this approach and constructed networks based on structural MRI data. By using GM volume maps, researchers were able to create networks that represent interregional structural connectivity and variations. They excluded certain Vermis regions with low signal-to-noise ratio and obtained 2D matrices that capture the relationships between different brain regions. In some studies, both the structural and functional characteristics of individuals with ASD were considered by combining information from sMRI and fMRI [18,40,74,86]. Instead of using only one type of connectivity matrix, researchers combined FC and SC matrices as features. For example, FC was generated using Pearson correlation, while SC was computed by examining connections between gray matter volumes of cortical parcels identified by a brain atlas [18]. Two strategies were proposed to merge FC and SC information: concatenating the two feature vectors or independently training the classifier on each one.

3.3.2. *Graph-based representations*

Graph theory principles have been employed to describe brain networks using graph-based representations. This approach allows researchers to extract valuable information from brain networks for various learning tasks, such as brain network classification and visualization. The main objective is to create a meaningful and useful representation of the intricate connections within the brain. This involves embedding the brain network into a lower-dimensional representation, which has been shown to improve the accuracy and performance of disease diagnosis, including ASD. Researchers have implemented graph-based representations from fMRI for ASD diagnosis [6,61,62,64,65,67,72,75,76,91–94]. The generation process involves constructing a graph composed of nodes representing specific brain ROIs and links representing the connectivity between these regions. The links can have weights indicating the strength of the connectivity, typically derived from FC analysis of fMRI data. These graph networks provide a comprehensive representation of the functional interactions and communication patterns within the brain. The general process for constructing graph-based representations from MRI modalities is depicted in Fig. 6. One proposed approach modeled brain regions as nodes and their FC as the edges of the graph [64,65], while another approach represented subjects from the dataset as nodes and utilized the similarity between subjects' fMRI and phenotypic features (such as age, sex, and site number) as edges [67].

3.3.3. *2D-slice images*

Multiple studies have been carried out on generating 2D images or slices from raw 3D or 4D MRI neuroimaging data. These slices serve as snapshots of the brain's structural or functional information in a 2D format. This approach is particularly useful when working with pre-trained models that require a 2D input. In the context of fMRI, researchers have explored the extraction of 2D slices to visualize brain activity for ASD diagnosis [19,82,95–97]. This involves selecting slices from different anatomical planes (sagittal, coronal, or axial views) and specific time points to provide localized representations of brain activity. Glass brain or stat map images are often generated to enhance the visualization of brain activity [19,95]. Similarly [4,98], studies have also focused on extracting 2D slices from sMRI data to visualize the brain's structural anatomy. These slices, obtained by slicing the 3D sMRI along specific planes (axial, sagittal, or coronal), offer detailed visualizations of the brain structure. An illustration of these 2D slices can be

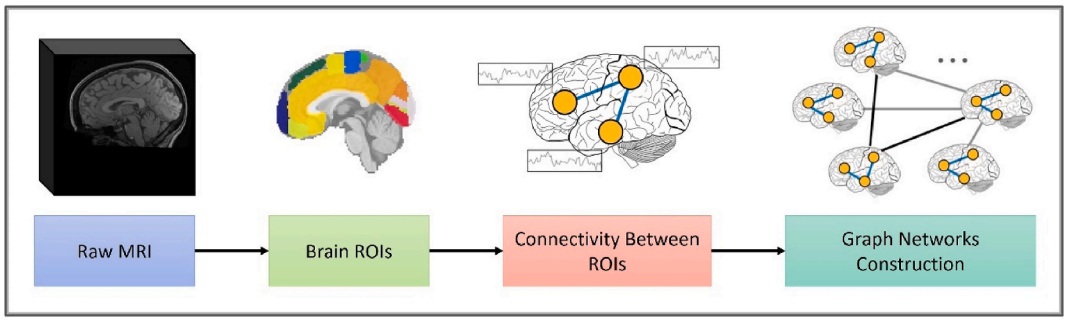


Fig. 6. Graph-based representations construction from MRI modalities.

seen in Fig. 7. These slices have been used as input for DL and pretrained models in the context of ASD diagnosis.

3.3.4. 3D-volume

The 3D brain volume refers to the three-dimensional representation of the entire brain obtained through MRI imaging. Fig. 7 shows the difference between 2D slices and the 3D brain volume data. Using the complete volume of 3D MRI images can be valuable when using DL or pretrained VT models for diagnosing ASD. This approach retains the spatial information and structural details of the original images, allowing for a comprehensive representation of the brain. As a result, the models can capture intricate patterns and relationships throughout the brain volume, potentially leading to a more accurate diagnosis of ASD. Several studies have demonstrated the effectiveness of using 3D MRI images as input for DL and pretrained models in diagnosing ASD. For example, in a study by Ref. [41], a 3D CNN was used to extract features from MRI scans, achieving a 70 % accuracy in distinguishing between ASD and non-ASD subjects. Another study utilized a pretrained 3D-ResNet model to extract representative and discriminative features from MRI data for ASD diagnosis [99].

3.3.5. Region-based representations

The discovery of brain imaging markers for autism is crucial for understanding the causes of the disorder. However, using raw MRI data from all brain regions as input presents challenges in extracting meaningful features for ASD diagnosis. Region-based MRI data representation involves focusing on specific areas or ROIs within the brain. This often requires segmenting or delineating specific brain regions or structures from the 3D brain volume to extract region-specific information or features. Fig. 8 presents a general process for extracting region-based data representation from MRI modalities. Several studies have explored different methods of extracting regions from MRI data for ASD diagnosis [1,11,41,66,68,100–103]. For example, in one study [41], raw MRI data was input into a 3D CNN to extract discriminative features. Higher-order analysis was then used to identify specific ROI by utilizing a brain atlas, and these regions were used as masks. The regions, along with the CNN extracted features, were input into a Genetic Algorithm for ASD diagnosis, resulting in a 73 % accuracy. In another study [100], the 3D cerebral cortex was mapped into 2D images using a geometry mapping approach that preserved the spatial information of different brain regions along the cortical surface. These 2D images served as input for ASD diagnosis. In yet another study [68], MRI data was segmented into 117 ROIs using a brain atlas. Each segmented brain region was transformed into 2D images, preserving their topological relationships. These 2D images were then processed using the fast discrete curvelet transform to obtain Curvelet space representations. This utilization of Curvelet space representations allowed for the characterization of brain regions, specifically capturing regional differences between ASD and non-ASD subjects.

3.3.6. Time series analysis

Studies investigating the diagnosis of ASD have placed a significant emphasis on examining the temporal aspects of BOLD time-series data derived from fMRI scans [57,71,84,104–106]. The objective is to capture temporal fluctuations and gain insights into the influence of anatomy on signal pattern transitions. Initially, fMRI signals are represented in two ways: temporal-level and spatial-level representations. The temporal-level representations involve signals obtained from various ROIs at each time point. In contrast, the spatial-level features comprise signals gathered from each ROI at different time points. To encompass the overall temporal characteristics of entire networks, the time series data from individual regions within each network are either averaged or subjected to different statistical measures. Fig. 9 displays the representations of a sample from time series data used for ASD detection. In Ref. [84], a total of nine statistical measures were employed to summarize the temporal dimension for each voxel. The mean and standard deviation calculations were performed on a per-voxel basis and used as input for the classification model. In Ref. [106], the time series data was extracted from a 4D fMRI scan using a brain atlas. Specifically, the representative time series was obtained by averaging the responses across voxels within the ROI. These representative time series data were then utilized as input for their transformer-based model for diagnosing ASD. Similarly, in other studies conducted by Refs. [57,105], brain atlases were used to extract brain ROIs and the average of the time-series data within each ROI was employed as input for their transformer-based models.

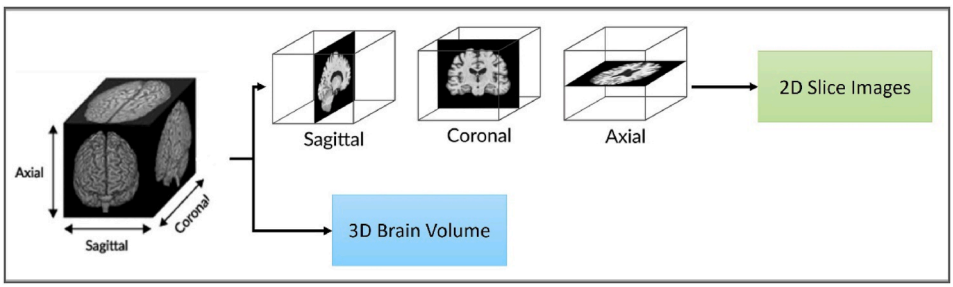


Fig. 7. 2D slices and 3D brain volume data representations from MRI modalities.

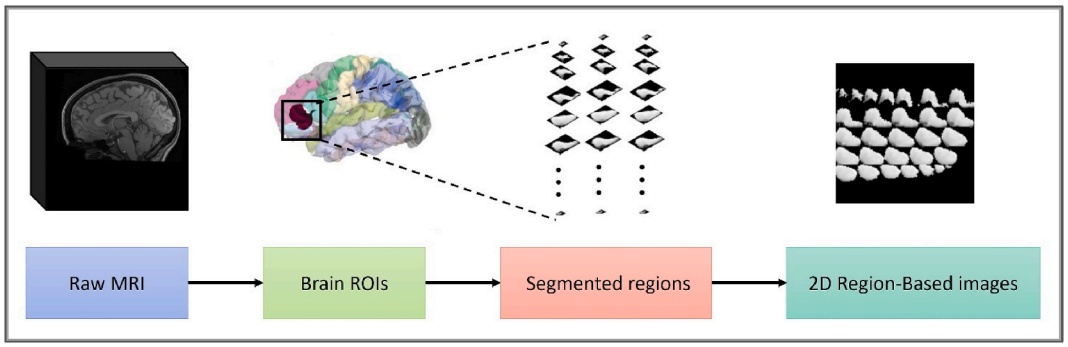


Fig. 8. Region-based data representation from MRI modalities.

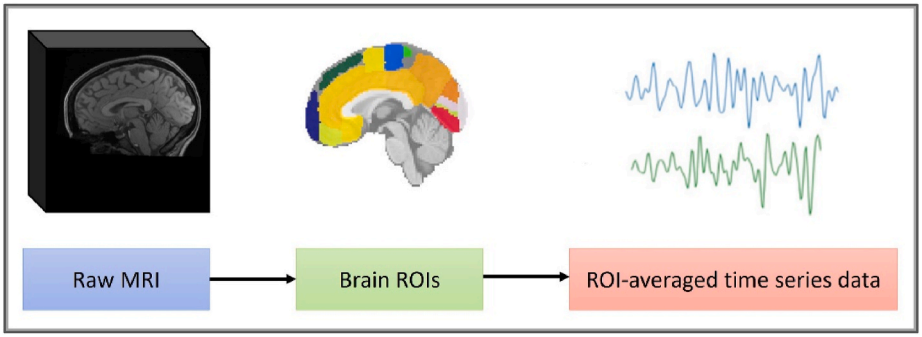


Fig. 9. Time series data extracted from MRI ROIs.

4. From machine learning to deep learning: advancing ASD diagnosis

Traditional research works have used various ML models for ASD diagnosis based on features extracted from MRI modalities. ASD is commonly a binary classification problem wherein the dataset contains two class labels, ASD and TC subjects. ASD can also be multi-classification problem wherein the dataset is labeled using ASD subtypes namely autistic disorder, Asperger's disorder, pervasive developmental disorder, and TC. The initial and important stage in utilizing ML models involves extracting and selecting features. In the context of ASD diagnosis, researchers extract representative data from raw MRI modalities to be inputted into ML models. However, this process often results in high-dimensional data, which can potentially reduce accuracy in diagnosing ASD. Feature reduction methods are used to transform the data from a high-dimensional space into a lower dimension without losing important information [39]. Feature reduction methods improve the performance of ML models, avoid overfitting, and reduces the running time in ASD diagnosis. Methods include *t*-test [12,58,63,76,85], recursive feature elimination (RFE) [6,11,18,66], analysis of variance (ANOVA) test [62], principal component analysis (PCA) [69], F-score [79], Fisher score [18] and autoencoder (AE) [2,7,70,71,78]. Support Vector Machine (SVM) algorithm has been widely used in ASD diagnosis, with promising results achieved in various studies [1,3,10–12,54,60,62,63,68,69,75]. For example, one study achieved an accuracy of 92.9 % using ASD-related functional connectivity (FC) data extracted from fMRI as input [81]. Another study utilized a cubic kernel function SVM to classify ASD using non-oscillatory FC data, resulting in an accuracy of 88.9 % [3]. A multi-view high order-FC network from fMRI was extracted which achieved 86 % accuracy [12]. One limitation of the aforementioned studies was the use of small datasets; 98 subjects [81], 144 subjects [3], and 92 subjects [12]. Random Forest (RF), another ML algorithm, has also been employed in ASD classification based on MRI data [1,70,78,86,88]. In one study, FC and volumetric features extracted from fMRI and sMRI were concatenated and fed to an RF classifier, resulting in a maximum accuracy of 80 % [86]. FC from fMRI and conditional RF were evaluated to reduce features dimensionality and to select optimal features, which attained 73 % accuracy [88]. K-Nearest Neighbors (KNN) is another ML model used in ASD classification. Features extracted from sMRI and fMRI were combined, achieving accuracies of 75 % and 79 % on a relatively large dataset [9]. KNN classifier with 11 neighbors and cosine metric achieved 78 % [3], and 58 % [1] accuracies on small datasets. Other ML models, such as Decision Tree (DT) [61], Logistic Regression (LR) [62], Ridge Classifier [54], Sparse Representation Classifier (SRC) [90], Ensemble learning (EL) [3,58], and Linear Discriminant Analysis (LDA) [1] have also been utilized in ASD diagnosis based on MRI data. The ensemble classifier achieved greater precision by utilizing several classifiers or classification models [15]. A study used self-weighted adaptive structure learning (SASL) method for ASD diagnosis [58], which utilized the FC derived from fMRI as weights and performed the SASL method achieving accuracy of 89 %. Table 4 outlines ML models used by traditional ASD diagnosis research, and summarized brain atlas, preprocessing techniques, feature extraction and selection methods from MRI modalities.

For advancing ASD diagnosis, researchers have explored DL models. In practice, DL models share some common basic properties and differ in their network architecture or in the way they are trained. A strength of deep networks is that the ability to automatically learn hidden features from the input data, understand the characteristics of unstructured data, and perform classification based on those characteristics [22]. Most DL models used in ASD diagnosis including Multi-Layer Perceptron (MLP) [1,5,7,14,17,18,59,65,66,77,80,82], Convolutional Neural Networks (CNN) [1,5,7,14,17,18,59,65,66,77,80,82], Autoencoder (AE) [2,4,7,18,70,71,77–79,107], Graph Convolutional Networks (GCN) [6,16,64,91], and Graph Attention Networks (GAT) [63,72]. Other variations of DL methods in relation to ASD including Deep Attention Neural Network (DANN) [89], Self-Attention Neural Network (SANN) [4,83], Graph Neural Network (GNN) [94], and Long-Short Term Memory (LSTM) [71] are discussed. More detailed explanations of DL methods in relation to ASD are provided in the subsequent subsections. These explanations are

Table 4
ASD diagnosis using MRI modalities and ML models: Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), Random Forest (RF), and K-Nearest Neighbors (KNN), Logistic Regression (LR), Decision Tree (DT), Linear Discriminant Analysis (LDA), and Sparse Representation Classifier (SRC).

Model	Data	Brain atlas	Pre-processing	Feature	Feature selection	Validation	Max Accuracy	No. of cases	Ref.
SVM	fMRI	HO	CPAC	Pearson's correlation-based spatial constraints	two sample t-tests, LASSO methods	10-CV	67.28 %	403 ASD. 468 TC	[63]
	fMRI	Glasser multimodal	CPAC	Graph-based on standard deviation of time series	ANOVA	CV	60.89 %	403 ASD, 468 TC	[62]
	fMRI, Phenotypic	CC400	–	FC with tangent Pearson	–	5-CV	71.1 %	505 ASD, 530 TC	[54]
	fMRI	BASC-064	DPARSF	FC with multiple Gaussian kernels	–	LOO-CV	68.42 %	280 ASD, 329 TC	[10]
	fMRI	AAL	SPM8	clustering-based multi-view high-order FC	*t*-test, LASSO	6-CV	86.2 %	45 ASD, 47 TC	[12]
	fMRI	AAL	SPM12, DPABI, CONN	non-oscillatory FC	P-value analysis	10-CV	88.9 % 78.08 %	36 ATD, 36 APD 36 PDD, 36 TC	[3]
	fMRI	AAL	CPAC	Dynamic FC	Multitask feature selection method	10-CV	76.8 %	403 ASD, 468 TC	[60]
	sMRI	HO	FSL	Brain geometric curvelet space	Generalized Gaussian distribution	LOO-CV	AUC 0.77 %	151 ASD, 151 TC	[68]
	fMRI, sMRI	AAL	–	Principal components of whole-brain phase synchrony, age feature	PCA	LOO-CV	78.89 %	49 ASD, 41 TC	[69]
	sMRI	DK	FreeSurfer	Brain cortical structures	RFE	LOO-CV	84.2 %	40 ASD, 36 TC	[11]
	fMRI	CC200	CPAC	Brain networks constructed with the mean of time series	Extra-trees	10-CV	72.2 %	506 ASD, 548 TC	[75]
	sMRI	–	–	Corpus callosum area, intracranial brain volume	Entropy, chi-square, symmetric uncertainty	5-CV	52.2 %	592 ASD, 571 TC	[1]
	fMRI	–	CPAC	FC	AE	CV	60.2 %	539 ASD, 573 TC	[78]
	fMRI, Genetic	AAL	MATLAB, SPIM12	FC + Gene Expression	*t*-test, SVM-RFE	LOO-CV	83.6 %	47 ASD, 24 TC	[85]
	fMRI	Dosenbach	DPABI, SPM. MATLAB	FC	Boruta method	LOO-CV	92.9 %	48 ASD, 50 TC	[81]
	fMRI	HO	CPAC, FSL	9-statistical summary measures	–	5-CV	66 %	620 ASD, 542 TC	[84]
	fMRI	AAL	DPARSF	Temporal dynamic of BOLD signals	–	–	78 %	41 ASD, 41 TC	[104]
RF	fMRI	CC200	DPARSF	FC	AE	10-CV	60.63 %	432 ASD, 556 TC	[70]
	fMRI, sMRI	DK		Morphological features, FC	–	4-CV	80 %	72 ASD, 113 TC	[86]
	fMRI	HO	FSL	FC	conditional RF	–	73.75 %	306 ASD, 350 TC	[88]
	sMRI	–	–	Corpus callosum area, intracranial brain volume	Entropy, chi-square, symmetric uncertainty	5-CV	54.79 %	592 ASD, 571 TC	[1]
	fMRI	–	CPAC	FC	AE	CV	60.2 %	539 ASD, 573 TC	[78]
KNN	sMRI	–	–	Corpus callosum area, intracranial brain volume	Entropy, chi-square, symmetric uncertainty	5-CV	54.79 %	592 ASD, 571 TC	[1]
	fMRI	AAL	SPM12, DPABI, CONN	Non-oscillatory FC	P-value analysis	10-CV	78.08 % 60.97 %	36 ATD, 36 APD 36 PDD, 36 TC	[3]
	sMRI. fMRI	AAL, DK	BET, CCS, FreeSurfer	Morphological features, FC	RF and grid search	CV	79 %	561 ASD, 521 TC	[9]
	fMRI	–	CPAC	FC	AE	CV	60.2 %	539 ASD, 573 TC	[78]
	fMRI	AAL	DPARSF	Temporal dynamic of BOLD signals	–	–	78 %	41 ASD, 41 TC	[104]
LR	fMRI, Phenotypic	CC200	–	FC with Tangent Pearson	–	5-CV	71.1 %	505 ASD, 530 TC	[54]
	fMRI	Glasser multimodal	CPAC	Graph-based on standard deviation of time series	ANOVA	CV	60.89 %	403 ASD, 468 TC	[62]
DT	fMRI	AAL	CPAC	Graph-based on BOLD time-series	Two sample *t*-test, LASSO methods	LOOCV	74.8 %	201 ASD, 251 TC	[61]
LDA	sMRI	–	–	Corpus callosum area, intracranial brain volume	Entropy, chi-square, symmetric uncertainty	5-CV	54.79 %	592 ASD, 571 TC	[1]
Ridge	fMRI. Phenotypic	CC200	–	FC with Tangent Pearson	–	5-CV	71.1 %	505 ASD, 530 TC	[54]
Ensemble	fMRI	–	CPAC	Population graphs of image, phenotypic features	Graph signal processing	10-CV	73.13 %	403 ASD, 468 TC	[67]

(*continued on next page*)

Table 4 (*continued*)

Model	Data	Brain atlas	Pre-processing	Feature	Feature selection	Validation	Max Accuracy	No. of cases	Ref.
	fMRI	AAL, CC200, Dosenbach	DPARSF	FC	*t*-test	10-CV	89.13 %	159 ASD, 197 TC	[58]
SRC	fMRI	AAL	DPARSF	FC	–	4-CV	–	175 ASD, 234 TC	[90]

Table 5
ASD diagnosis using MRI modalities and DL models: Convolutional Neural Network (CNN), Graph Attention Network (GAT), Graph Convolutional Networks (GCN), Graph Neural Network (GNN), Deep Attention Neural Network (DANN), Self-Attention Neural Network (SANN), Multi-Layer Perceptron (MLP), Autoencoder (AE), and Long-Short Term Memory (LSTM).

Model	Dataset	No. of cases	Brain Atlas	Preprocessing	Feature selection	Max Accuracy	Ref.
CNN	fMRI	529 ASD, 573 TC	–	CPAC	–	83 %	[19]
	sMRI, fMRI	1555 (Both subjects)	ALL	SpeedyPP	–	69.39 %	[40]
	fMRI	539 ASD, 573 TC	CC200	CPAC	–	80 %	[15]
	sMRI	518 ASD, 567 TC	SRI24	DRAMMS	–	71 %	[42]
	sMRI	500 ASD, 500 TC	FSL	HO	Genetic algorithm	73 %	[41]
	sMRI	592 ASD, 571 TC	–	–	Entropy, chi-square, symmetric uncertainty	66 %	[1]
	fMRI	539 ASD, 573 TC	–	CPAC	AE	84.05 %	[78]
	fMRI	505 ASD, 530 TC	CC400	CPAC	–	70.22 %	[5]
	fMRI	505 ASD, 530 TC	AAL	CPAC	–	74 %	[44]
	fMRI	79 ASD, 105 TC	CC200	CPAC	–	77.74 %	[43]
	fMRI	79 ASD, 105 TC	–	CCS	–	94.70 %	[82]
	fMRI	620 ASD, 542 TC	HO	CPAC, FMRIB	–	64 %	[84]
	fMRI	1711 ASD, 42147 TC	AAL	SPT	–	67.02 %	[108]
	fMRI	539 ASD, 573 TC	CC400	CPAC	–	95 %	[103]
	sMRI	–	–	Wiener filter	–	95.7 %	[101]
GAT	fMRI	403 ASD, 468 TC	HO	CPAC	Two sample *t*-test, LASSO	72.40 %	[63]
	fMRI	505 ASD, 530 TC	HO	CPAC	–	95 %	[72]
GCN	fMRI	402 ASD, 464 TC	AAL	PCP	–	73.1 %	[64]
	fMRI	403 ASD, 468 TC	HO	CPAC, MNI152, FSL	Deep feature selection	79.5 %	[73]
	fMRI	403 ASD, 468 TC	HO	CPAC	RFE	73.71 %	[6]
	fMRI	402 ASD, 464 TC	MODL	–	–	72.5 %	[91]
	fMRI	403 ASD, 468 TC	MA	CPAC	–	86.3 %	[16]
GNN	fMRI	408 ASD, 476 TC	AAL	DPARSF	–	79.78 %	[94]
DANN	fMRI, Phenotypic	408 ASD, 401 TC	AAL, HO, CC200	CPAC	–	73 %	[89]
SANN	sMRI	518 ASD, 567 TC	SRI24	DRAMMS	–	72.48 %	[83]
MLP	fMRI	493 ASD, 530 TC	CC200, AAL	CPAC, FSL	–	84 %	[65]
	fMRI	432 ASD, 556 TC	CC200	DPARSF	AE	71.35 %	[70]
	fMRI	419 ASD, 530 TC	CC200, AAL, Dosenbach	CPAC	AE	74.52 %	[7]
	fMRI	505 ASD, 530 TC	CC200	CPAC	Sparse AE	70.8 %	[2]
	fMRI	403 ASD, 468 TC	AAL	CPAC	–	69.81 %	[14]
	fMRI	539 ASD, 573 TC	AAL	CPAC, FMRIB, SPM, Artifact detection tools	Sparse AE	81.5 %	[77]
	fMRI	402 ASD, 464 TC	BASC	CPAC	–	88 %	[17]
	fMRI	505 ASD, 530 TC	CC200	–	F-score, AE	70.9 %	[79]
	fMRI	505 ASD, 530 TC	CC400	CPAC	–	75.27 %	[80]
	fMRI	79 ASD, 105 TC	–	CCS	–	94.70 %	[82]
	sMRI, fMRI	368 ASD, 449 TC	AAL, CC200, Destrieux	CPAC	Fisher score, AE	85.06 %	[18]
	fMRI	506 ASD, 532 TC	CC200, AAL90, DOS160	DPARSF	–	79.13 %	[59]
AE	sMRI	20 ASD, 15 TC	–	FreeSurfer	–	86.95	[4]
	fMRI	432 ASD, 556 TC	CC200	DPARSF	AE	71.35 %	[70]
	fMRI	322 ASD, 352 TC	CC200	CPAC	AE	71.3	[71]
	fMRI	505 ASD, 530 TC	CC200	CPAC	Sparse AE	70.8 %	[2]
	fMRI	419 ASD, 530 TC	CC200, AAL, Dosenbach	CPAC	AE	74.52 %	[7]
	fMRI	539 ASD, 573 TC	AAL	CPAC, FMRIB, SPM, Artifact detection tools	Sparse AE	81.5 %	[77]
	fMRI	539 ASD, 573 TC	–	CPAC	AE	84.05 %	[78]
	fMRI	505 ASD, 530 TC	CC200	–	F-score	70.9 %	[79]
	sMRI, fMRI	368 ASD, 449 TC	AAL, CC200, Destrieux	CPAC	Fisher score	85.06 %	[18]
LSTM	fMRI	322 ASD, 352 TC	CC200	CPAC	AE	71.3	[71]

further summarized in Table 5.

4.1. Multi-layer perceptron (MLP)

MLP is a type of artificial neural network (ANN) that consists of interconnected layers of neurons for information processing [1]. Each layer receives inputs from the previous layer, and the final layer predicts the corresponding class [39]. Several research studies have utilized MLP for ASD diagnosis using MRI modalities [1,5,7,14,17,18,59,65,66,77,80,82]. For instance, in one study [7], discriminative feature representation was extracted from multi-Atlas FC derived from fMRI using autoencoders. The weights from the hidden layers in the autoencoders were used as initial weights in the MLP. The Softmax regression method was employed in the output layer to determine the label for each subject. This proposed method achieved an accuracy of 74.5 % using 949 subjects. Another study utilized a sparse autoencoder with Pearson's correlations from 200 brain regions as input, followed by MLP with two hidden layers and a softmax function in the output layer. The MLP and autoencoder were trained simultaneously, resulting in a classification accuracy of 70.8 % [2]. In a different study, MLP with two hidden layers and FC was used as input features, and a sigmoid function was applied to the output layer for classifying ASD subjects. This model achieved an accuracy of 69.8 % using 871 subjects [14]. Additionally, a method for interpreting the trained model was proposed based on a linear formula for each input subject. The fMRI data with three different brain atlases were utilized as input for MLP, with feature extraction layers learning the three sets of features and concatenating them using a fusion layer. The Softsign function was used for classification, resulting in a model with 79 % accuracy [59]. Another study used the corpus callosum area and intracranial brain volume extracted from sMRI as input to MLP for classifying ASD subjects, achieving an accuracy of 56 % [1]. Lastly, MLP with two hidden layers was trained using graph theoretical measures extracted from fMRI and achieved an accuracy of 84 % [65].

4.2. Convolutional neural network (CNN)

CNN is a type of feedforward neural network that includes convolutional layers, which apply filters to incoming signals by sliding them across the input image [22]. These filters are used to extract relevant features from the image by considering spatial dependencies. Numerous researchers have employed CNN to classify individuals with ASD versus typical control subjects using MRI images [1,5,15,19,40–44,78,82,84,101,103,108]. In one study, SC and FC derived from sMRI and fMRI were combined, and a CNN was trained, tested, and validated using independent samples [40]. The softmax layer was utilized for subject classification, and the average accuracy achieved was 69.4 %. Another study proposed an enhanced CNN specifically designed for ASD diagnosis using fMRI [15]. This network architecture included temporal convolutional layers with filter sizes of 32 and 64, respectively, and employed causal convolutions and dilations, making it ideal for processing time series data extracted from fMRI. The model was evaluated using 10-fold cross-validation and achieved an average accuracy of 80 %. A 3D-CNN was designed for ASD diagnosis using an sMRI dataset and trained from scratch [41]. This network consisted of three convolutional layers, three max pooling layers, and two fully connected layers for data classification. The model was trained using the cross-entropy loss function and the Adadelta optimizer, resulting in a classification accuracy of 73 %.

4.3. Autoencoder (AE)

AE is a type of unsupervised deep learning model that aims to reproduce its input, allowing the network to learn fundamental data representations [22]. It comprises input, hidden, and output layers, as well as two essential components: the encoder and decoder. Recent studies in ASD diagnosis have explored the use of deep AE in an unsupervised manner on the ABIDE dataset [2,4,7,18,70,71,77–79,107]. Unsupervised feature learning has the advantage of not assuming any specific patterns in the input data that may assist in accurate classification, enabling the identification of neural patterns without human intervention [2]. Based on these learned representations, a generative adversarial network (GAN) was employed to classify ASD and non-ASD subjects using sMRI, achieving a relatively high accuracy of 86.95 % [4]. The GAN was trained exclusively on non-ASD scans and treated ASD scans as outliers. Three adjacent sMRI slices were used as input to the GAN for reconstructing the next three slices, and a self-attention mechanism was incorporated to capture global dependencies among image regions. Other studies have used AE to reduce the dimensionality of input scans. For example, a hybrid classifier that combined an AE with ML models was proposed for ASD diagnosis using fMRI data [78]. The AE was constructed by reducing the number of neurons in the hidden layer to extract essential features from the data. The highest classification accuracy of 84 % was achieved when the AE was combined with a CNN.

4.4. Graph convolutional networks (GCN)

Graphs are data structures composed of interconnected nodes. GCN is a type of neural network that is specifically designed to process graph-structured data [109]. It combines node features and structural information to learn deep representations of graphs. GCN has shown superior performance compared to other methods for relational learning. Incorporating a brain network into a concise and low-dimensional representation improves the accuracy of disease diagnosis. Recent research has applied GCN to extract latent features from functional network graphs constructed from fMRI data for ASD diagnosis [6,16,64,91]. In one study [91], the FC matrix was transformed into a graph, where each row vector represented a node and the FC weights between different brain regions were represented as edges. The model consisted of several layers, including a functional graph construction layer, two graph convolutional layers, a fully connected layer, and an output layer. The Adam optimizer was used to minimize the loss, resulting in a 72.5 % accuracy in five-fold cross-validation. Another study proposed a novel approach for ASD diagnosis using GCN [6]. In this approach, a graph structure was created by combining phenotypic measures with neuroimaging data from the ABIDE database. The FC of each subject was represented as nodes, and the connections between nodes were determined based on similarity, calculated using gender, collection site, and age. The GCN model architecture included multiple hidden layers, a ReLU activation layer, an output layer, and a softmax layer. A dropout of 0.3 was also applied to prevent overfitting, and the model achieved an accuracy of 73.7 %.

4.5. Graph attention network (GAT)

The GAT was originally introduced for classifying graph-structured data [110]. GAT combines the attention mechanism with GCN, allowing the network to focus on important nodes by calculating attention coefficients from the current node and its neighbors. This weighted attention coefficient helps reduce the influence of noisy edges in the network. In the context of ASD and TC classification, a GAT model was proposed to classify functional brain networks [72]. The model employed graph attention layers to learn node representations and an attention pooling layer to obtain the overall functional brain network representation based on these node representations. The brain networks were constructed from fMRI data, where each node represented ROI identified by the HO brain atlas, and the edges represented the functional connectivity between these ROIs. The GAT model achieved an accuracy of 95 % in classifying 1035 subjects in the ABIDE-I dataset. Another study utilized the GAT model to extract node features from fMRI data, using two layers and batch normalization for feature normalization [63]. These extracted features were then fed into MLP for

classification, achieving a maximum accuracy of 72.4 %.

4.6. Other DL models for ASD diagnosis

Some literature works used other DL models to classify ASD subjects based on MRI modalities including Deep Attention Neural Network (DANN) [89], Self-Attention Neural Network (SANN) [4,83], Graph Neural Network (GNN) [94], and Long-Short Term Memory (LSTM) [71]. For example, SANN was used for ASD diagnosis using sMRI dataset using two self-attention layers followed by Leaky ReLU activation and normalization layers which achieved 72.48 % accuracy [83]. ASD diagnosis with LSTM reported an accuracy of 71.3 % [71]. More details on the datasets and preprocessing techniques used by each model are given in Table 5.

5. Vision transformers (VT) for ASD diagnosis

5.1. Architectures of VT for ASD diagnosis

Transfer learning has gained popularity in the field of deep learning, particularly in computer vision. This approach involves utilizing a pre-trained model that has been trained on a general task such as image classification, and then fine-tuning it for a more specific task like object detection or semantic segmentation [22,111]. By leveraging the prior knowledge of the pre-trained model, transfer learning saves time and resources compared to starting from scratch. It has proven to be highly successful, often achieving top performance in benchmark tests and real-world applications. Convolutional neural networks (CNNs) have played a significant role in this success, thanks to their ability to learn structured image representations through expanding receptive fields. Techniques such as fine-tuning and freezing layers have been used with CNN architectures to further improve image classification and outperform traditional machine learning models. Computer vision models frequently use CNNs architectures with different variants include the following: VGG [21,22], ResNet [23,24], Xception [25], Inception [26,27], DenseNet [28], MobileNet [29,30], NASNet [31], and EfficientNet [32,33]. Large transformer-based methods to solve computer vision tasks include ViT [34], Swin [35], MobileViT [36], and ConvNeXt [38] transformers. These models often serve as the backbone and have their final layers adjusted to fit the new task. Thus, transfer learning can offer solutions to challenges in brain disorders research by allowing the replication of computer vision models using larger, more heterogeneous datasets [19].

Fig. 10 demonstrates a systematic framework and architecture for vision transformers (VT) used in the diagnosis of ASD using MRI modalities. This architecture is based on the existing literature that utilizes MRI without any prior feature extraction for classification. The architecture consists of three stages. In the first stage, the input data can be either structural (sMRI) or functional (fMRI) images, each having different characteristics. These MRI modalities are typically stored in 3D or 4D formats. However, vision transformer models require 2D input data, so a technique has been proposed in the literature to extract 2D representative data by extracting 2D image slices from the raw 3D/4D MRI modalities. This allows the volumetric data to be transformed into a format compatible with vision transformer models, enabling their use in ASD diagnosis. A choice needs to be made regarding the use of either sMRI or fMRI and selecting one plane (sagittal, coronal, or axial), a combination of two planes, or all three planes. Additionally, MRI scans are acquired as a series of slices, with each slice acquired at a specific time and duration required to move the radiofrequency coil between slices. Therefore, the number of slices taken from each plane needs to be determined. Researchers have used varying numbers of slices, ranging from one to 60 slices per plane [4,82]. Different image pre-processing operations are performed on the brain images using Brain Atlases. Fig. 11 shows a visualization of human brains after applying different human brain atlases. These atlases provide comprehensive information about the structure and function of the human brain by combining data from multiple sources, such as anatomy, neuroimaging, and genetics. They offer details about the location and function of different brain regions and how they are interconnected. Brain atlases are utilized by researchers and clinicians to enhance understanding of brain function and development, diagnose neurological disorders, and plan surgical interventions. Some commonly used human brain atlases include the Allen Human Brain Atlas, the Talairach Atlas, and the Harvard-Oxford Atlas. Because most of the researchers downloaded images from pipelines like CPAC, other preprocessing methods such as brain extraction, noise reduction, etc. are performed directly through the pipeline. The second stage includes the utilization of popular and powerful state-of-the-art vision transformers with transfer learning architecture for ASD diagnosis. Once the input images are resized based on the transformer, different transfer learning-based architectures can be proposed and the transformer's hyperparameters should be specified. The structure of transfer learning can be simple using one flatten layer and an output layer or can be deep by employing several convolution layers before the output layer. The pre-trained model can be modified by replacing the last layer with a new layer that has as many neurons as there are classes in the dataset. Then, the model can be trained on the new dataset with the objective of minimizing the cross-entropy loss function. Once the model has been trained, it can be used for classifying new brain images into different classes of brain disorders. The input image is fed into the network, and the output of the last layer is interpreted as a probability distribution over all possible classes. The class with the highest probability is then considered as the predicted class for the input image. In the third stage of the architecture in Fig. 10, the output layer can serve as a binary classifier where the neuroimaging dataset has two classes ASD and TC subjects. The prediction of unseen data can be considered as ASD diagnosis whereby a decision is made as either having ASD or not. On the other hand, the architecture of using vision transformers can be used to construct *n*-classifier whereby the dataset is labeled into subtypes of ASD such as autistic disorder, Asperger's disorder, and pervasive developmental disorder [3].

5.2. Implemented VT for ASD diagnosis

5.2.1. Very deep convolutional network (VGG) models

VGG, developed by the Oxford vision team, are composed of multiple convolutional layers combined with maxpooling layer [21,22]. The VGG pre-trained models consist of a series of convolutional layers, each of which has filters of size 3x3 and a stride of 1. These convolutional layers are followed by pooling layers with a stride of 2 and a filter size of 2x2. The architecture includes 13 convolutional layers and 5 pooling layers, with the final layer being a fully connected layer with 1000 nodes (one for each class in the ImageNet dataset). It is commonly used in computer vision tasks such as image classification and object detection. VGG16 pre-trained model was utilized to extract features from 2D images generated from fMRI. These features were fed into CNN which used sigmoid activation function to classify the subjects [19]. The average 5-fold cross-validation accuracy was 83 %, using 1102 subjects (529 ASD, 573 TC) fMRI from ABIDE dataset. In another work, the last fully connected layer in VGG was replaced and re-trained with a cropped portion of sMRI that contains the corpus callosum region in the subject's brain [1]. The model was trained for ASD diagnosis with softmax function and ADAM optimizer which achieved 66 % classification accuracy. VGG with 3D convolution was proposed to classify ASD from TC based on fMRI wherein several blocks of squeeze and excitation modules were integrated with the VGG structure to boost classification performance [43]. According to their study, the squeeze and excitation module can highlight important features through the learning of channel-specific descriptors. The model named MCSE-VGG [43] was originally designed on multi-channel input to squeeze and excitability VGG. The VGG-net was used for ASD diagnosis, and the 3D convolution was replaced for 2D convolution. The architecture used batch

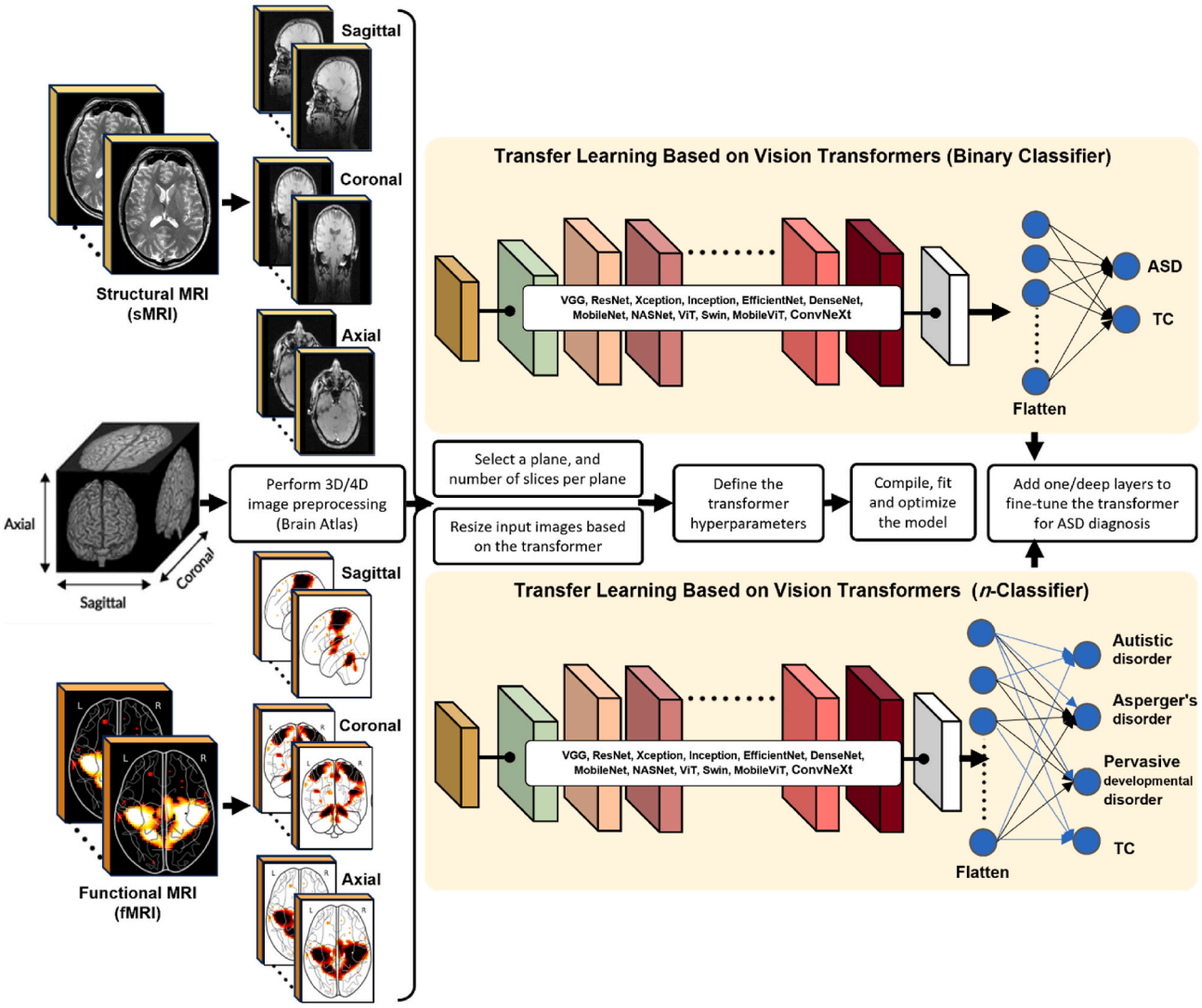


Fig. 10. Systematic architectures for ASD diagnosis and classification using different vision transformers.

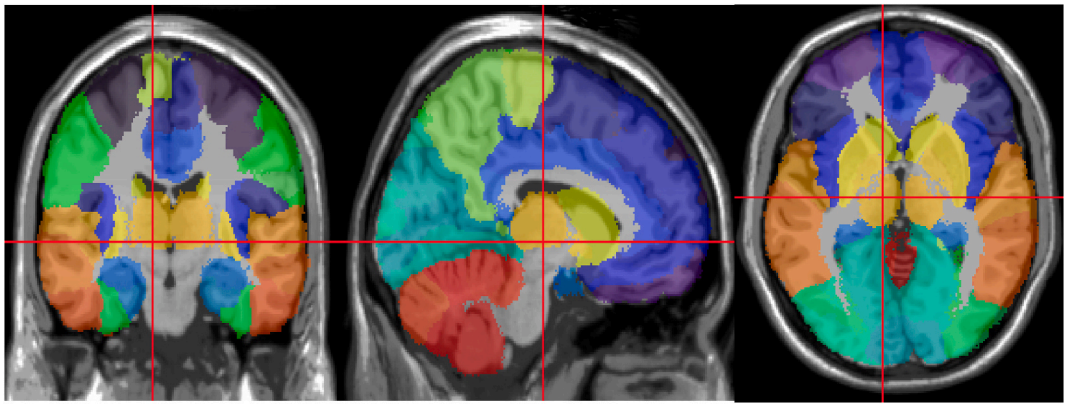


Fig. 11. Examples of MRI brain images after applying human brain atlases[5]
5 http://www.thomaskoenig.ch/Lester/ibaspm.htm.

normalization and leaky ReLU as activation functions. ADAM optimizer was initialized with the learning rate of 0.0005 in the training phase. Decay of 0.1 in the learning rate was applied if the validation loss didn't decrease after 10 epochs. LeakyReLU was set to a negative slope of 0.01 with a batch size of 12. Convolution kernel parameters were regularized using L2 norms with 1e-5 weights. The cross-entropy loss function was used to determine whether a subject belongs to the TC or ASD class. They evaluated their model with 10-cross validation and achieved 77.74 % mean accuracy. In a study referenced as [96], researchers converted 3D fMRI images into 2D brain images and used a pre-trained

VGG-16 model for classifying subjects into individuals with ASD and typically developing individuals (TCs). The study achieved a maximum accuracy of 63.4 % in accurately distinguishing between the two groups. Another study referenced as [98] proposed a concatenated CNN model for classifying sMRI images into ASD and TC subjects. This model used both VGG and ResNet models to extract features from 2D sMRI images and combined these features to improve classification accuracy. The proposed model achieved a maximum accuracy of 88.12 % in accurately diagnosing ASD.

5.2.2. Deep residual network (ResNet) models

The ResNet founded by Refs. [23,24], is constructed of residual blocks, consisting of two or three sequential convolutional layers, and a parallel identity shortcut connecting the first layer's input to the last layer's output [22]. ResNet models consist of several convolutional layers followed by residual blocks. In each residual block, the input is passed through two or three convolutional layers, and then added to the original input. This allows the model to learn the residual mapping between input and output, making it easier to train much deeper neural networks. The output of each residual block is then fed to the next one until the final output is obtained. In Ref. [42], ResNet network with five bottleneck layers was used for ASD diagnosis based on morphological features extracted from sMRI. Weights were randomly initialized and re-trained ResNet from scratch with Adam optimization method. The model evaluated using 10-fold cross-validation and achieved mean classification accuracy of 71 %. Two types of fMRI activation maps were calculated and fed into two separate classifiers, 3D ResNet-18 network and MLP for feature extraction [44]. Features were extracted from MRI using a multimodal model: ResNet-18 in the first phase, and MLP in the second phase. In ResNet-18, convolutional layers are stacked on top of each other, followed by average pooling and a fully connected layer. In the first two convolutional blocks, $1 \times 3 \times 3$ convolutional filters were also used. ReLU activation function was utilized after each layer. After completing the convolutional blocks, average pooling was performed, resulting in 512-D vectors. The model was trained from scratch using minibatch stochastic gradient descent and softmax cross-entropy loss with a batch size of 8 and momentum of 0.9. The two networks were trained independently and combined with four fully connected layers to perform the classification. Integrating both types of input features in their networks achieved classification accuracy of 74 % which was superior to the performance of training each model independently. In a study referred to as [100], researchers introduced an innovative approach where 3D cortical meshes obtained from sMRI scans were converted into 2D cortical meshes. These 2D meshes were then used as input for a pretrained ResNet model to diagnose ASD. To adapt the ResNet model for the classification task, the fully connected layer was modified. The ResNet model was trained for 125 epochs with a batch size of 64 and a learning rate of 0.01. The training process utilized a 10-fold cross-validation strategy, and the hyperparameters were optimized through grid search. The study reported a maximum accuracy of 67.7 % in diagnosing ASD using the trained ResNet model. Other research studies have also employed ResNet models for ASD diagnosis, utilizing fMRI and sMRI modal [96,98,99,104]. These studies focus on transforming MRI data into appropriate 2D representations that can be fed into ResNet models. For 3D MRI images, the standard 2D convolution layers in the ResNet architecture are replaced with 3D convolution layers, allowing for direct input of the MRI images and extraction of relevant features. The reported accuracies in these studies range from 77 % to 88 %.

5.2.3. Inception, InceptionResNet, and Xception models

GoogleNet (or Inception V1) [112] is a deep CNN architecture that was developed by researchers at Google. It was introduced in 2014 and gained popularity for its innovative design and impressive performance in image classification tasks. The key idea behind the Inception v1 architecture is the use of multiple parallel convolutional layers of different sizes within a single module, known as an inception module. This allows the network to capture information at different scales and resolutions, enabling it to learn both local and global features effectively. Inception [26,27] pre-trained model is a CNN model designed to address the problem of deciding the optimal filter size for neural networks. In traditional CNNs, the filter size is fixed and is typically chosen to be small (e.g., 3x3 or 5x5) to capture local image features. However, this approach can result in a very deep network with many parameters. The Inception model addresses this by using a combination of filters with different sizes (1x1, 3x3, 5x5) in parallel. This allows for both local and global features to be captured without the need for a very deep network. Additionally, the model uses a technique called "dimensionality reduction" that reduces the number of channels in the output of the filters to further reduce the number of parameters. The Inception pre-trained model can be used for image classification by fine-tuning it on a specific dataset. This involves taking the pre-trained model and training it on a new set of images with labels. The last few layers of the network are usually replaced with new layers to match the number of classes in the new dataset, and these layers are then trained on the new images. Once the model has been retrained, it can be used to classify new images into specific classes.

On the other hand, Xception [25] is a pre-trained image classification model that has been trained on the ImageNet dataset. The architecture of Xception is based on the Inception model, which is another popular convolutional neural network. However, Xception takes this architecture a step further by using depthwise separable convolutions, which are a type of convolutional operation that separates the process of convolution into two steps: depthwise convolution and pointwise convolution. This allows for more efficient computation and reduces the number of parameters required for training. The Xception model is made up of 36 convolutional layers and 2 fully connected layers. The input to the model is a 224x224 RGB image, and the output is a classification label for the image. The model is pre-trained on the ImageNet dataset, which contains over a million images and thousands of categories. Xception is known for its high accuracy and efficiency, as it beats ResNet accuracy results, making it one of the popular choices for image classification tasks in industry and academia. A study conducted by Ref. [95] utilized the InceptionV3 model for diagnosing ASD using 2D images obtained from raw fMRI data of both ASD and TC subjects. To adapt the InceptionV3 model for their classification task, the researchers made modifications by replacing the top layers. Two different approaches were used for training initialization: one with ImageNet weights and the other with random weights. The last dense layer was configured with the softmax function for classification. During training, hyperparameters were adjusted, including the use of the ADAM optimizer with a learning rate of 0.0001. L2 regularization techniques were employed to address overfitting concerns. The study achieved an impressive maximum accuracy of 98.35 % using a small dataset obtained from a single site within the ABIDE dataset. In another study by Ref. [97], the hybrid InceptionResNetV2 model was used for ASD diagnosis. This model combines features from both the Inception and ResNet architectures. 2D slices were extracted from raw fMRI data and used as input for the model. Pre-trained weights from the ImageNet dataset were transferred for model initialization, and a dropout rate of 0.8 was applied. The classification task involved using softmax activation for ASD/TC classification. The model was trained for 20 epochs with a batch size of 32. The study reported a maximum accuracy of 70.22 % using this approach. Additionally, a study by Ref. [104] employed the GoogLeNet pre-trained model, which is a specific instance of the Inception architecture, for ASD diagnosis. 2D image data was generated from raw fMRI scans, and the pre-trained GoogLeNet model was used to extract meaningful features for ASD diagnosis. These features were then inputted into machine learning classifiers to classify subjects into ASD and TC groups. The model achieved an impressive performance with a maximum accuracy of 80 %.

5.2.4. Densely connected convolutional network (DenseNet) models

DenseNet [28] is a pre-trained deep learning model that has been extensively used for image classification tasks. It is a CNN architecture that employs a unique strategy of connecting each layer to every other layer in a feed-forward fashion. This arrangement allows the network to learn highly discriminative features by reusing all the feature maps produced in the previous layers. The DenseNet model utilized several so-called dense layers that have been trained on the ImageNet dataset. In a study mentioned as [104], researchers utilized the DenseNet201 model for diagnosing ASD based on MRI data. They converted raw fMRI data into 2D images and used a pre-trained DenseNet model to extract representative features for ASD diagnosis. These features were then input into machine learning classifiers to classify subjects as either ASD or TCs. The model achieved a performance with accuracy, sensitivity, and specificity of 86.0 %. Another study referenced as [100] used the DenseNet-121 model for ASD diagnosis. They created 2D planar meshes from sMRI data to feed into their 2D model. To adapt the DenseNet model for the binary classification task, the fully connected layer was replaced. The training process involved 125 epochs, a batch size of 64, and a learning rate of 0.01. A 10-fold cross-validation strategy was employed to enhance training, and hyperparameters were optimized using grid search. The study found that the trained DenseNet model achieved a maximum accuracy of 67.85 % in diagnosing ASD.

5.2.5. Swin transformer

A study suggested the Swin transformer as a means of decreasing the computational expense involved in calculating attention for images with high resolution [35]. Swin's recent models have been developed by researchers at Microsoft for computer vision tasks. It is designed to handle both spatial and temporal information in an efficient way, making it suitable for tasks such as image classification, object detection, and video analysis. Swin is based on the idea of hierarchical feature extraction, where features are extracted at different scales, and then combined to form a holistic representation of the image. It uses a multi-scale transformer architecture where a series of transformer blocks are applied at different levels of the feature hierarchy. At each level, the features are split into non-overlapping patches, which are then processed by the transformer blocks. The output of each block is then fed into the next level of the hierarchy. One of the key benefits of the Swin architecture is its scalability, which allows it to handle large image sizes and input resolutions. This makes it ideal for tasks such as image classification and object detection, where high resolution images need to be processed quickly and accurately. In a study conducted by Ref. [102], Swin transformer was utilized for diagnosing ASD using fMRI data. The fMRI data underwent preprocessing and was divided into patches. These patches were then transformed into embeddings to be used as input for the Swin transformer. To improve feature extraction and classification performance, modifications were made to the Swin architecture. The hierarchical structure of the Swin transformer was leveraged to enhance features by combining output maps from different stages. This led to improved network performance without significantly increasing computational costs or training time. Additionally, a cross-patch module was designed to learn fine-grained local features by shuffling and grouping patch embeddings. The impact of the model's architecture on performance was investigated by adjusting the number of Swin Transformer blocks in different stages. The study achieved a maximum accuracy of 78 % in accurately classifying subjects into individuals with ASD and TCs.

5.3. Other VT

Numerous transformers have been recently developed and documented in the literature. In this paper, we have examined a selection of these transformers to explore the potential for utilizing them in the diagnosis of ASD. The mentioned transformers encompass BiT (Big Transfer) [37], MobileViT [36], EfficientNet [32,33], MobileNet [29,30], ViT [34], ConvNeXt [38]. The BiT transformer [37] is a type of neural network architecture, specifically a variation of the popular Transformer model. It was developed by Google Research as part of the Big Transfer (BiT) project, which aims to explore large-scale transfer learning in computer vision. The BiT transformer leverages the concept of transfer learning by pre-training on a large-scale dataset containing a diverse range of visual concepts. The pre-training process involves predicting the correct labels for a large number of images, which helps the model learn to recognize various visual patterns and features. This pre-training phase is performed on a powerful computational infrastructure, allowing the model to learn from billions of images. The key advantage of the BiT transformer is its ability to transfer learned knowledge from diverse visual concepts to new tasks. This makes it particularly useful in scenarios where labeled data may be scarce or expensive to obtain. By leveraging the pre-trained BiT transformer, developers can achieve state-of-the-art performance on various computer vision tasks with minimal data and computational resources.

MobileViT [36] is a vision transformer architecture that is designed for efficient and scalable image recognition on mobile devices. It is a modified version of the popular ViT architecture that uses transformers as convolutions. MobileViT uses a similar self-attention mechanism as ViT, but with a few modifications. First, it uses a smaller input image size to reduce the memory and compute requirements. Second, it uses an efficient convolutional layer at the beginning of the network to extract local features from the input image. Third, it uses a multihead self-attention mechanism that reduces the number of attention heads compared to ViT. MobileViT also uses distillation techniques to improve its accuracy and reduce its memory requirements. It leverages a larger teacher ViT model to distill knowledge into a smaller student MobileViT model. Overall, the MobileViT architecture is designed to offer a trade-off between accuracy and efficiency for mobile devices, with a focus on reducing the model size, computation requirements, and memory footprint.

EfficientNet [32,33] is a family of CNNs for image classification aims to achieve higher accuracy while also reducing the computational cost and number of parameters compared to traditional CNN architectures. EfficientNet models use a compound scaling method that scales all dimensions of the network in a balanced way. This means that instead of focusing solely on increasing the depth, width or resolution of the network, these dimensions are scaled together to achieve optimal performance. The architecture of EfficientNet is based on a backbone network, similar to other CNN models. However, it incorporates novel components such as the use of squeeze-and-excitation blocks, which enhance the representation power of the network and allow it to better capture the dependencies between different parts of the image. EfficientNet models have achieved state-of-the-art performance on various image classification tasks, including the ImageNet dataset. They have also been proven to work well in transfer learning, where a pre-trained model is fine-tuned on a smaller dataset for a specific task, such as object detection or segmentation.

MobileNet [29,30] is a type of neural network architecture designed for mobile devices and other low-power applications. It uses depth-wise separable convolutions, which consists of a depth-wise convolution followed by a point-wise convolution. This allows for a significant reduction in the number of parameters required for the network, and thus, a reduction in the computation required for inference and training. MobileNet pre-trained models have been trained on large datasets such as ImageNet and are commonly used for image classification tasks. These pre-trained models can be fine-tuned on new datasets, allowing for the transfer of knowledge learned from the original dataset to new tasks. For image classification tasks, the pre-trained MobileNet models take an image as input, and then output a class prediction. This makes them useful for a variety of computer vision applications, such as object detection, face recognition, and scene understanding. Additionally, MobileNet models can be optimized for specific hardware platforms, allowing for efficient deployment on mobile devices and other

low-power devices.

The Vision Transformer (ViT) [34] is a deep learning architecture which provides new insight for vision-related tasks from CNN-based state-of-the-art approaches. Although the original transformer model combines both encoders and decoders, ViT model only has encoders. It is a variant of the popular Transformer architecture used for natural language processing. ViT's architecture consists of a sequence of transformer encoder layers, followed by a pooling layer and a simple MLP for classification. ViT operates on patches of the input image, which are flattened into sequences of tokens that can be processed by the Transformer encoder layers. These tokens are fed into the model as input, rather than individual pixels or convolutional feature maps like in traditional CNN architectures. During training, the model learns to attend to relevant tokens to classify the image. The attention mechanism allows the model to focus on important regions of the image, rather than treating each token equally. One of the main advantages of ViT is its ability to handle images of varying sizes without requiring resizing or cropping. This makes it more flexible compared to traditional CNNs that require a fixed input size. ViT can also be pre-trained on large image datasets and fine-tuned for specific image classification tasks, resulting in state-of-the-art performance on several benchmark datasets. The main difference between ViT [34] and Swin [35] is how they process images. ViT processes the entire image as a sequence of patches, while Swin Transformer uses a hierarchical approach that divides the image into smaller local regions and processes them separately. Overall, both ViT and Swin have demonstrated the effectiveness of using vision transformers for image processing, but Swin Transformer offers a more advanced and flexible approach that may be more suitable for complex image recognition tasks.

ConvNeXt [38] is a deep architecture that is based on ResNet architecture designed for image classification tasks. Pre-trained ConvNeXt models are models that have been trained on a large dataset of images, and their weights have been saved. These pre-trained models can be used for various image-related tasks such as object detection, image segmentation, and image recognition. ConvNeXt pre-trained models are available in various configurations such as different number of layers, different input sizes, and different training datasets. Popular pre-trained ConvNeXt models are ResNeXt-101 and ResNeXt-152, which have been trained on the ImageNet dataset containing over a million labeled images. The architecture of ConvNeXt is composed of three main components. First, the input image is fed into a series of convolutional layers, which extract features from the image. Second after the convolutional layers, the network has a transition module, which consists of a 1x1 convolutional layer followed by a pooling layer. The 1x1 convolutional layer is used to reduce the number of feature maps, and the pooling layer is used to reduce the spatial dimensions of the feature maps. Finally, the output of the transition module is fed into a series of fully connected layers, which perform the actual classification of the input image. ConvNeXt architecture also includes the use of residual connections, which improves the overall accuracy.

Transformers like BiT (Big Transfer) [37], MobileViT [36], EfficientNet [32,33], MobileNet [29,30], ViT [34], ConvNeXt [38] can be used for ASD diagnosis and classification by leveraging their powerful image recognition capabilities. As transformers are typically pretrained on large-scale datasets, transfer learning can be applied. The knowledge gained during pre-training significantly improves the model's performance on these downstream tasks, even with less training data. The transformer model can be fine-tuned on the ASD-specific dataset. Fine-tuning involves training the model on the labeled images from the ASD dataset to learn the specific patterns and features relevant to ASD diagnosis. This step helps the model adapt to the ASD domain. Once the model is trained, it can be used to classify new, unseen images. Given an input image, the model predicts whether the person is likely to have ASD or not. The model's output can be a probability score or a binary classification. Additional iterative improvements can be applied if the performance is not satisfactory, by adjusting hyperparameters, adding more training data, or exploring different transformer models. It is important to note that the success of ASD diagnosis using transformers heavily relies on the availability of a well-labeled data through different data representations discussed earlier in Section 3.3.

To conclude about VT in relation to ASD research, Table 6 presents a summary of vision pretrained transformers described including VGG [21,22], ResNet [23,24], Xception [25], Inception [26,27], DenseNet [28], MobileNet [29,30], NASNet [31], EfficientNet [32,33], and large transformer-based models such as ViT [34], Swin [35], MobileViT [36], BiT [37], and ConvNeXt [38] transformers. This review paper discusses the potential of vision transformers in the field of brain disorder classification, particularly for ASD diagnosis, using brain MRI image datasets. VT with their deep neural network learning capacity, serve as powerful tools for brain disorder classification tasks, enabling the recognition of complex patterns in brain images that may indicate the existence of ASD. The models are trained on a large dataset of labeled brain images representing different brain disorders. Once trained, these models can be used to predict the presence or absence of ASD in new brain images. By passing the new images through the VT model, it can generate probability scores for each class. Based on the highest probability score, the model classifies the new brain image as either a specific ASD or as normal/healthy. This approach has shown promising results in recent studies for identifying and classifying ASD.

6. Brain transformers (BT) for ASD diagnosis

6.1. Architectures of BT for ASD diagnosis

Brain transformers (BT) are created using deep learning techniques, which involve training neural networks on large datasets of brain scan images, such as computed tomography (CT), diffusion tensor imaging (DTI), positron emission tomography (PET), and MRI, for general brain-related tasks. These BT models can be applied to analyze brain imaging data and classify brain disorders. The emerging trend of BT models has the potential to greatly enhance the diagnosis of complex and time-consuming brain disorders like ASD by utilizing MRI modalities. ASD diagnosis and classification often rely on analyzing MRI scans to identify specific brain patterns and abnormalities associated with the disorder. By training BT models specifically on MRI images, they can capture and learn relevant features and representations that are crucial for distinguishing individuals with ASD from those without ASD. One advantage of using BT models for ASD diagnosis is that they can be pretrained on datasets consisting of MRI scans with similar features. Unlike VT pretrained on ImageNet, which contains a diverse range of object classes and visual features, BT models can be pretrained on datasets that are more tailored to the target task. This allows the model to learn representations that are specifically suited to the characteristics of brain imaging data, potentially leading to improved performance in ASD diagnosis and classification.

To build a BT, researchers begin by collecting a large dataset of brain scan images that represent different types of brain disorders. These images are then labeled with information about the specific disorder they represent. Once the datasets are collected and labeled, the researchers use DL techniques to train a neural network to recognize patterns in the data and classify the images into different brain disorders. Once the BT has been trained, it can be fine-tuned to analyze new brain scan images and classify them into different disorders based on the patterns it has learned. This can be a useful tool for diagnosing and monitoring brain disorders, as it can provide automated and accurate classifications that can help clinicians better understand and treat their patients. Another on-hand method to build a brain transformer is based on one of the vision transformers that is fine-tuned on neuroimaging data. Using such a method, different transformer-based architecture pretrained on MRI data and designed for ASD diagnosis have been proposed recently including standard transformer-based model [113], BN-Transformer [93], METAFormer [114], Com-BrainTF [92],

Table 6
Summary of vision transformers and pre-trained models for image classification tasks and its relation and usage for ASD diagnosis.

Model	Variations	Min- Max Size (MB)	Performance on ImageNet		Input pre-processing	Input shape	Source	Used for ASD
			Top-1	Top-5				
VGG	VGG16, VGG19	528–529	71.3 %	90.1 %	convert input images from RGB to BGR, zero-center each color channel with no scaling	224x224	[21,22]	[1,19,43, 96,98]
ResNet	ResNet50, ResNet50V2, ResNet101, ResNet101V2, ResNet152, ResNet152V2	98-232	78.0 %	94.2 %	convert input images from RGB to BGR, zero-center each color channel with no scaling	224x224	[23,24]	[42,44,96, 98–100, 104]
Xception	Xception	88	79.0 %	94.5 %	scale input pixels between −1 and 1	299x299	[25]	×
Inception	InceptionV3, InceptionResNetV2, GoogLeNet	92–251	80.3 %	95.3 %	scale input pixels between −1 and 1	299x299	[26,27, 112]	[95,97, 104]
DenseNet	DenseNet121, DenseNet169, DenseNet201	33–80	76.2 %	93.6 %	–	224x224	[28]	[100,104]
MobileNet	MobileNet, MobileNetV2	14–16	71.3 %	90.1 %	scale input pixels between −1 and 1	224x224	[29,30]	×
EfficientNet	EfficientNetB0, EfficientNetB1, EfficientNetB2, EfficientNetB3, EfficientNetB4, EfficientNetB5, EfficientNetB6, EfficientNetB7. EfficientNetV2B0, EfficientNetV2B1, EfficientNetV2B2, EfficientNetV2B3, EfficientNetV2S, EfficientNetV2M, EfficientNetV2L	29–479	85.7 %	97.5 %	inputs to be float tensors of pixels with values in range [0–255]	Optional shape tuple	[32,33]	×
ViT	base – hybrid – large - huge	–	88.55 %	–	inputs to be float or uint8 tensors of pixels with values in range [0–255]	224x224 256x256 384x384	[34]	×
Swin	base – small – tiny (v1 & v2)	87.3 %	98.2 %	–	inputs to be float or uint8 tensors of pixels with values in range [0–255]	v2: 256x256 v1: 384x384	[35]	[102]
MobileViT	Lightweight, general-purpose, and mobile-friendly	–	78.4 %	–	inputs to be float tensors of pixels with values in range [0–255]	Optional shape tuple	[36]	×
ConvNeXt	ConvNeXtTiny, ConvNeXtSmall, ConvNeXtBase, ConvNeXtLarge, ConvNeXtXLarge	109–1310	86.7 %	–	inputs to be float or uint8 tensors of pixels with values in range [0–255]	Optional shape tuple	[38]	×
BiT	BiT-L model: state-of-the-art transfer learning method for image classification	–	87.5 %	98.6 %	Normalize the pixel values of the image to a standardized range	224x224	[37]	×

ST-Transformer [57], STCAL [105], BolT [106], and BrainFormer [115]. Additionally, several brain-related disorders transformers have been proposed such as MSGTN [116], STAGIN [117], and the medical transformer [118]. BT have shown promise in ASD diagnosis, with high accuracy depending on the task requirements and available data. Fig. 12 illustrates a comprehensive framework for using brain transformers in ASD diagnosis, which is similar to the framework discussed earlier with VT. The main difference lies in the structure of the top layers used. In the recommended framework, since BTs have already been trained on a large neuroimaging dataset, zero-shot learning can be employed. Zero-shot learning is an algorithmic approach where a model learns to recognize objects or classes it has never encountered before based on the semantic relationships between different classes. As BT have been trained to recognize images of brain disorders, they can also identify images of TC and ASD, even without specific training on those images, by leveraging the shared semantic relationship of "the brain". A few additional layers can be added to the BT for training using transfer learning techniques, as explained earlier, specifically tailored for ASD diagnosis. This ensures that the model adapts to the unique characteristics and patterns of ASD in brain images. Table 7 summarized BTs for ASD and brain-related applications.

6.2. Transformer-based models designed for ASD diagnosis

6.2.1. Standard transformer-based model

The researchers in Ref. [113] conducted a study where they proposed a transformer-based model for analyzing fMRI data. Their model utilizes the FC matrix for positional decoding, enabling the analysis of relationships between different brain regions. The architecture of the model is similar to ViT-B/16 and includes patch embedding, a transformer encoder, and an MLP head. Patch embedding divides the segmented sample into patches and extracts features using 3D-CNN. The transformer encoder utilizes self-attention to analyze relationships between features, and the MLP head employs fully connected layers to obtain the classification result. The performance of the model was evaluated on the ABIDE dataset, achieving a classification accuracy of 74.18 % through 10-fold cross-validation.

6.2.2. Brain Network Transformer

The Brain Network Transformer, also known as BN-Transformer, was introduced by Ref. [93] as a method for analyzing brain network data. It aims to predict certain characteristics of brain subjects, such as their biological sex or the presence of a disease, using graph-based representations of fMRI brain networks. The framework consists of two main components: the multi-head self-attention (MHSA) module and the orthonormal clustering readout (OCREAD) operator. In the MHSA module, the model learns to enhance node features through attention mechanisms by mapping the brain network to node embeddings. The initial node features capture positional information based on the connections between nodes. The attention mechanism in the MHSA module calculates attention scores without considering edge weights or relative positions in dense brain networks. Multiple non-linear mappings are used to generate expressive node features. The OCREAD operator generates embeddings at the graph level by utilizing the similarity between

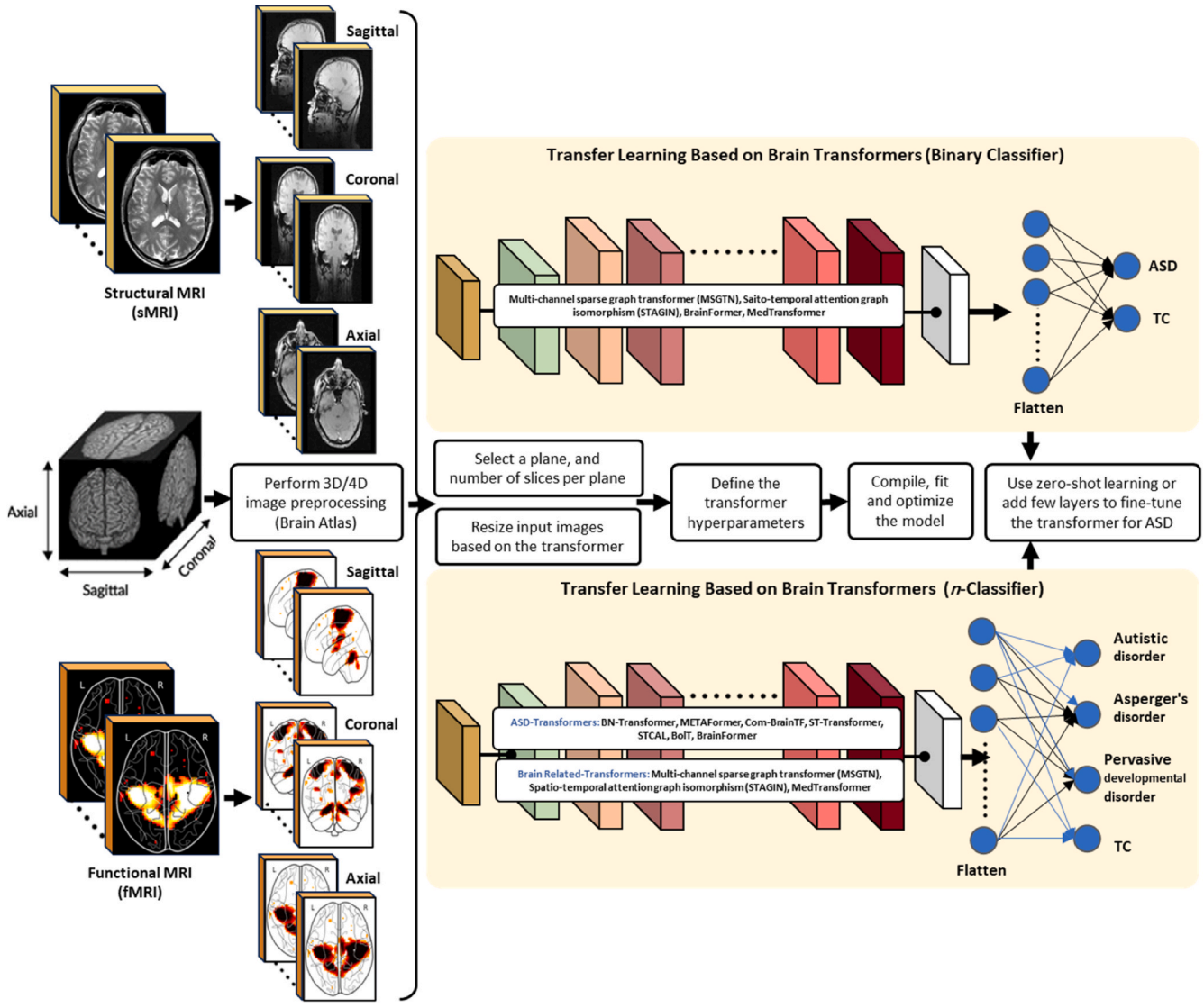


Fig. 12. Systematic architectures for ASD diagnosis and classification using recently developed brain transformers.

nodes within functional modules. Soft assignments to clusters are calculated using a softmax projection operator. Graph-level embeddings are obtained by aggregating node embeddings based on the cluster information. The performance of the proposed Brain Network Transformer is evaluated on the ABIDE dataset, achieving a maximum accuracy of 71 % in diagnosing ASD.

6.2.3. *METAFormer*

METAFormer was introduced as a method for classifying ASD using fMRI data [114]. The framework was evaluated on the ABIDE I dataset, which consisted of 406 ASD subjects and 476 typically developing (TC) subjects. METAFormer employed a multi-atlas approach, where flattened connectivity matrices from three different atlases (AAL, CC200, and DOS160) were used as input to the transformer encoder. In their study, the transformer architecture only utilized the encoder component and did not make use of the decoder part. To balance the impact of input features, each input to the transformer underwent embedding and scaling operations. The embedded input then underwent processing through a transformer encoder with multiple layers, attention heads, and feed-forward units, following a BERT-style architecture. The final output was obtained using linear layers and a softmax activation. The proposed METAFormer framework surpassed the performance of existing methods on the ABIDE I dataset, achieving an average accuracy of 83.7 % and an AUC-score of 83 %.

6.2.4. *Com-BrainTF*

The Com-BrainTF transformer was developed for analyzing brain connectomes in the context of ASD diagnosis [92]. This approach utilizes a hierarchical local-global transformer architecture. It incorporates functional information into the transformer encoder by integrating community-aware ROI embeddings. Its framework consists of a local transformer and a global transformer. The local transformer learns community-specific embeddings by leveraging ROI-level information. It employs shared parameters across communities and personalized prompt tokens to differentiate embedding functions. The outputs of the local transformer, including class tokens and node embeddings, are then fed into the global transformer. The global transformer combines community-specific node embeddings and prompt tokens to capture inter-community dependencies. Both the local and global transformers utilize a transformer encoder with a multi-head attention mechanism to capture the relationships between nodes. Finally, a graph readout layer is employed to aggregate global-level node embeddings and generate a high-level representation of the brain graph. When evaluated on the ABIDE I dataset, Com-BrainTF transformer achieved a maximum accuracy of 72.50 % for ASD diagnosis.

Table 7
Summary of brain vision transformers for different ASD classification.

Category	Transformer	Task	Dataset	Modalities	No. subject	Max Accuracy	Ref.
Pre-trained vision transformers used for ASD	VGG	ASD Classification	ABIDE I	2D Slice fMRI	1102	83 %	[19]
			ABIDE I	fMRI Connectivity	184	77.74 %	[43]
			ABIDE I + II	2D Slice fMRI	2206	63.4 %	[96]
			ABIDE I	2D Slice sMRI	1112	88.12 %	[98]
	ResNet	ASD Classification	ABIDE I	fMRI Connectivity	1035	74 %	[44]
			ABIDE I	sMRI Connectivity	1085	71 %	[42]
			ABIDE I	3D sMRI	500	75 %	[99]
			ABIDE I	2D Slice sMRI	1112	88.12 %	[98]
	Inception	ASD Classification	ABIDE I	fMRI Time-series	82	80 %	[104]
			ABIDE I + II	2D fMRI	138	98.35 %	[95]
			ABIDE I	2D fMRI	172	70.22 %	[97]
	DenseNet	ASD Classification	ABIDE I + II	3D sMRI	2206	67.85 %	[100]
	Swin	ASD Classification	ABIDE I	fMRI Regions	800	78 %	[102]
Transformer-based architecture models designed for ASD	METAFormer	ASD Classification	ABIDE I	fMRI Connectivity	882	83.70 %	[114]
	Com-BrainTF	ASD Classification	ABIDE I	fMRI Graph	1009	72.50 %	[92]
	BN-Transfomer	ASD Classification	ABIDE I	fMRI Graph	1009	71 %	[93]
	ST-Transformer	ASD Classification	ABIDE I + II	fMRI Time-series	1009	71 %	[57]
	Standard transformer	ASD Classification	ABIDE I	fMRI Connectivity	871	74.18 %	[113]
	STCAL	ASD Classification	ABIDE I + II	fMRI Time-series	1035	73 %	[105]
	BolT-Transformer	ASD Classification	ABIDE I	fMRI Time-series	871	71.28 %	[106]
	BrainFormer	Brain Disease Classification (MDD, AD, MCI, ASD, AS, ADHD, MD, and MOH)	ABIDE, ADNI, MPILMBB, ADHD-200, and ECHO	3D fMRI	5015	97.2 %	[115]
Brain-related disorder transformers	MSGTN	AD Classification	ADNI	fMRI Connectivity DTI	170	92.12 %	[116]
	STAGIN	Gender Classification	HCP-Rest HCP-Task	fMRI Graph	8543	99.19 %	[117]
	Med-Transformer	Brain disease diagnosis, brain age prediction, and brain tumor segmentation	IXI, Cam-CAN, and ABIDE	3D sMRI	1550	83 %	[118]

Task includes Major Depressive Disorder (MDD), Alzheimer's Disease (AD), Mild Cognitive Impairment (MCI), Autism Spectrum Disorder (ASD), Asperger's Syndrome (AS), Attention Deficit Hyperactivity Disorder (ADHD), Migraine Disease (MD), Medication-Overuse Headache (MOH).

6.2.5. ST-transformer

ST-Transformer was suggested as a method for distinguishing individuals with ASD from typically developing (TCs) using fMRI data [57]. This transformer model incorporates a linear spatial-temporal multi-headed attention unit to capture both spatial and temporal representations of the fMRI data. The structure of the ST-Transformer consists of a transformer-encoder that includes a multi-headed self-attention module, a position feed-forward network, residual connectivity, and layer normalization. The self-attention mechanism is replaced by a linear spatial-temporal multi-headed attention unit. This unit performs spatial self-attention followed by temporal self-attention to obtain a representation of the spatial-temporal features in the fMRI data. To accelerate the training process, a linear attention module replaces the conventional self-attention unit in the ST-Transformer. The ST-Transformer was evaluated on a large dataset consisting of subjects from two independent datasets (ABIDE I and ABIDE II). It achieved robust accuracies of 71.0 % and 70.6 % for ASD diagnosis on the ABIDE I and ABIDE II datasets, respectively.

6.2.6. STCAL

The STCAL (Spatial-temporal co-attention learning) transformer was proposed by Ref. [105] as a method for diagnosing ASD and Attention Deficit Hyperactivity Disorder (ADHD) using fMRI data. The STCAL model consists of a guided co-attention network that incorporates self-attention, guided co-attention, and sliding cluster attention modules. The self-attention module, inspired by the transformer's encoder layer, utilized multi-head attention and pointwise feedforward layers to capture correlations within the feature representation. The guided co-attention module focuses on modeling interactions between spatial and temporal features. To overcome the limitations of global attention, the sliding cluster attention module was introduced. This module enabled the detection of short-time local signals in the fMRI data by constructing independent self-attention layers based on sliding window clusters. The local attention representations were then aggregated using a sliding fusion block. The sliding cluster attention module significantly enhanced the model's capability to capture similar short-time local signals in the data. The STCAL model achieved competitive accuracies on different datasets, with accuracies of 73.0 % on the ABIDE I dataset, 72.0 % on the ABIDE II dataset, and 72.5 % on the ADHD-200 dataset.

6.2.7. BolT transformer

The research paper [106] introduced BolT, a transformer model specifically designed for analyzing multivariate fMRI time series. It utilized a cascade of transformer encoders and a fused window attention mechanism to process the data efficiently. BolT begins by encoding temporally overlapped windows within the fMRI time series, capturing local representations. It then incorporates cross-window attention, which integrates information from neighboring windows by considering both the base tokens within each window and the fringe tokens from adjacent windows. To capture both local and global patterns effectively, BolT gradually increases the extent of window overlap and the number of fringe tokens in its cascade. This approach allows the model to capture fine-grained details as well as broader patterns in the data. Additionally, BolT incorporates cross-window regularization to align

high-level classification features across the entire time series. This regularization promotes consistency and reduces noise in the analysis. The performance of the BolT transformer model was evaluated on the ABIDE I dataset and achieved a maximum accuracy of 71.28 %.

6.2.8. BrainFormer

BrainFormer was proposed to develop a universal brain disease classification model based on a single MRI volume that does not require ROI labeling [115]. The researchers tested their model using independently acquired datasets, including ABIDE, ADNI,[4] MPILMBB,[5] ADHD-200,[6] and ECHO that were collected in their study. These datasets have several disorders, including major depressive disorder (MDD), Alzheimer's disease (AD) and mild cognitive impairment (MCI), ASD and Asperger's syndrome (AS), attention deficit hyperactivity disorder (ADHD), migraine disease (MD) and medication-overuse headache (MOH). After collecting images from different sites, the MRI images were normalized to remove inconsistencies. To extract features, a 3D convolution layer was used as a backbone. Then, shallow global attention (SGA) and deep global attention (DGA) were used as attention blocks. The SGA block was created to enable the exchange of information across shallow layers through fully connected layers, while the DGA block's global attention mask was designed specifically for deep layers to precisely combine global information. In unified framework for 3D MRI volume classification, the two types of global attention blocks were combined. Softmax function was used to map each brain MRI feature into normal and diseased classes. Moreover, gradient-based localization-map visualization approach was proposed for finding possible brain biomarkers for identifying the disease. Through this visual technique, class-dependent saliency maps were calculated from gradients back-propagated from specific classes, further enhancing the learnable feature's discriminability. According to the BrainFormer's developers, the model can capture both local and distant feature correlation at various levels, thanks to its specially designed SGAs and DGAs. Additionally, it uses gradient-based localization-map visualization to restore the original information within MRI volume, allowing to locate disease-related biomarkers accurately and facilitating the diagnosis of brain diseases. In brain-related transformers such as MSGTN [116] and STAGIN [117], MRI images were used as graphs, while the BrainFormer [115] used raw 3D MRI images as input without any feature extraction. To construct a graph based on MRI images, the image features that can be represented as nodes and the edges between them as connections were defined. The Harvard Oxford (HO) [119] brain atlas was utilized to parcellate the MRI into ROIs. These ROIs were treated as nodes in a graph, with edges defined based on the connectivity between ROIs. Using Pearson's correlation coefficient, FC between ROIs were computed by extracting the mean time series of each ROI.

6.3. Brain-related disorder transformers

6.3.1. MSGTN

MSGTN transformer was proposed for Alzheimer's disease (AD) classification [116]. They utilized MRI and DTI provided by public AD Neuroimaging Initiative (ADNI).[7] In their study, the image (i.e., MRI and DTI) and non-image information (i.e., gender and age) for each subject were obtained and combined using a graph. Locally weighted clustering coefficients were used to extract the image information from MRI and DTI. Features were constructed as FC and SC using Pearson's correlation factors and white fiber binding between each pair of brain regions of DTI. Locally weighted clustering coefficients were used to fuse and stack the multimodal features to form a brain network. The graph is constructed by representing the non-image information as edges and the image information as vertices. The resulting graph is fed into Graph Transformer Network (GTN) to perform the classification. MSGTN consists of stacked multiple graph transformer layers, graph convolution layer, pooling layer, and softmax function to perform the classification.

6.3.2. STAGIN

STAGIN transformer [117] was developed to address the challenge of learning dynamic brain connectome graph representations and provide temporal analysis of data. The transformer was applied to MRI neuroimaging data collected from the Human Connectome Project (HCP)[8] for gender classification. STAGIN employed graph neural networks (GNNs) to operate on dynamic FC features, formalized as mapping graphs sequence $G_{dynamic} = (G(1), \ldots, G(T))$ with T times/steps to an embedding sequence called $h\_G_{dynamic}$. Then, two-step embedding was generated using a graph isomorphism network and transformer encoder combined with self-attention. Their network employed a four-layer GNN to process shifted windows across the time series. In this approach, ROIs were treated as nodes, with edges derived from FC features, and node features calculated using recurrent networks. Graph features were then extracted from each graph using a squeeze-excitation readout module and consolidated into a single latent feature using a transformer. Latent features were linearly projected to class logits. FC features were computed using Pearson's correlation with a window size of 50. Hyperparameters used for cross-validation included learning rate of 2×10^{-4}, 40 epochs, and batch size of 8. There are claims that the transformer has not only increased the classification performance of models but also improved their interpretability in spatial and temporal dimensions. As a result, more attention is being given to using transformer-based approaches for analyzing features and relationships in intricate graphs.

6.3.3. Medical transformer

The Medical Transformer, or shortly MedTransformer, is a deep learning model designed to analyze 3D MRI images of the brain using a universal encoder-decoder architecture [118]. The model employed a novel transfer learning framework by extracting meaningful features from the MRI images and encoding them into a compact representation that can be used for a variety of downstream tasks, such as regression, classification, and segmentation. The model also employed attention mechanisms to selectively focus on the most relevant regions of the brain for each task, making it more efficient and accurate than traditional models. MedTransformer has been trained on a large and diverse set of brain MRI images including Information eXtraction from Images (IXI), Cambridge Center for Ageing and Neuroscience (Cam-CAN), and ABIDE. This pre-training helps the model to capture robust and meaningful features that are transferable across datasets. MedTransfomer was evaluated on three brain-related tasks namely brain disease diagnosis, brain age prediction, and brain tumor segmentation, and it achieved state-of-the-art results: 92 % of classification accuracy for brain diseases, 92 % of regression accuracy for age prediction, and 97 % for brain tumor segmentation. MedTransformer uses a comprehensive method by dividing a 3D volumetric image into 2D slices from three different planes (sagittal, coronal, axial) and utilizing these slices as inputs for the network. Three stages were shown. Firstly, a convolutional encoder using a sample network was pre-trained. Secondly, a transformer for a masked encoding vector prediction acting as a self-supervised learning (SSL) proxy task was pre-trained. Thirdly, the backbone prediction network was fine-tuned for medical tasks.

[4] https://adni.loni.usc.edu
[5] https://www.neuroconnlab.org/data
[6] http://fcon_1000.projects.nitrc.org/indi/adhd200
[7] https://adni.loni.usc.edu
[8] http://www.humanconnectomeproject.org/data

7. Do it the transformer way: vision and brain transformers

The direction of previous research on ASD diagnosis which emphasized ML methods especially SVM and MLP, have almost discontinued. Comparative data of VT methods presented in Table 7 include pretrained vision transformers, VGG, ResNet, Inception, and DenseNet for ASD diagnosis. The evaluation was performed on different datasets, such as ABIDE I and ABIDE II, using various types of brain imaging data, including 2D/3D fMRI, sMRI, and connectivity data. VT's performances were encouraging as they ranged from 70% to 88 %. We ignored the highest accuracy results (98 % in Ref. [95]) as this could be results of data leakages of using MRI slices [120].

Prominent role has been played by transformers in the processing of sequence-to-sequence problems in natural language processing (NLP) [20]. By abandoning the sequence-dependence characteristic of Recurrent Neural Networks (RNN), transformers' success is largely determined by the use of multi-head attention, which is the stacking and integration of multiple layers of self-attention [121]. A self-attention layer is applied to the input sequence to capture the internal correlation of data or features, thereby reducing the reliance on external data. Additionally, it allows unit-level global information (long-term dependencies) to be captured, which greatly facilitates the development of time series data processing models. Transformers are efficient since they solve sequential learning problems in one shot, as well as effective because they require a lot less time for training [121]. Image processing and recognition have investigated many transformer architectures [34, 35]. Because CNNs have a limitation in ignoring long-term relationships between objects in the image, studies have shown that adding the attention mechanism, which has been successful in NLP, to CNNs capture these long-term dependencies and improve image classification accuracy by treating it as a sequence prediction task [35,122]. Recently, computer vision has begun to explore transformer architectures for image classification. There is a shortage of research on transformer-based models in classifying individuals with ASD despite their recent success in various natural language processing tasks. Using transformer models, we can overcome the complexities of ASD diagnosis, including its heterogeneity and limited availability of data. As a result, we concluded that using transformers can provide valuable insights into ASD diagnosis, especially when they are trained on brain images. This can serve as a new contribution to the literature on ASD diagnosis based on MRI neuroimaging data.

Referring to Tables 7 and it demonstrates the performance of different transformer-based models in ASD classification using various fMRI data types and datasets. It highlights the range of accuracies achieved and the potential of these models in improving our understanding and diagnosis of ASD. Each model was evaluated on different datasets and utilizes different types of fMRI data, such as connectivity matrices, graph representations, or time-series. METAFormer achieved an accuracy of 83.7 % on the ABIDE I dataset using fMRI connectivity data [114]. Com-BrainTF, on the other hand, achieved a slightly lower accuracy of 72.5 % on the same dataset using fMRI graph data [92]. Brain Network Transformer also used fMRI graph data from ABIDE I and achieved an accuracy of 71 % [93]. ST-Transformer achieved an accuracy of 71 % on the ABIDE I + II datasets using fMRI time-series data [57]. One of the basic transformer-based architecture achieved an accuracy of 74 % on ABIDE I using fMRI connectivity data [113]. STCAL, which also used fMRI time-series data from ABIDE I + II, achieved an accuracy of 73 % [105]. BolT-Transformer achieved an accuracy of 71.28 % on ABIDE I used fMRI time-series data [106]. In contrast to the other models, BrainFormer focused on classifying multiple brain diseases, including ASD, using 3D fMRI data from various datasets. It achieved an impressive accuracy of 97.2 % on a combined dataset [115].

Additionally, brain-related disorder transformers MSGTN [116], STAGIN [117], and MedTransformer [118] for brain related problems and their corresponding classifications can be further extended for ASD. Their results have shown great impact: MSGTN transformer achieved classification accuracies of 83 % [118], 92 % [116], and 99 % [117]. These transformers utilize various imaging techniques and demonstrate promising results in terms of classification accuracy for different brain-related disorders. We believe that they can be further fine-tuned on ASD neuroimaging data.

To such extent, this review study considers VTs as general classification models, and BTs as specific-purpose brain disorders classification models. The direction of developing BTs is promising and can initiate new trends and contributions for ASD diagnosis. We have proposed two systematic frameworks in Figs. 10 and 12 to pave the way for future research and investigation in this field. In order to implement these frameworks, the input images need to undergo preprocessing based on the specific shape requirements of the transformer being used. Typically, the images should be adjusted to have an average value of zero and a standard deviation of one. Additionally, the images should be converted from a single-channel grayscale MRI image to a three-channel input, where the grayscale image is duplicated in the other two channels. After preprocessing, the input MRI images are resized to the appropriate resolution specified by the transformer's input shape. The dataset is then divided into patches, with each image patch being flattened and fed into a linear embedding model. To fine-tune the vision/brain transformers (VT/BT) for the downstream task of classifying ASD into two categories (TC and ASD), or four categories (TC, autistic disorder, Asperger's disorder, and pervasive developmental disorder), the top layer of the base transformer should be removed. This can be achieved through either transfer learning or zero-shot learning techniques. To obtain the final output, the softmax function is applied to estimate the class probabilities for each subject in the classification task.

8. Challenges

Using vision transformers (VT) and brain transformers (BT) for the diagnosis and classification of ASD comes with several challenges. One challenge is the categorization of subjects into ASD and TC groups in the available datasets. The diagnostic criteria used in different studies can vary, leading to classification difficulties. Variations in diagnostic standards across settings and clinical practices contribute to heterogeneity within the ASD group, making it challenging for transformer models to identify distinct patterns that differentiate individuals with ASD from TC subjects.

Using MRI modalities, such as structural MRI (sMRI) and functional MRI (fMRI), for ASD diagnosis also presents challenges. The discernible differences between brain imaging data of individuals with ASD and others are still under investigation. Using raw brain imaging images as inputs to VT/BT without segmentation or explicit data representation raises concerns about feasibility and accuracy. Specific brain regions, such as the prefrontal cortex, amygdala, and corpus callosum, have been identified as potential biomarkers for ASD using sMRI. On the other hand, fMRI data is commonly used to visualize brain activity or connectivity patterns. However, the performance of VT/BT in ASD diagnosis solely based on raw brain imaging images remains uncertain.

The need for an effective MRI data representation space is another challenge when using VT/BT for ASD diagnosis. While VT/BT excel at capturing global information and long-term dependencies, incorporating local context from visual data, such as MRI images, can be challenging. This limitation may hinder the comprehensive understanding of the data and potentially affect the accuracy of ASD diagnosis. To address this, researchers have explored combining the transformer architecture with convolutional layers to leverage both local and global information simultaneously.

Interpretable predictions pose another critical challenge when using VT/BT for ASD diagnosis. Although attention maps have been used effectively to identify significant regions in medical images, the attention maps generated by transformers can lack clarity, leading to inaccurate associations with tokens. This vagueness of attention maps makes it challenging to establish reliable associations between attention

weights and specific regions within the image, potentially resulting in misleading or unreliable predictions.

Furthermore, pretrained vision and brain (VT/BT) transformers have specific requirements for 2D input data, while MRI data is typically collected in 4D or 3D formats. Researchers must explore techniques to transform 4D/3D MRI data into a 2D format or update transformer architectures to handle 3D input using 3D convolution layers. However, updating the architecture may result in the loss of pretrained weights and require additional computational resources. Careful consideration of this challenge is important to weigh the trade-offs between transforming MRI data or updating the transformer architecture.

9. Conclusion and future trends

ASD is a neurological disorder characterized by communication difficulties, impaired social relationships, and repetitive behaviors in children. Clinicians rely on various methods to detect ASD, including neuroimaging techniques such as magnetic resonance imaging (MRI). MRI-based structural and functional modalities are commonly used to diagnose ASD, as they provide valuable insights into the brain's structure and function. However, accurate diagnosis using MRI modalities can be time-consuming and challenging. There are factors that may contribute to misdiagnosis in ASD using MRI, such as fatigue or external noise during imaging. To improve the quality of care for individuals with ASD, it is crucial to enhance the diagnostic capabilities of existing tools, enabling faster and earlier intervention. This can be achieved by reducing diagnosis time or improving the accuracy of predictions. This review work summarizes the field of ASD diagnosis based on MRI neuroimaging datasets, covering more than 78 papers published between 2020 and 2023. Many studies in the literature have utilized MRI data from the ABIDE dataset for ASD diagnosis. Most of these studies have focused on using functional MRI (fMRI), while only a limited number have employed structural MRI (sMRI) for classification purposes. Additionally, many researchers have predominantly relied on a small number of subjects from a single site within the ABIDE repository for their classification models. To ensure more reliable and generalizable results, there is a need to explore the combination of multiple data types, such as MRI-phenotypic information, for ASD diagnosis. Incorporating phenotypic information such as age, weight, and family history can enhance the features used for classification and lead to more promising outcomes.

Few studies, including references [54,67,89], have investigated the use of MRI modalities in combination with phenotypic data for diagnosing ASD. Additionally, multimodal neuroimaging, such as MRI-EEG [123,124], has been shown to be crucial in clinical studies for ASD diagnosis. Some studies, like reference [85], have explored the potential of integrating MRI modalities with genetic data for diagnosing ASD. By combining datasets with different modalities, new AI-based methods for ASD diagnosis can be developed. Connectivity techniques are important for extracting features from structural and functional neuroimaging data. It would be beneficial to propose new feature extraction methods based on connectivity for different neuroimaging modalities. Recent studies have predominantly focused on extracting connectivity matrices as representations for each subject and using them as inputs for AI models. However, there is limited exploration by researchers regarding the direct utilization of MRI modalities as image data in the context of ASD diagnosis. ASD datasets available for research are typically collected from various clinical sites, resulting in variations in image acquisition parameters such as resolution, contrast, and scanning techniques. These discrepancies can pose challenges for traditional deep learning models to effectively learn patterns from image data across different sites.

The use of vision transformer-based architectures shows great potential in identifying reliable biomarkers for ASD. Vision transformers (VT) and brain transformers (BT) have demonstrated promising results in diagnosing ASD by discovering consistent biomarkers. The attention mechanisms in these models enable them to focus on relevant regions of interest in an image. Future trends involve developing specialized attention mechanisms tailored to analyze specific brain regions or connectivity patterns in MRI images. These mechanisms can aid in identifying crucial brain regions or networks associated with ASD, thus enhancing diagnostic accuracy. The review also highlights the importance of transfer learning and zero-shot learning and emphasized them as important techniques that can improve the performance of ASD diagnosis models. Leveraging pre-trained vision transformers and brain transformers on datasets rich in medical information and neuroimaging MRI data is an intriguing development that has the potential to enhance the accuracy of ASD and other brain disorder diagnoses. The review also explores various transformer models that employ transfer learning principles to extract pertinent features from brain MRI images. Promising research directions for the automated diagnosis of ASD using MRI modalities include investigating graph theory, representation learning, transfer learning, zero-shot learning, and Q-learning techniques. These approaches offer promising avenues for gaining valuable insights and advancements in ASD diagnosis using MRI neuroimaging data.

Funding

None.

Declaration of competing interest

The author declares that they have no known conflict of interests, and no competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The researchers would like to acknowledge Deanship of Scientific Research, Taif University, Saudi Arabia for awarding this work.

References

[1] H. Sharif, R.A. Khan, A Novel Machine Learning Based Framework for Detection of Autism Spectrum Disorder (ASD), Applied Artificial Intelligence, 2021, pp. 1–33.
[2] F. Almuqhim, F. Saeed, ASD-SAENet, A sparse autoencoder, and deep-neural network model for detecting autism spectrum disorder (ASD) using fMRI data, Front. Comput. Neurosci. 15 (2021).
[3] A. Sadiq, M.I. Al-Hiyali, N. Yahya, T.B. Tang, D.M. Khan, Non-oscillatory connectivity approach for classification of autism spectrum disorder subtypes using resting-state fMRI, IEEE Access 10 (2022) 14049–14061.
[4] K. Devika, D. Mahapatra, R. Subramanian, V.R.M. Oruganti, Outlier-based autism detection using longitudinal structural MRI, IEEE Access 10 (2022) 27794–27808.
[5] Z. Sherkatghanad, M. Akhondzadeh, S. Salari, M. Zomorodi-Moghadam, M. Abdar, U.R. Acharya, R. Khosrowabadi, V. Salari, Automated detection of autism spectrum disorder using a convolutional neural network, Front. Neurosci. 13 (2020).
[6] M. Cao, M. Yang, C. Qin, X. Zhu, Y. Chen, J. Wang, T. Liu, Using DeepGCN to identify the autism spectrum disorder from multi-site resting-state data, Biomed. Signal Process Control 70 (2021), 103015.
[7] Y. Wang, J. Wang, F.-X. Wu, R. Hayrat, J. Liu, AIMAFE: autism spectrum disorder identification with multi-atlas deep feature representation and ensemble learning, J. Neurosci. Methods 343 (2020), 108840.
[8] S. Wang, R.M. Summers, Machine learning and radiology, Med. Image Anal. 16 (2012) 933–951.
[9] P. Moridian, N. Ghassemi, M. Jafari, S. Salloum-Asfar, D. Sadeghi, M. Khodatars, A. Shoeibi, A. Khosravi, S.H. Ling, A. Subasi, R. Alizadehsani, J.M. Gorriz, S. A. Abdulla, U.R. Acharya, Automatic autism spectrum disorder detection using artificial intelligence methods with MRI neuroimaging: a review, Front. Mol. Neurosci. 15 (2022).
[10] N. Wang, D. Yao, L. Ma, M. Liu, Multi-site clustering and nested feature extraction for identifying autism spectrum disorder with resting-state fMRI, Med. Image Anal. 75 (2022), 102279.
[11] L. Squarcina, G. Nosari, R. Marin, U. Castellani, M. Bellani, C. Bonivento, F. Fabbro, M. Molteni, P. Brambilla, Automatic classification of autism spectrum disorder in children using cortical thickness and support vector machine, Brain and Behavior 11 (2021) e2238.
[12] F. Zhao, X. Zhang, K.H. Thung, N. Mao, S.W. Lee, D. Shen, Constructing multi-view high-order functional connectivity networks for diagnosis of autism

spectrum disorder, IEEE (Inst. Electr. Electron. Eng.) Trans. Biomed. Eng. 69 (2022) 1237–1250.
[13] M.M. Rahman, O.L. Usman, R.C. Muniyandi, S. Sahran, S. Mohamed, R.A. Razak, A review of machine learning methods of feature selection and classification for autism spectrum disorder, Brain Sci. 10 (2020).
[14] J. Hu, L. Cao, T. Li, B. Liao, S. Dong, P. Li, Interpretable learning approaches in resting-state functional connectivity analysis: the case of autism spectrum disorder, Comput. Math. Methods Med. 2020 (2020), 1394830.
[15] R. Kashef, ECNN: enhanced convolutional neural network for efficient diagnosis of autism spectrum disorder, Cognit. Syst. Res. 71 (2022) 41–49.
[16] L. Li, H. Jiang, G. Wen, P. Cao, M. Xu, X. Liu, J. Yang, O. Zaiane, TE-HI-GCN, An ensemble of transfer hierarchical graph convolutional networks for disorder diagnosis, Neuroinformatics 20 (2022) 353–375.
[17] F.Z. Subah, K. Deb, P.K. Dhar, T. Koshiba, A Deep Learning Approach to Predict Autism Spectrum Disorder Using Multisite Resting-State fMRI, 11, Applied Sciences, 2021.
[18] M. Rakić, M. Cabezas, K. Kushibar, A. Oliver, X. Lladó, Improving the detection of autism spectrum disorder by combining structural and functional MRI information, Neuroimage: Clinical 25 (2020), 102181.
[19] M.R. Ahmed, Y. Zhang, Y. Liu, H. Liao, Single volume image generator and deep learning-based ASD classification, IEEE Journal of Biomedical and Health Informatics 24 (2020) 3044–3054.
[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention Is All You Need, arXiv e-prints, 2017. arXiv 1706.03762.
[21] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, 2014 arXiv:1409.1556.
[22] I. Vasilev, D. Slater, G. Spacagna, P. Roelants, V. Zocca, Python Deep Learning, second ed., 2019.
[23] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, 2015 arXiv:1512.03385.
[24] K. He, X. Zhang, S. Ren, J. Sun, Identity Mappings in Deep Residual Networks, 2016, 05027 arXiv:1603.
[25] F. Chollet, Xception: Deep Learning with Depthwise Separable Convolutions, 2016 arXiv:1610.02357.
[26] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the Inception Architecture for Computer Vision, 2015 arXiv:1512.00567.
[27] C. Szegedy, S. Ioffe, V. Vanhoucke, A. Alemi, Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning, 2016, 07261 arXiv:1602.
[28] G. Huang, Z. Liu, L. van der Maaten, K.Q. Weinberger, Densely Connected Convolutional Networks, 2016, 06993 arXiv:1608.
[29] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, MobileNets, Efficient Convolutional Neural Networks for Mobile Vision Applications, 2017, 04861 arXiv:1704.
[30] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, MobileNetV2: Inverted Residuals and Linear Bottlenecks, 2018, 04381 arXiv:1801.
[31] B. Zoph, V. Vasudevan, J. Shlens, Q.V. Le, Learning Transferable Architectures for Scalable Image Recognition, 2017, 07012 arXiv:1707.
[32] M. Tan, Q.V. Le, EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, 2019 arXiv:1905.11946.
[33] M. Tan, Q.V. Le, EfficientNetV2: Smaller Models and Faster Training, 2021, 00298 arXiv:2104.
[34] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale, arXiv e-prints, 2020. arXiv 2010.11929.
[35] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin Transformer, Hierarchical Vision Transformer Using Shifted Windows, arXiv e-prints, 2021. arXiv 2103.14030.
[36] S. Mehta, M. Rastegari, MobileViT: Light-Weight, General-Purpose, and Mobile-friendly Vision Transformer, 2021, 02178 arXiv:2110.
[37] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, N. Houlsby, Big Transfer (BiT): General Visual Representation Learning, 2020 arXiv pre-print server.
[38] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A ConvNet for the 2020s, 2022, 03545 arXiv:2201.
[39] U.S. Shanthamallu, A. Spanias, Machine and Deep Learning Algorithms and Applications, Springer International Publishing, 2022.
[40] M.J. Leming, S. Baron-Cohen, J. Suckling, Single-participant structural similarity matrices lead to greater accuracy in classification of participants than function in autism in MRI, Mol. Autism. 12 (2021) 34.
[41] H. Shahamat, M.S. Abadeh, Brain MRI analysis using a deep learning based evolutionary approach, Neural Network. 126 (2020) 218–234.
[42] J. Gao, M. Chen, Y. Li, Y. Gao, Y. Li, S. Cai, J. Wang, Multisite autism spectrum disorder classification using convolutional neural network classifier and individual morphological brain networks, Front. Neurosci. 14 (2021).
[43] M. Yang, M. Cao, Y. Chen, Y. Chen, G. Fan, C. Li, J. Wang, T. Liu, Large-scale brain functional network integration for discrimination of autism using a 3-D deep learning model, Front. Hum. Neurosci. 15 (2021).
[44] M. Tang, P. Kumar, H. Chen, A. Shrivastava, Deep multimodal learning for the diagnosis of autism spectrum disorder, Journal of Imaging 6 (2020).
[45] A. Chaddad, J. Li, Q. Lu, Y. Li, I.P. Okuwobi, C. Tanougast, C. Desrosiers, T. Niazi, Can autism Be diagnosed with artificial intelligence? A narrative review, Diagnostics 11 (2021) 2032.
[46] M. Khodatars, A. Shoeibi, D. Sadeghi, N. Ghaasemi, M. Jafari, P. Moridian, A. Khadem, R. Alizadehsani, A. Zare, Y. Kong, A. Khosravi, S. Nahavandi, S. Hussain, U.R. Acharya, M. Berk, Deep learning for neuroimaging-based diagnosis and rehabilitation of Autism Spectrum Disorder: a review, Comput. Biol. Med. 139 (2021), 104949.
[47] L. Meijie, L. Baojuan, H. Dewen, Autism spectrum disorder studies using fMRI data and machine learning: a review, Front. Neurosci. 15 (2021).
[48] H.S. Nogay, H. Adeli, Machine learning (ML) for the diagnosis of autism spectrum disorder (ASD) using brain imaging, Rev. Neurosci. 31 (2020) 825–841.
[49] R.A. Bahathiq, H. Banjar, A.K. Bamaga, S.K. Jarraya, Machine learning for autism spectrum disorder diagnosis using structural magnetic resonance imaging: promising but challenging, Front. Neuroinf. 16 (2022).
[50] M. Xu, V. Calhoun, R. Jiang, W. Yan, J. Sui, Brain imaging-based machine learning in autism spectrum disorder: methods and applications, J. Neurosci. Methods 361 (2021), 109271.
[51] F. Ahmad, I. Ahmad, Y. Guerrero-Sánchez, Classification of schizophrenia-associated brain regions in resting-state fMRI, The European Physical Journal Plus 138 (2023) 58.
[52] E.N. Pitsik, V.A. Maximenko, S.A. Kurkin, A.P. Sergeev, D. Stoyanov, R. Paunova, S. Kandilarova, D. Simeonova, A.E. Hramov, The topology of fMRI-based networks defines the performance of a graph neural network for the classification of patients with major depressive disorder, Chaos, Solitons & Fractals 167 (2023), 113041.
[53] N.A. Baghdadi, A. Malki, H.M. Balaha, M. Badawy, M. Elhosseini, A3C-TL-GTO: alzheimer automatic accurate classification using transfer learning and artificial Gorilla troops optimizer, Sensors 22 (2022).
[54] M. Kunda, S. Zhou, G. Gong, H. Lu, Improving multi-site autism classification via site-dependence minimization and second-order functional connectivity, IEEE Trans. Med. Imag. (2022) 1.
[55] T. Eslami, F. Saeed, Auto-ASD-network: a technique based on deep learning and support vector machines for diagnosing autism spectrum disorder using fmri data, Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (2019) 646–651.
[56] N. Lazar, The Statistical Analysis of Functional MRI Data, 1 ed., Springer, New York, NY, 2008.
[57] X. Deng, J. Zhang, R. Liu, K. Liu, Classifying ASD based on time-series fMRI using spatial–temporal transformer, Comput. Biol. Med. 151 (2022), 106320.
[58] F. Huang, E.-L. Tan, P. Yang, S. Huang, L. Ou-Yang, J. Cao, T. Wang, B. Lei, Self-weighted adaptive structure learning for ASD diagnosis via multi-template multi-center representation, Med. Image Anal. 63 (2020), 101662.
[59] T.M. Epalle, Y. Song, Z. Liu, H. Lu, Multi-atlas classification of autism spectrum disorder with hinge loss trained deep architectures: ABIDE I results, Appl. Soft Comput. 107 (2021), 107375.
[60] J. Liu, Y. Sheng, W. Lan, R. Guo, Y. Wang, J. Wang, Improved ASD classification using dynamic functional connectivity and multi-task feature selection, Pattern Recogn. Lett. 138 (2020) 82–87.
[61] S. Itani, D. Thanou, Combining anatomical and functional networks for neuropathology identification: a case study on autism spectrum disorder, Med. Image Anal. 69 (2021), 101986.
[62] A. Brahim, N. Farrugia, Graph Fourier transform of fMRI temporal signals based on an averaged structural connectome for the classification of neuroimaging, Artif. Intell. Med. 106 (2020), 101870.
[63] C. Yang, P. Wang, J. Tan, Q. Liu, X. Li, Autism spectrum disorder diagnosis using graph attention network based on spatial-constrained sparse functional brain networks, Comput. Biol. Med. 139 (2021), 104963.
[64] H. Jiang, P. Cao, M. Xu, J. Yang, O. Zaiane, Hi-GCN, A hierarchical graph convolution network for graph embedding learning of brain network and brain disorders prediction, Comput. Biol. Med. 127 (2020), 104096.
[65] A. Kazeminejad, R.C. Sotero, The importance of anti-correlations in graph theory based classification of autism spectrum disorder, Front. Neurosci. 14 (2020).
[66] M.T. Ali, Y. ElNakieb, A. Elnakib, A. Shalaby, A. Mahmoud, M. Ghazal, J. Yousaf, H.A. Khalifeh, M. Casanova, G. Barnes, A. El-Baz, The role of structure MRI in diagnosing autism, Diagnostics 12 (2022).
[67] Z. Rakhimberdina, X. Liu, T. Murata, Population graph-based multi-model ensemble method for diagnosing autism spectrum disorder, Sensors (2020) 20.
[68] C. Alvarez-Jimenez, N. Múnera-Garzón, M.A. Zuluaga, N.F. Velasco, E. Romero, Autism spectrum disorder characterization in children by capturing local-regional brain changes in MRI, Med. Phys. 47 (2020) 119–131.
[69] X. Ma, X.-H. Wang, L. Li, Identifying individuals with autism spectrum disorder based on the principal components of whole-brain phase synchrony, Neurosci. Lett. 742 (2021), 135519.
[70] M. Ingalhalikar, S. Shinde, A. Karmarkar, A. Rajan, D. Rangaprakash, G. Deshpande, Functional connectivity-based prediction of autism on site harmonized ABIDE dataset, IEEE (Inst. Electr. Electron. Eng.) Trans. Biomed. Eng. 68 (2021) 3628–3637.
[71] Y. Liu, L. Xu, J. Yu, J. Li, X. Yu, Identification of autism spectrum disorder using multi-regional resting-state data through an attention learning approach, Biomed. Signal Process Control 69 (2021), 102833.
[72] J. Hu, L. Cao, T. Li, S. Dong, P. Li, GAT-LI: a graph attention network based learning and interpreting method for functional brain network classification, BMC Bioinf. 22 (2021).
[73] L. Shao, C. Fu, Y. You, D. Fu, Classification of ASD based on fMRI data with deep learning, Cognitive Neurodynamics 15 (2021) 961–974.
[74] O. Dekhil, M. Ali, R. Haweel, Y. Elnakib, M. Ghazal, H. Hajjdiab, L. Fraiwan, A. Shalaby, A. Soliman, A. Mahmoud, R. Keynton, M.F. Casanova, G. Barnes, A. El-Baz, A comprehensive framework for differentiating autism spectrum disorder from neurotypicals by fusing structural MRI and resting state functional MRI, Semin. Pediatr. Neurol. 34 (2020), 100805.

[75] Y. Liu, L. Xu, J. Li, J. Yu, X. Yu, Attentional connectivity-based prediction of autism using heterogeneous rs-fMRI data from CC200 atlas, Experimental Neurobiology 29 (2020) 27–37.
[76] N. Chaitra, P.A. Vijaya, G. Deshpande, Diagnostic prediction of autism spectrum disorder using complex network measures in a machine learning framework, Biomed. Signal Process Control 62 (2020), 102099.
[77] L. He, H. Li, J. Wang, M. Chen, E. Gozdas, J.R. Dillman, N.A. Parikh, A multi-task, multi-stage deep transfer learning model for early prediction of neurodevelopment in very preterm infants, Sci. Rep. 10 (2020), 15072.
[78] H. Sewani, R. Kashef, An autoencoder-based deep learning classifier for efficient diagnosis of autism, Children 7 (2020).
[79] J. Zhang, F. Feng, T. Han, X. Gong, F. Duan, Detection of Autism Spectrum Disorder Using fMRI Functional Connectivity with Feature Selection and Deep Learning, Cognitive Computation, 2022.
[80] X. Yang, P.T. Schrader, N. Zhang, A deep neural network study of the ABIDE repository on autism spectrum classification, IJACSA) International Journal of Advanced Computer Science and Applications (2020) 11.
[81] L. Zhao, Y.-K. Sun, S.-W. Xue, H. Luo, X.-D. Lu, L.-H. Zhang, Identifying boys with autism spectrum disorder based on whole-brain resting-state interregional functional connections using a boruta-based support vector machine approach, Front. Neuroinf. 16 (2022).
[82] M.S. Ahammed, S. Niu, M.R. Ahmed, J. Dong, X. Gao, Y. Chen, DarkASDNet: classification of ASD on functional MRI using deep neural network, Front. Neuroinf. 15 (2021).
[83] Z. Wang, D. Peng, S. Yongbin, G. Jingjing, Autistic spectrum disorder detection and structural biomarker identification using self-attention model and individual-level morphological covariance brain networks, Front. Neurosci. 15 (2021).
[84] R.M. Thomas, S. Gallo, L. Cerliani, P. Zhutovsky, A. El-Gazzar, G.v. Wingen, Classifying autism spectrum disorder using the temporal statistics of resting-state functional MRI data with 3D convolutional neural networks, Front. Psychiatr. 11 (2020).
[85] P. Lu, X. Li, L. Hu, L. Lu, Integrating Genomic and Resting State fMRI for Efficient Autism Spectrum Disorder Classification, Multimedia Tools and Applications, 2021.
[86] O. Dekhil, M. Ali, Y. El-Nakieb, A. Shalaby, A. Soliman, A. Switala, A. Mahmoud, M. Ghazal, H. Hajjdiab, M.F. Casanova, A. Elmaghraby, R. Keynton, A. El-Baz, G. Barnes, A personalized autism diagnosis CAD system using a fusion of structural MRI and resting-state functional MRI data, Front. Psychiatr. 10 (2021).
[87] N. Bhagwat, A. Barry, E.W. Dickie, S.T. Brown, G.A. Devenyi, K. Hatano, E. DuPre, A. Dagher, M. Chakravarty, C.M.T. Greenwood, B. Misic, D.N. Kennedy, J.-B. Poline, Understanding the impact of preprocessing pipelines on neuroimaging cortical surface analyses, GigaScience 10 (2021).
[88] M.A. Reiter, A. Jahedi, A.R.J. Fredo, I. Fishman, B. Bailey, R.-A. Müller, Performance of machine learning classification models of autism using resting-state fMRI is contingent on sample heterogeneity, Neural Comput. Appl. 33 (2021) 3299–3310.
[89] K. Niu, J. Guo, Y. Pan, X. Gao, X. Peng, N. Li, H. Li, Multichannel deep attention neural networks for the classification of autism spectrum disorder using neuroimaging and personal characteristic data, Complexity 2020 (2020), 1357853.
[90] J. Wang, L. Zhang, Q. Wang, L. Chen, J. Shi, X. Chen, Z. Li, D. Shen, Multi-class ASD classification based on functional connectivity and functional correlation tensor via multi-source domain adaptation and multi-view sparse representation, IEEE Trans. Med. Imag. 39 (2020) 3137–3147.
[91] J. Li, F. Wang, J. Pan, Z. Wen, Identification of autism spectrum disorder with functional graph discriminative network, Front. Neurosci. 15 (2021).
[92] A. Bannadabhavi, S. Lee, W. Deng, X. Li, Community-Aware Transformer for Autism Prediction in fMRI Connectome, 2023 arXiv:2307.10181.
[93] X. Kan, W. Dai, H. Cui, Z. Zhang, Y. Guo, C. Yang, Brain Network Transformer, 2022, 06681 arXiv:2210.
[94] Z. Wang, Y. Xu, D. Peng, J. Gao, F. Lu, Brain functional activity-based classification of autism spectrum disorder using an attention-based graph neural network combined with gene expression, Cerebr. Cortex 33 (2022) 6407–6419.
[95] L. Herath, D. Meedeniya, M.A.J.C. Marasingha, V. Weerasinghe, Autism Spectrum Disorder Diagnosis Support Model Using Inception V3, 2021 International Research Conference on Smart Computing and Systems Engineering, SCSE), 2021, pp. 1–7.
[96] R. Nur Syahindah Husna, A.R. Syafeeza, N. Abdul Hamid, Y.C. Wong, R. Atikah Raihan, Functional magnetic resonance imaging for autism spectrum disorder detection using deep learning, Jurnal Teknologi 83 (2021) 45–52.
[97] N. Dominic, Daniel, T.W. Cenggoro, A. Budiarto, B. Pardamean, Transfer learning using inception-ResNet-v2 model to the augmented neuroimages data for autism spectrum disorder classification, Communications in Mathematical Biology and Neuroscience (2021).
[98] T. Wadhera, M. Mahmud, D.J. Brown, A Deep Concatenated Convolutional Neural Network-Based Method to Classify Autism, Springer Nature Singapore, Singapore, 2023, pp. 446–458.
[99] X. Chen, Z. Wang, Y. Zhan, F.A. Cheikh, M. Ullah, Interpretable Learning Approaches in Structural MRI: 3D-ResNet Fused Attention for Autism Spectrum Disorder Classification, SPIE, 2022.
[100] K. Gao, Z. Fan, J. Su, L.-L. Zeng, H. Shen, J. Zhu, D. Hu, Deep transfer learning for cerebral cortex using area-preserving geometry mapping, Cerebr. Cortex 32 (2021) 2972–2984.
[101] V. Prasad, G.V. Sriramakrishnan, I. Diana Jeba Jingle, Autism spectrum disorder detection using brain MRI image enabled deep learning with hybrid sewing training optimization, Signal, Image and Video Processing 17 (2023) 4001–4008.
[102] H. Zhang, Z. Wang, Y. Zhan, Classification and diagnosis of autism spectrum disorder using Swin transformer, in: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), 2023, pp. 1–4.
[103] A. Othmani, T. Bizet, T. Pellerin, B. Hamdi, M.-A. Bock, S. Dev, Significant CC400 Functional Brain Parcellations Based LeNet5 Convolutional Neural Network for Autism Spectrum Disorder Detection, Springer Nature Switzerland, Cham, 2023, pp. 34–45.
[104] M.I. Al-Hiyali, N. Yahya, I. Faye, Z. Khan, Autism spectrum disorder detection based on wavelet transform of BOLD fMRI signals using pre-trained convolution neural network, International Journal of Integrated Engineering 13 (2021) 49–56.
[105] R. Liu, Z.A. Huang, Y. Hu, Z. Zhu, K.C. Wong, K.C. Tan, Spatial–temporal Co-attention learning for diagnosis of mental disorders from resting-state fMRI data, IEEE Transact. Neural Networks Learn. Syst. (2023) 1–15.
[106] B. Hasan Atakan, Ş. Irmak, D. Onat, D. Salman Ul Hassan, Ç. Tolga, BolT: Fused Window Transformers for fMRI Time Series Analysis, arXiv e-prints, 2022 arXiv 2205.11578.
[107] W. Yin, S. Mostafa, F.-x. Wu, Diagnosis of autism spectrum disorder based on functional brain networks with deep learning, J. Comput. Biol. 28 (2021) 146–165.
[108] M. Leming, J.M. Górriz, J. Suckling, Ensemble deep learning on large, mixed-site fMRI datasets in autism and other tasks, Int. J. Neural Syst. 30 (2020), 2050012.
[109] T.N. Kipf, M. Welling, Semi-Supervised Classification with Graph Convolutional Networks, arXiv e-prints, 2016. arXiv 1609.02907.
[110] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, Y. Bengio, Graph Attention Networks, arXiv e-prints, 2017. arXiv 1710.10903.
[111] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (2010) 1345–1359.
[112] C. Szegedy, L. Wei, J. Yangqing, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1–9.
[113] W. Li, S. Wang, G. Liu, Transformer-based Model for fMRI Data: ABIDE Results, 2022 7th International Conference on Computer and Communication Systems, ICCCS), 2022, pp. 162–167.
[114] L. Mahler, Q. Wang, J. Steiglechner, F. Birk, S. Heczko, K. Scheffler, G. Lohmann, Pretraining Is All You Need: A Multi-Atlas Enhanced Transformer Framework for Autism Spectrum Disorder Classification, 2023, 01759 arXiv:2307.
[115] W. Dai, Z. Zhang, L. Tian, S. Yu, S. Wang, Z. Dong, H. Zheng, First Glance Diagnosis: Brain Disease Classification with Single fMRI Volume, arXiv e-prints, 2022 arXiv 2208.03028.
[116] Y. Qiu, S. Yu, Y. Zhou, D. Liu, X. Song, T. Wang, B. Lei, Multi-channel sparse graph transformer network for early Alzheimer's disease identification, in: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), 2021, pp. 1794–1797.
[117] B.-H. Kim, J.C. Ye, J.-J. Kim, Learning Dynamic Graph Representation of Brain Connectome with Spatio-Temporal Attention, arXiv e-prints, 2021 arXiv 2105.13495.
[118] E. Jun, S. Jeong, D.-W. Heo, H.-I. Suk, Medical Transformer, Universal Brain Encoder for 3D MRI Analysis, 2021, 13633. ArXiv, abs/2104.
[119] R.S. Desikan, F. Ségonne, B. Fischl, B.T. Quinn, B.C. Dickerson, D. Blacker, R. L. Buckner, A.M. Dale, R.P. Maguire, B.T. Hyman, M.S. Albert, R.J. Killiany, An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest, Neuroimage 31 (2006) 968–980.
[120] I.E. Tampu, A. Eklund, N. Haj-Hosseini, Inflation of test accuracy due to data leakage in deep learning-based classification of OCT images, Sci. Data 9 (2022) 580.
[121] A. Kedia, M. Rasu, Hands on Python Natural Language Process.
[122] K. He, C. Gan, Z. Li, I. Rekik, Z. Yin, W. Ji, Y. Gao, Q. Wang, J. Zhang, D. Shen, Transformers in medical image analysis, Intelligent Medicine (2022).
[123] M. Radhakrishnan, K. Ramamurthy, K.K. Choudhury, D. Won, T.A. Manoharan, Performance analysis of deep learning models for detection of autism spectrum disorder from EEG signals, Trait. Du. Signal 38 (2021) 853–863.
[124] R. Menaka, R. Karthik, S. Saranya, M. Niranjan, S. Kabilan, An improved AlexNet model and cepstral coefficient-based classification of autism using EEG, Clin. EEG Neurosci. (2023), 15500594231178274.