Biological Psychology 194 (2025) 108976

Contents lists available at ScienceDirect

Biological Psychology

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

MCBERT: A multi-modal framework for the diagnosis of autism spectrum disorder

Kainat Khan, Rahul Katarya *

Big Data Analytics and Web Intelligence Laboratory, Department of Computer Science & Engineering, Delhi Technological University, New Delhi, India

ARTICLE INFO

Keywords:
Autism spectrum disorder
BERT
Convolutional neural network
Deep learning
Medical imaging
Multi-modal architecture

ABSTRACT

Within the domain of neurodevelopmental disorders, autism spectrum disorder (ASD) emerges as a distinctive neurological condition characterized by multifaceted challenges. The delayed identification of ASD poses a considerable hurdle in effectively managing its impact and mitigating its severity. Addressing these complexities requires a nuanced understanding of data modalities and the underlying patterns. Existing studies have focused on a single data modality for ASD diagnosis. Recently, there has been a significant shift towards multimodal architectures with deep learning strategies due to their ability to handle and incorporate complex data modalities. In this paper, we developed a novel multimodal ASD diagnosis architecture, referred to as Multi-Head CNN with BERT (MCBERT), which integrates bidirectional encoder representations from transformers (BERT) for meta-features and a multi-head convolutional neural network (MCNN) for the brain image modality. The MCNN incorporates two attention mechanisms to capture spatial (SAC) and channel (CAC) features. The outputs of BERT and MCNN are then fused and processed through a classification module to generate the final diagnosis. We employed the ABIDE-I dataset, a multimodal dataset, and conducted a leave-one-site-out classification to assess the model's effectiveness comprehensively. Experimental simulations demonstrate that the proposed architecture achieves a high accuracy of 93.4 %. Furthermore, the exploration of functional MRI data may provide a deeper understanding of the underlying characteristics of ASD.

1. Introduction

Autism spectrum disorder (ASD) is a persistent neurological condition marked by repetitive and restricted behaviors, alongside challenges in communication and social skills, encompassing a diverse range of sensory-motor deficits (Abbas et al., 2023). A recent survey indicates that approximately 1.5 % of the global population falls within the autism spectrum, with many cases going unnoticed (Miaoyan Wang et al., 2023)(Luo et al., 2023). Signs such as avoidance of eye contact, unresponsiveness to their names, repetitive movements, and limited speech compared to peers serve as indicators for identifying autism in children. Various innovative techniques have emerged to enhance cognitive abilities and alleviate ASD symptoms (Çetintaş et al., 2023). However, behavioral therapy and early diagnosis at the initial stage are some of the best solutions for elevating the patient's interpersonal interaction and reducing aggression and anxiety. Early intervention plays a pivotal role in improving the lives of those with ASD, beginning with the identification of affected children (Park & Cho, 2023)(Sun et al., 2023). Despite advancements in research and treatment, there is currently no accurate clinical treatment available to cure ASD (Alsaade & Alzahrani, 2022). Understanding cognition as a combination of processes in the brain underscores the intricate nature of ASD (Timms et al., 2022) (Gracia, 2022)(Qayyum et al., 2022). Early detection, paired with comprehensive intervention strategies, remains pivotal in addressing the multifaceted challenges associated with autism and promoting optimal cognitive development (Landowska et al., 2022) (Parlett-Pelleriti et al., 2022)(Egger et al., 2022).

As discussed in the above paragraph, detecting autism spectrum

Abbreviations: AI, Artificial Intelligence; ASD, Autism Spectrum Disorder; ABIDE, Autism Brain Imaging Data Exchange; BERT, Bidirectional Encoder Representation Transformers; CNN, Convolutional Neural Network; DL, Deep Learning; DT, Decision Tree; ECG, Electrocardiography; EOG, electrooculography; FP, False Positive; FN, False Negative; KNN, K-Nearest Neighbor; LSTM, Long Short-Term Memory; ML, Machine Learning; MLP, Multi-Layer Perceptron; SMRI/fMRI, Structural/functional magnetic resource imaging; SVM, Support Vector Machine; MM, Multi-Modal; NB, Naïve Bayes; NN, Neural Network; PDA, Principal Component Analysis; TD, Typically Developed; TP, True Positive; TN, True Negative; ReLU, Rectified Linear Unit; RF, Random Forest.

* Corresponding author.
E-mail addresses: Khankainat388@gmail.com (K. Khan), rahuldtu@gmail.com (R. Katarya).

https://doi.org/10.1016/j.biopsycho.2024.108976
Received 12 June 2024; Received in revised form 28 November 2024; Accepted 16 December 2024
Available online 23 December 2024

disorder at an early stage is crucial for preventing the deterioration of the individual's condition (Nogay & Adeli, 2024)(Ciceri et al., 2023) (Joudar et al., 2023). Therefore, there is a need for convenient, accurate, and time-efficient methods to diagnose autistic traits. Presently, predictive models utilizing extensive datasets play a pivotal role in diagnosing and predicting ASD, contributing to the enhancement of the quality of life for individuals affected by ASD (Khor et al., 2023) (Loganathan et al., 2023). However, the intricate neural patterns underpinning the spectrum of autism behaviors and the severity of the disorder remain inadequately understood (Khodatars et al., 2021)(Akter et al., 2021). Several traditional strategies are employed in the diagnosis and treatment of ASD. These include the utilization of modalities such as functional magnetic resource imaging (fMRI), blood tests, structural magnetic resource imaging (sMRI), interviews (questionnaires), and electroencephalograms (EEG) (Shoeibi et al., 2021) (Emon et al., 2021). In recent developments, non-invasive brain imaging has provided a more comprehensive understanding of the neural circuitry linked with neural developmental disorders (Lin et al., 2020). Notably, fMRI enables the visual evaluation of the functional characteristics of the brain. This offers precise insights into various neurological disorders (Khan & Katarya, 2023). For example, in the diagnosis of ASD, rather than solely depending on observational methods and patient interactions, physicians leverage neuroimages to detect anomalies in brain activity. This approach enhances the efficiency and precision of identifying differences in neural pathways among patients. Developing an architecture that takes both images and its meta-data (multi-modal) into account, can enhance the efficacy and association between both modalities. This multi-modal strategy can prove to be effective in diagnosing ASD.

1.1. Major contribution

There has been a substantial accumulation of non-imaging (meta-features) datasets. Elements like gender, behavioral characteristics, patient history, and genetic sequences significantly influence disease diagnosis. The integration of non-imaging and imaging data through multimodal architecture is crucial for enhancing the efficacy of algorithms. Nonetheless, non-imaging/meta-features exhibit high dimensionality, constraining the representational capabilities of conventional machine-learning approaches (Waizbard-Bartov et al., 2023). Deep learning strategies present an avenue for efficiently amalgamating multimodal data to facilitate the diagnosis of autism spectrum disorder (ASD). The research framework built in this paper primarily concentrates on adopting deep-learning image and text-processing techniques to establish a multi-modal framework for diagnosing ASD. This is achieved by employing various techniques and fusing their outputs at the end. In the developed architecture, we introduced blocks/ components for image and meta-features modalities to diagnose ASD (detailed explained in Section 3).

The key contributions of the work are:

- Introducing a novel fusion technique for the diagnosis of ASD in a multimodal setting. This technique simultaneously integrates information from brain magnetic resource imaging (MRI) images and its meta-features by fusing the output of multi-head CNN and bidirectional encoder representation from transformers (BERT).
- Developing an efficient approach for feature fusion, utilizing convolutional block attention component (CBAC) to extract spatial and channel dimension attributes.
- Creating a BERT-based architecture that incorporates layers for efficiently handling the meta-features. This enables the retrieval of important features from the meta-features modality.
- Conducting a comprehensive evaluation of the proposed multimodal fusion architecture by comparing the experimental findings with existing works, ablation study, and by performing the LOSO test. The assessment is performed on a multimodal dataset called as autism brain imaging data exchange (ABIDE) to validate the performance of the MCBERT architecture.

1.2. Paper organization

The following sections comprise the structure of this research: A detailed literature review that incorporates current methods in the diagnosis of autism spectrum disorder is included in Section 2. The technique of the suggested architecture is explained in Section 3, mainly covering the topics of multi-head CNN, channel attention component (CAC), and spatial attention component (SAC). Section 4 presents a thorough examination of the experimental results using an ablation study, a leave-one-site-out classification test, challenges, and future directions. To demonstrate the effectiveness of the suggested model, this part is further subdivided. Finally, Section 5 concludes the entire work and outlines its potential and limitations.

2. Literature survey

In recent years, there has been a rapid boost in employing Artificial Intelligence techniques, specifically well-known techniques like machine learning (ML), computational intelligence, and the deep learning (DL), for building multimodal autism spectrum disorder (ASD) diagnosis systems. All the mentioned techniques have been used in recent years. Still, there is a significant shift towards multimodal architectures and deep learning due to its capability of handling and incorporating complex data modalities. However, these technique showcases several drawbacks that need to be fulfilled. So, to gain a comprehensive comprehension of the fundamental principles and methodologies employed in prior investigation, we extensively reviewed a multitude of papers related to the diagnosis of autism spectrum disorder. We also incorporated papers from other domains that have worked on multimodal architectures. In order to synthesize our current understanding and clarify knowledge gaps, this section explores the existing research on ASD and other areas. The section's discussion is essential to laying the foundation for our suggested study. Table 1 elucidates several noteworthy and pertinent works undertaken in recent years.

It is evident from the literature analysis (Table 1 and Table 2) above that many methods have been developed by researchers to diagnose autism spectrum disorder (ASD). Others combined ideas of special optimization algorithms and deep learning networks to diagnose ASD, while others concentrated on ASD detection to identify facial indicators using machine learning. Each research project contributed a different viewpoint and unique insights to this field of study. All the papers reviewed in Table 1 and Table 2 incorporate multi-modality datasets. Most of the papers are focused on the medical domain while few of them are cross-domain researches.

To get a deeper understanding of the previous architecture on the single modality for ASD, Table 3 and Table 4 outline some of the noteworthy research specifically conducted for ASD. Table 3 focuses on fMRI studies, while Table 4 includes those employing sMRI and facial image datasets. Deep learning strategies have been adopted widely to handle the intricate ASD dataset and to develop an effective architecture to analyze ASD.

3. Proposed Architecture

In this segment, we discussed the preliminaries of the standard techniques with the methodology adopted in the work and the novelty introduced. We explained the problem statement and the techniques mathematically as shown in further sub-sections.

3.1. Problem statement

Autism spectrum disorder (ASD) presents a multifaceted challenge, with its prevalence escalating and traditional diagnostic procedures

Table 1
A literature survey performed on the multimodal architectures developed for the diagnosis of ASD, various diseases (Medical), and other domains using machine learning/deep learning, where, ***M*$_1$**: Image; ***M*$_2$**: Text; ***M*$_3$**: Meta-features/Sensor data; ***M*$_4$**: Videos; ***M*$_5$**: Audio; ***M*$_6$**: Signals, denotes different data modalities.

Author, Year	Objective	Techniques	Included Modalities						Target Domain	Outcomes	Limitations/ Future work
			***M*$_1$**	***M*$_2$**	***M*$_3$**	***M*$_4$**	***M*$_5$**	***M*$_6$**			
(Mingzhi Wang et al., 2023)	Multimodal ASD diagnosis architecture	Weight learning network; Graph CNN; DeepGCN	✓ (ABIDE I)		✓				Medical	Acc: 77.27 %; Pre:77.7 %; Recall: 80.96 %	Small data size; lack of interpretability; imbalanced gender ratio
(J. Li et al., 2023)	Building multimodal dataset for autism analysis		✓			✓			Medical	Consists of 1315 videos for social and movement behavior analysis	
(Song et al., 2023)	Multimodal technique based on response towards name behavior of children suffering from autism	Human pose tracking; Automatic name detection; Head pose estimation	Dataset of 30 participants				✓	✓	Medical	Acc: ~93.3 %	Small dataset; limited generalizability; dependency on body movements; sensitivity to reaction speed
(Han et al., 2022)	Proposed multimodal architecture for diagnosing ASD in children	Stacked denoising encoder (SDAE)	(EEG, ET data of 90 individuals)					✓	Medical	Acc: ~93.56 %; Sen: ~92.50 %; Spec: ~98.0 %	High computational cost; advanced NN algorithms can be explored
(Herath et al., 2024)	Multimodal ASD identification	Multimodal + multisite ensemble classifier (Inception V3; MobileNet, DensetNet, ResNet50)	✓ (ABIDE-I and II)		✓				Medical	Best acc: 97.82 % (improvement of ↑3.25 %)	Other data modalities can be incorporated; more number of training images can be added
(Haputhanthri et al., 2020)	Classification of ASD using different modalities	EEG and thermographic feature extraction; Naïve Bayes; neural net; logistic regression; random forest	✓ (ADOS−2)					✓	Medical	Best acc: 94 %	Small dataset
(M. Tang et al., 2020)	Multimodal architecture for diagnosing ASD	3D-ResNet; MLP	✓		✓				Medical	Acc: 74 %; Recall: 95 %; F1: 0.805	Limited amount of data; Low accuracy; overfitting issues; incorporate optimization techniques with large dataset
(Chen et al., 2022)	Autism identification via adversarial-graph learning networks	Adversarial learning; Garph networks	✓ (sMRI and fMRI)						Medical	Accuracy: 74.7 %; Specificity: 77.4 %	Accuracy needs to be enhanced; No incorporation of meta-attributes

proving inadequate in effectively addressing the complexities of identification. The insufficiency of suitable tools, medical assessments, and therapies compounds the difficulty in diagnosing ASD accurately. This predicament has spurred the emergence of advanced ASD diagnostic frameworks integrating deep learning. Nevertheless, due to the intricate nature of the disorder, conventional ASD identification frameworks encounter hurdles in achieving a precise diagnosis. Researchers are exploring pivotal biomarkers such as eye-tracking metrics, functional/anatomical brain attributes, neurophysiological patterns, behavioral indicators, and genetic markers. Among these, brain attributes stand out as a widely adopted and significant biomarker, given the direct involvement of the nervous system. The ability to discern brain patterns holds promise for distinguishing brain asymmetry. Efficient ASD diagnosis necessitates a comprehensive history (meta-features) of the child's development coupled with the ASD biomarker. Consequently, researchers have provided diverse diagnostic frameworks leveraging ASD biomarkers and DL algorithms on a single modality data. However, despite significant research in this realm, diagnosing ASD via multimodal architecture is still very limited. This work endeavors to address the limitation by proposing a multi-modal ASD diagnostic architecture, specifically focusing on brain MRIs.

3.2. Convolution neural network (CNN)

In contemporary deep learning for image recognition/classification, CNNs stand out as a prominent neural network architecture. The architecture of CNN is structured into three layers: (i) the entry layer, (ii) the hidden (latent) layer, and (iii) the output layer. The hidden layers, alternatively termed pooling, or completely connected layers, play a major role in the overall architecture (Kwon et al., 2022) (Herath et al., 2022).

3.2.1. The convolutional layer

The convolutional strategy is applied recurrently within this layer to induce changes in the output function. Comprising the neuronal maps, also known as the "filter/feature maps" or "characteristic maps", the discrete convolution of receptors quantifies neural activity (Fig. 1, block A). This process involves computing the overall neural weights of the input and activation function assignments (Wan et al., 2022). Fig. 1 provides a visual representation of a generic discrete convolutional layer.

3.2.2. Max pool layer

The max pool layer forms a multitude of grids from the segmented convolutional layer's output. Sequential matrices are created using the maximum grid value. Operators are employed to derive the average or maximum value for each matrix. Fig. 1, block B, illustrates the construction of the max pool layer (Wawer et al., 2022).

Table 2
A literature survey performed on the multimodal architectures developed for the diagnosis of ASD, various diseases (Medical), and other domains using machine learning/deep learning, where, ***M*$_1$**: Image; ***M*$_2$**: Text; ***M*$_3$**: Meta-features/Sensor data; ***M*$_4$**: Videos; ***M*$_5$**: Audio; ***M*$_6$**: Signals, denotes different data modalities.

Author, Year	Objective	Techniques	Included Modalities						Target Domain	Outcomes	Limitations/ Future work
			***M*$_1$**	***M*$_2$**	***M*$_3$**	***M*$_4$**	***M*$_5$**	***M*$_6$**			
(Hasan et al., 2024)	Multimodal drowsiness detection architecture through explainable machine learning	KNN; SVM; RF; SHAP; PDA	(EEG, ECG, EOG)					✓	Medical	Acc: 80.1 %; Sen: 70.35; Spec: 82.2 %	Small dataset; limited number of features; Deep learning techniques can be explored
(Du et al., 2022)	Multimodal classification technique to analyze the uniqueness of ASD and Schizophrenia	Functional and structural connectivity measures	✓ (fMRI, sMRI)						Medical	Acc: 83.08 %	Lack of assessment of symptoms; limited neuro-imaging measures; model-level fusion only
(Chan et al., 2023)	Behavior change prediction in students via multimodal architecture	Feature engineering; sampling techniques; SVM; NB; DT; RF; KNN; MLP; XGboost			✓	✓			Behavior change	Acc: 98 %; Precision: 97 %	Prediction target is binary; costly setup
(Fu et al., 2023)	Survival prediction via multimodal graph-based framework	Region-based via multimodal module; Embedding module; deep MM graph-based network	✓ (METABRIC; BASEL dataset)		✓				Medical	Prediction performance: METABRIC = 0.7484; BASEL = 0.7479	Lack of interpretability; spatial simplification impact; less generalization to new data
(Passos et al., 2023)	Multimodal architecture for energy-efficient speech enhancement	Self-supervised framework integrating graph NN and canonical correlation analysis (CCA)	✓ (AV ChiMe3 dataset)			✓			Speech enhancement	Proposed framework ensures improved feature learning	Did not quantify the amount of energy saving; biologically realistic neuronal architecture can be developed
(Le et al., 2023)	Multilabel and multimodal emotion recognition	Feature extraction (CNN, ALBERT); multimodal fusion (transformers); emotion-level embedding (multi-head attention)		✓ (IEMOCAP; CMU-MOSEI)	✓	✓	✓		Emotion recognition	Developed framework outperforms existing methods with an accuracy of 85.9 %	High computational cost; time-consuming; redundant frames in videos
(Jaafar & Lachiri, 2023)	Detection of aggression in surveillance	Multiple deep neural networks; 3D-CNN		✓ (Dataset of aggression in trains)	✓	✓	✓		Medical	Unweighted average acc = 85.66 %; Weighted average acc = 86.35 %	Results cannot be generalized on a huge dataset; the model cannot specify all aggressive situations
(De Silva et al., 2021)	Decision support system for ADHD	Seed-based correlation; data augmentation; CNN	✓ (Eye movement data + fMRI)		✓				Medical	Acc: 82 %	Accuracy can be enhanced
(Meiwei Zhang et al., 2024)	Multimodal machine learning-based Alzheimer diagnosis framework	Extreme learning machine; entropy-based polynomial function; attention mechanism	✓ (ADNI)		✓				Medical	Acc~ 98 %	Lack of generalizability; does not address the issue of missing data
(Yu et al., 2024)	Multimodal transformer-based framework for Alzheimer	Transformers	✓ (ADNI)		✓				Medical	AUC: 0.993	Lack of generalizability; lack of result validation
(Sheng et al., 2024)	Multimodal hybrid framework for Alzheimer's diagnosis	Harris hawks optimization; kernel extreme learning	✓ (ADNI: MRI + CSF + PET)						Medical	Acc: 99.2 %	Parameter sensitivity; computationally intensive; overfitting
(Z. Li et al., 2023)	Supervised and self-supervised learning on multimodal data	Self-attention; latent feature extraction; cross-modality feature learning	✓ (CINEPS and COEPS datasets)		✓				Medical	The developed model performed significantly well on various parameters	Imbalance in the data ratio; more data modalities can be considered
(Mengyi Zhang et al., 2024)	Multimodal Alzheimer's disease diagnosis using brain images	Pyramid attention strategy with GAN	✓ (ADNI dataset: MRI and PET images)						Medical	Acc: 89.9 %	Small dataset; accuracy can be enhanced
(Moon et al., 2022)	Proposed MedVill for Multimodal representation learning	BERT; Multimodal attention strategy	✓ (MIMIC-CXR; Open-I; VQA-RAD)	✓					Medical	MedVill performed well against various considered techniques	Scope on accuracy improvement; Need to work on diverse multi-view studies
(Alpar, 2023)	Multimodal tumor segmentation using a mathematical fuzzy framework	Nakagami imaging; Fuzzy fusion; Segmentation	✓ (Two types of images: Binary segmented and FLAIR images)						Medical	Average dice score: 92.78 %	High number of training parameters

Table 3
Literature survey for autism spectrum disorder on single modality architectures incorporating fMRI dataset.

Authors, Year	Objective	Techniques	Data Modality	Achievements	Limitations
(Elakkiya & Dejey, 2024)	Deep learning integrated activation function for the screening of autism	Developed two models, namely MinAutiNet and AutiNet for processing fMRI	fMRI	Max Accuracy: AutiNet = 77.78 %; MinAutiNet = 88.89 %	Need for automatic feature extraction techniques: accuracy needs to be improved
(N. Li et al., 2024) 2024	Multi-level joint learning network for the brain to diagnose ASD	Graph networks	fMRI	Accuracy: 81.5 %	High model complexity; large number of parameters
(Y. Tang et al., 2023),	Multi-site ASD diagnosis	LSTM; Two-stage adversarial approach	fMRI	Accuracy: 0.80; Specificity: 0.80; Sensitivity: 0.81	Accuracy needs to be improved; single-modality architecture
(Parui et al., 2023),	ASD diagnosis via sensor-based and AI approach	Brain connectivity analysis	fMRI	Accuracy: 84.79 %	Need to improve accuracy
(Kang et al., 2022)	ASD recognition via multi-view ensemble and multi-site fMRI	LSTM-Conv architecture; SDAE; PCA	fMRI	Accuracy: 72.0 %	No incorporation of sMRI; lower accuracy

Table 4
Literature survey for autism spectrum disorder on single modality architectures incorporating sMRI and facial images.

Authors, Year	Objective	Techniques	Data Modality	Achievements	Limitations
(Nogay & Adeli, 2024)	Deep learning-based ASD classification via age and gender factors	CNN, data augmentation, grid search optimization with multiple classifications	sMRI	Max Accuracy: 85.42 %	Only two factors were taken into consideration
(Mishra & Pati, 2023),	ASD classification framework	Deep CNN; data augmentation; optimization	sMRI	Max Accuracy: 81.35 %	Accuracy can be enhanced; single modality model
(El Mouatasim & Ikermane, 2023)	ASD diagnosis via facial imaging	Control sub-gradient approach with deep CNN; DenseNet model	Face images	The developed approach with DenseNet enhanced the overall results with Precision = 98 %; Recall = 97 %; F1-score = 97 %	Single modality architecture; limited exploration of hyperparameters;
(Dc et al., 2022)	ASD severity detection via ML	KNN, DT, SVM, NB, RF; GLCM	Face images	Max Accuracy: 91.2 %	Deep learning strategies can be adopted

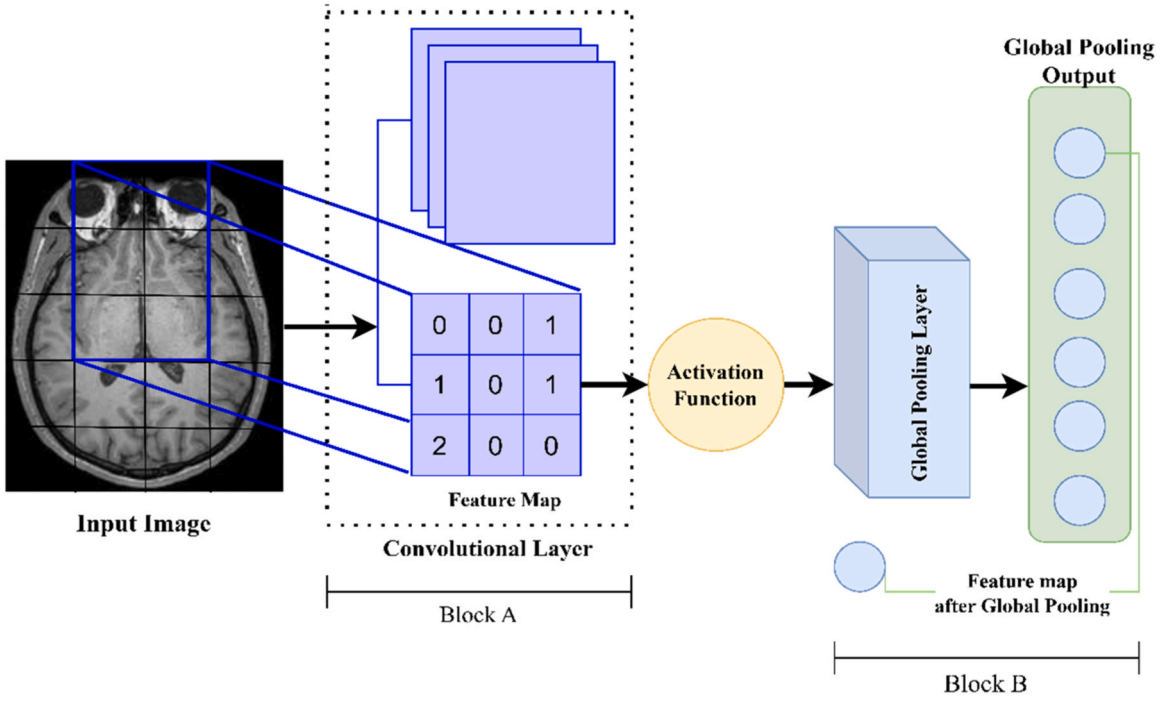


Fig. 1. Visual representation of the workflow of generic convolution neural network.

3.2.3. *Fully connected layer (FCL)*

Constituting 90 % of the entirety of the structural elements of the CNN, the FCL allows the transmission of the input across the network with a pre-configured vector length. Data is transformed within this portion before grading. The convolutional layer is also transformed to conserve information integrity. Neurons from every preceding layer are utilized in these FCLs, serving as the network's ultimate layer (Khodatars et al., 2021).

3.3. Multi-head CNN

In this research paper, we introduced a three-headed convolutional neural network specifically crafted to extract pertinent patterns from input images. The convolutional layer comprises multiple convolutional filters that, through convolution operations, generate the output feature map (mainly explained in the above section) from input images. Within the convolutional layers, the obtained feature maps via the preceding layer undergo convolution via various kernels (Lasantha et al., 2024). Additionally, bias is incorporated to augment the outcome of the convolution operation, which subsequently passes via an activation function, giving rise to the feature maps for the subsequent layers. Mathematically, the mth feature map at the lth layer of the eth head of the multi-head CNN is represented as a matrix, with the value at the kth row denoted as $R_{lm}^{k,e}$. The calculation of this value follows the formula presented in Eq. (1).

$$R_{lm}^{k,e} = f_{Relu}\left(f_{conv2d}^{e}\left(R_{l-1}^{k+j}\right)\right), \quad \forall e = 1, \quad 2, \quad 3 \tag{1}$$

Here, f_{Relu} denotes the activation function that replaces all negative values with 0 (zero) in the feature map, while f_{conv2d}^{e} represents the convolution function of the eth head in our multi-head CNNs, s articulated in the Eq. (2).

$$f_{conv2d}^{e}\left(R_{l-1}^{k+j}\right) = b_{lm} + \sum_{i}\sum_{j=0}^{\eta_l^e-1} W_{lmi}^{je} R_{(l-1)}^{k+j,\ e} \tag{2}$$

b_{lm} represents the bias for a particular feature map, where i is the index of the feature map at the $(l-1)$ layer. Additionally, W_{lmi}^{je} signifies the weight matrix present at the position j of the convolution kernels, and η_l^e represents the length of the kernel of the eth head in our multi-head CNN. In the developed multi-head CNN architecture, a crucial element is the pooling layer. This layer plays a pivotal role in reducing the parameter count and computations by decreasing the spatial size of the feature representation. Among the various pooling techniques, max pooling stands out as the most popular and widely utilized method.

$$P_{hlm} = \max_{(y,\ z)\ \in\ \mathbb{R}_{l,m}} \nu_{hyz} \tag{3}$$

Here P_{hlm} represents the pool operation of the hth feature maps. ν_{hyz} signifies the component at position (y, z) enclosed by the pool region $\mathbb{R}_{l,m}$. This region shows a receptive field around (l, m).

3.4. Convolution block attention component (CBAC)

In the developed architecture, each head incorporates two CBACs to optimize training performance by accentuating both spatial and channel features with brain MRI images. The CAC network empowers the MCBERT framework to concentrate on crucial channel features while disregarding others. These channel features contain intricacies intrinsic to the individual color or feature channels, delineating distinct aspects such as textual nuances and color variations. This channel-level scrutiny is necessary for capturing fine-grain details by facilitating a comprehensive characterization of image content. To assess the importance of every channel, diverse weight information is utilized to various feature channels and feature dimensions of the visual data. SAC enables the architecture (Fig. 2, block A) to prioritize spatial dimension information on the feature map. The features encapsulate the spatial relationships, structural configurations, and overall layout of the image, presenting a holistic perspective on the contextual arrangement of visual elements. The analysis of spatial features is pivotal for decoding the spatial semantics and intrinsic geometry embedded within the visual data. For feature extraction, CBAC sequentially extracts a 1-D channel attention map $A_C \in \mathbb{R}^{c\times1\times1}$ and a 2-D spatial attention map $A_s \in \mathbb{R}^{1\times H\times W}$ from the provided intermediate feature map $I \in \mathbb{R}^{c\times H\times W}$ of the MRI visual data. The comprehensive attention mechanism is articulated in Eqs. (4) and (5).

$$I' = A_C(I) \otimes I \tag{4}$$

$$I'' = A_s(I) \otimes I' \tag{5}$$

In this context, the symbol $\otimes$ denotes element-wise multiplication, producing the refined feature I''. The channel attention features undergo

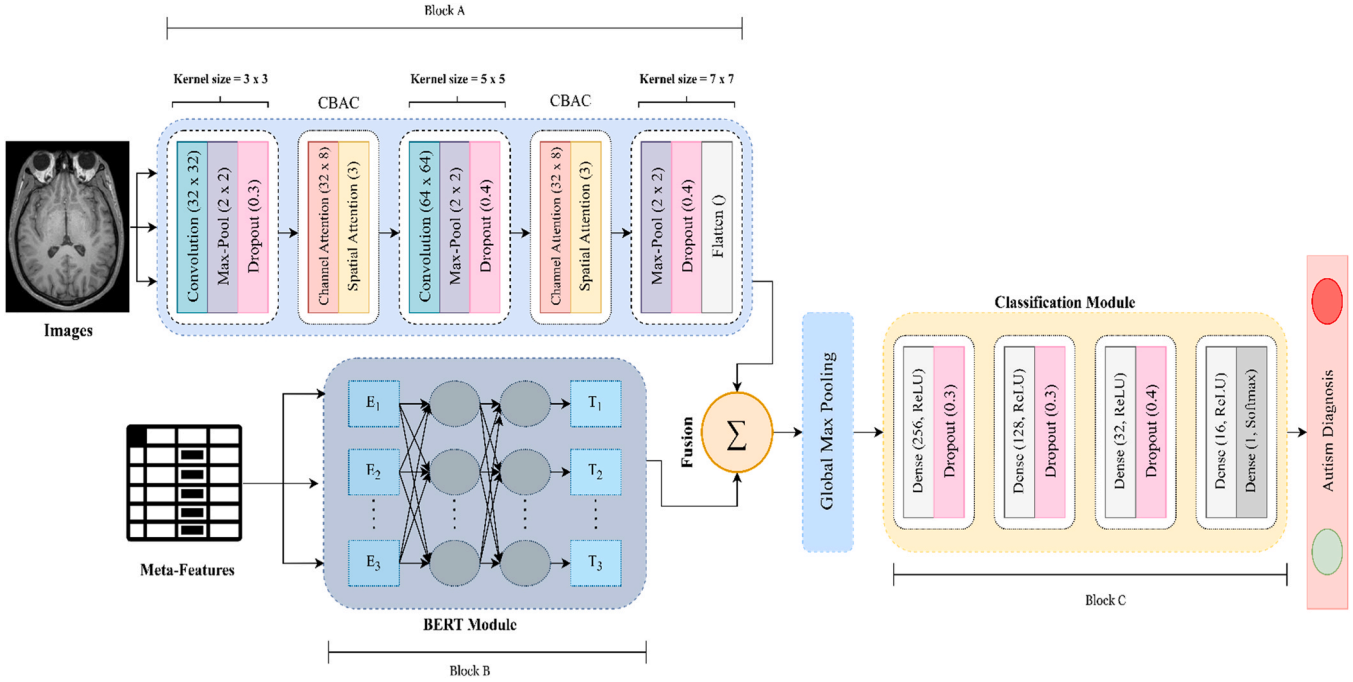


Fig. 2. Multimodal architecture of MCBERT incorporating convolutional layers with channel block attention component (Block A) for image modality, a BERT module (Block B) for the meta-features, fusing the output of the block A and block B, and passing it through global max pooling and the final classification module (Block C) to diagnose ASD.

compression along the spatial dimension, and reciprocally. The CAC network, illustrated in Fig. 3, augments the significance of relevant information while diminishing the weight of unnecessary details in the feature channel. Consequently, the developed module accentuates channels within the MRI images.

Within the CAC, average-pooled patterns and max-pooled attributes are extracted (as in Fig. 3) from the aggregated feature map, employing both average-pooling and max-pooling operations on spatial information (Lyra et al., 2024)(Tan et al., 2024). These high-level patterns undergo processing in a shared multi-layer perceptron (MLP) model, featuring a hidden layer. The outcome of the shared network traverses a pipeline involving additional max-pooling and average-pooling operations, coupled with a non-linear activation function (ReLU) (Vasanthakumari et al., 2023), to generate the channel attention map $A_C \epsilon \mathbb{R}^{c\times1\times1}$. The utilization of two pooling operations enhances the extraction of high-level features (Özbay & Altunbey Özbay, 2023). The mathematical calculation of channel attention is expressed in Eq. (6), where σ represents the sigmoid function.

$$A_C(I) = \sigma(MLP(AvgPool(I))) + (MLP(MaxPool(I))) \tag{6}$$

$$A_s(I') = \sigma\left(f^{7\times7}\left([AvgPool(I'); MaxPool(I')]\right)\right) \tag{7}$$

Moving to the SAC network, depicted in Fig. 3, enhances the spatial dimension features in the feature map through feature filtering on pixels at different positions within the same spatial dimension, assigning weights to significant features. SAC executes average-pooling and max-pooling operations on the feature map I' along the channel dimension, producing two feature maps that are subsequently fused and convolved by a 7 × 7 kernel size. This convolution operation yields the final spatial attention map $A_s \epsilon \mathbb{R}^{1\times H\times W}$. Where 7×7 denotes the convolution operation with a filter size of 7 × 7. The mathematical formulation of SAC is provided in Eq. (7). The extracted feature maps from the multi-head CNN, refined through CBAC, are then combined with patient meta-features for further analysis.

3.5. Bidirectional encoder representations from transformers

BERT or Bidirectional encoder representations from transformers stand out as an efficient and revolutionary model for feature extraction in various tasks. For the meta-features (patient information), we employed the BERT model, a pre-trained language representation model (Fig. 2, Block B). The employed BERT model transforms the input patient data into vector representations. These vectors capture both inter-feature relationships and sentence-level features from the patient information. Its primary function is to transform input into vectors (Tseng et al., 2024). In contrast to conventional language pre-training models, BERT incorporates two tasks for model pre-training. Consequently, the word vectors produced by BERT not only convey inter-word features but also encompass features at the sentence level (Min et al., 2024). The pivotal component in BERT is the Bi-transformer, utilizing a self-attention mechanism and fully connected (FCL) layer to model input, diverging from the use of recurrent neural networks and CNN for feature extraction. The self-attention mechanism, paramount in transformers, computes relationships between the data, adjusting the weight of importance based on these relationships. Thus, each word's vector not only signifies its meaning but also provides insights into relationships with other features (Muizelaar et al., 2024). The BERT module's output is then used as input for the multi-head self-attention mechanism, as described in Eq. (8). The computational process is depicted in Eq. (8), where Q denotes the query vector, K is for the representation vector, the value vector is V, and the input vector dimension is denoted by d_k. Here, the input vectors are derived from the meta-features encoded by BERT, and the attention mechanism computes the relationships between the features, adjusting their weights based on their relevance.

$$Attention(Q, K, V) = softmax\left(\frac{Qk^T}{\sqrt{d_k}}\right)V \tag{8}$$

The transformer also incorporates a multi-attention mechanism as the self-attention mechanism alone is limited to capturing information in a single dimension. Initially, the vectors Q, K, and V undergo linear mapping h times. Finally, the resulting attention matrices are concatenated, enabling the acquisition of multi-dimensional information. The formula describing the process is as follows:

$$Multihead(V, Q, K) = concat(head_1, \quad head_h) \tag{9}$$

$$head_i = attention(QW_i^Q, \quad KW_i^K, \quad VW_i^V) \tag{10}$$

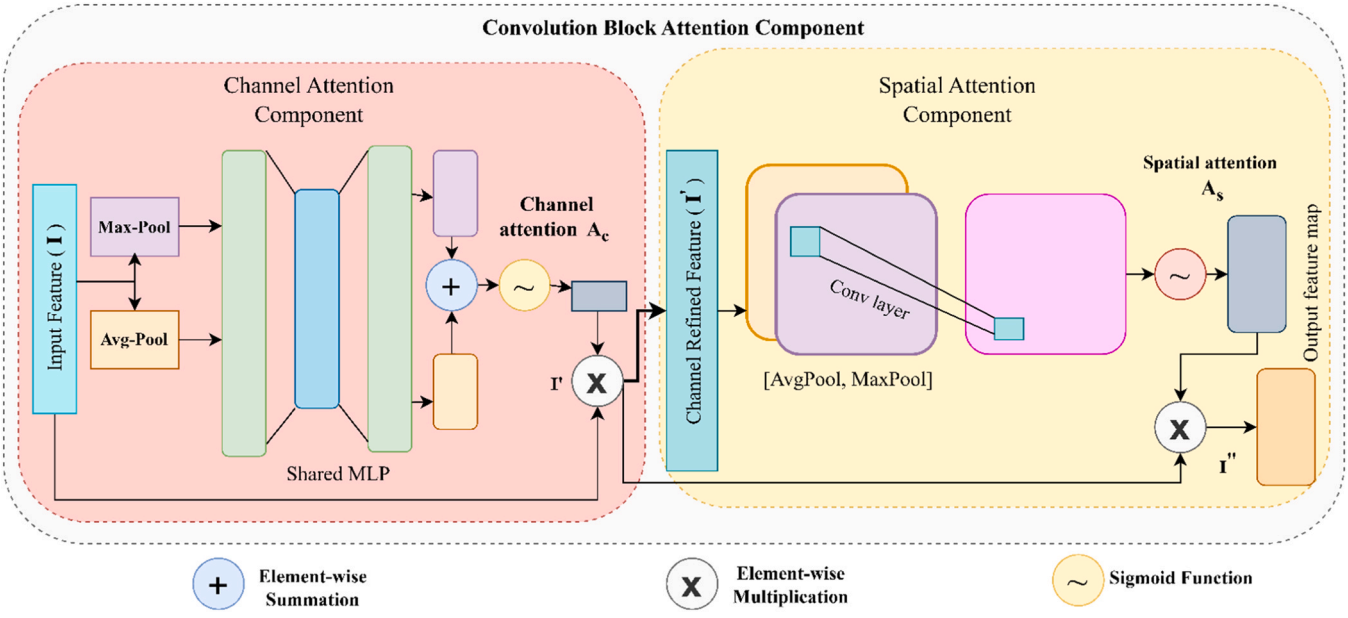


Fig. 3. Detailed architecture of convolution block attention component (CBAC) incorporating the visual representation of channel attention component (CAC) and spatial attention component (SAC).

3.5.1. Pre-processing

To address the distinct statistical properties in our multimodal data during training, we adopted a standardization and normalization approach for meta-features (non-imaging data). Specifically, dictionaries are crafted from the data, encompassing age, site, and gender information for each sample. Age values fall within the range of (6,64), while the 17 sites are encoded as (0, 1, ……,15, 16). And gender is represented as (0,1). A normalization process is applied to both sites and ages, transforming their values to lie within the standardized interval of (0, 1). This meticulous preprocessing ensures that the non-imaging input data is appropriately rescaled and ready for integration with the multimodal data, promoting improved convergence and effectiveness during the training process.

3.6. Classification Module

The combined output vectors generated by the multi-head CNN (Block A) and the BERT module are integrated and fed into the classification module for the diagnosis of ASD through the utilization of global max pooling (GMP). The GMP layer efficiently selects the most salient features and produces a feature map for both the target classes, contributing to the reduction of trainable parameters. Following this, fully connected layers are employed, incorporating neurons with ReLU activation function, specifically 256, 128, 32, and 16 neurons in each layer. To address potential overfitting, dropout layers are strategically inserted in conjunction with these fully connected (FCL) layers. The final stage involves applying the softmax activation function to calculate the class score for both target classes, determining the correct diagnosis result with the high probability score. The representation of the softmax

```
Function normalize_meta_features(u):
    Create an empty dictionary for normalized data
        Normalized_data = {}
    For data_type in ['site', 'age', 'gender']:
        Extract the data for the current_type
            X = u[data_type]
    Normalize the data to the range [0, 1]
            X_normalized = (X - min (X)) / (max (X) – min (X))
    Standardization of the data
            X_standardized = (X – np.mean (X)) / np.std (X)
    Add the normalized data to the dictionary
            Normalized_data [data_type] = X_normalized

Return normalized_data
```

Due to the inherent limitation of the self-attention mechanism in capturing the sequential order of input, BERT introduces position embedding and segment embedding to discern between adjacent sentences. Within the BERT framework, each input variable in its input sequence is derived through the summation of a word vector, a position vector, and a segment vector. The ultimate word vector is produced via a process of deep bidirectional coding, following which it is sent into the classification module mentioned in the section below.

function during diagnosis is formulated in Eqs. (11) and (12). Where ϕ denotes the output features from the preceding FCLs. During training, the cross-entropy loss function ($\mathscr{L}$) is employed to minimize the loss value, as depicted in Eq. (13). Here, y_i signifies the actual classes, and $\widehat{y}_i$ indicates the outcomes through the developed architecture.

$$P = softmax(\phi) = \frac{\exp(\phi)}{f_0^1 \exp(\phi)} \tag{11}$$

Table 5
Phenotypic measure summary of the ABIDE-I dataset.

Site	ASD	TD	Male count	Female count	Average age
CMU	14	13	21	6	26
Caltech	19	18	29	8	27
Leuven	29	34	55	8	18
KKI	20	28	36	12	10
NYU	75	100	139	36	15
MaxMun	24	28	48	4	25
OLIN	19	15	29	5	16
OHSU	12	14	26	0	10
SBL	15	15	30	0	34
PITT	29	27	48	8	18
Stanford	19	20	31	8	9
SDSU	14	22	29	7	14
UCLA	54	44	86	12	13
Trinity	22	25	47	0	16
USM	46	25	71	0	22
UM	66	74	113	27	14
Yale	28	28	40	16	12

$$\widehat{y} = argmax(P) \tag{12}$$

$$\mathscr{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \quad \log(\widehat{y}_i) + (1 - \quad y_i)\log(1 - \quad \widehat{y}_i)] \tag{13}$$

N signify the data samples.

4. Experimental setup and result analysis

In this segment, we explained the experimental outcomes obtained from the multimodal autism spectrum disorder (ASD) diagnosis architecture to present the efficacy of the developed architecture. This segment incorporates sub-sections giving brief descriptions of the experimental configuration, the ASD multimodal dataset employed, performance evaluation metrics considered, the quantitative analysis of results, and the leave-one-site-out-classification test. Furthermore, we conducted a comparative analysis to contrast the findings of our work with various existing state-of-the-art approaches.

4.1. Experimental configuration

All experiments in this study were conducted on a laptop with an Intel Core i5 10th Generation processor, 8 GB of RAM, 512 GB of storage, and running the Windows 11 operating system. The system was also equipped with an NVIDIA GTX 1650 graphics card with 4 GB of VRAM, which was utilized to enhance computational performance, particularly during model training. The experiments were implemented using Python, and several libraries were employed for data analysis and model development. Numpy was used for numerical computations and matrix operations, while Pandas handled data manipulation tasks, including loading and preprocessing datasets. For visualization, Matplotlib and Seaborn were used to plot training results and statistical graphics, respectively. Scikit-learn was applied for model evaluation and computation of performance metrics. These tools and frameworks formed the core of the experimental setup and were integral to the development and evaluation of the proposed model.

4.2. Dataset description

Our research conducted experiments on ABIDE-I, a publicly available data repository. This multi-modal data is gathered from 1112 participants across 17 sites worldwide (64dece7304f468474ed45342880865a4e7eb119d @ fcon_1000.projects.nitrc.org., n.d.). This multimodal dataset comprises (a) MRI scans and comprehensive (b) phenotypic information (as in Table 5) for each subject. These phenotypic measures included demographic information (age, gender), and count of autism spectrum disorder (ASD) to TD participants. We selected these specific measures as they are clinically relevant to understanding ASD, capturing all dimensions essential for a comprehensive analysis. Table 3 represents the phenotypic measure summary of the employed ABIDE dataset. The inclusion of these phenotypic measures complements the imaging data, allowing us to address the heterogeneity observed in ASD. This multimodal approach aligns with our objective of developing a model that integrates both neuroimaging and non-imaging data for better diagnostic accuracy. Our study focuses on resting-state structural MRI (rs-MRI) scans for the imaging part. To maintain data quality and maintain methodological comparability, we meticulously excluded data with missing series (non-imaging part), incomplete brain coverage, and other scanning artifacts. Our analysis ultimately focused on 875 participants, including 403 participants diagnosed with autism spectrum disorder (ASD) and 472 typically developed (TD). Table 6 describes the dataset description with a training and test split ratio of 80:20.

Table 6
Dataset description of the employed ABIDE-I dataset.

Images with ASD	14,105	**Total**	**Training: Testing**
Images of typically developed (without ASD)	16,520	= 30,625	**(80:20)** 24,500: 6125

Table 7
Key classification metrics employed to evaluate the proposed work.

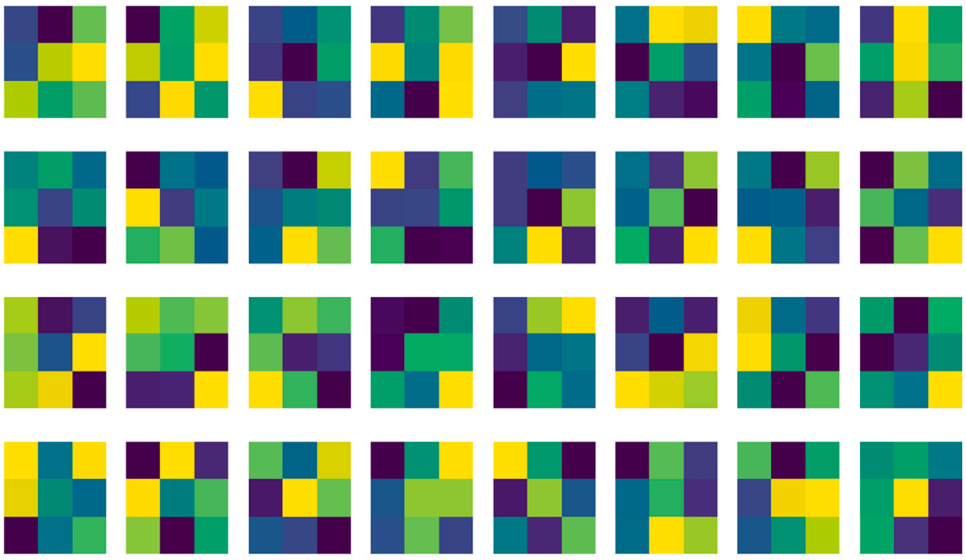
Performance Metric	Formula	Value Range	Cases Assumed
Accuracy	$\frac{TP+TN}{TP+FP+TN+FN}$	[0,1]	***TP***: Autistic individuals identified as autistic individuals. ***TN***: Non-autistic/Healthy individuals identified as non-autistic ***FP***: Non-autistic/Healthy individuals identified as autistic ***FN***: Autistic individuals identified as non-autistic
Sensitivity (Recall)	$\frac{TP}{TP+FN}$		
Specificity	$\frac{TN}{TN+FP}$		

4.3. Performance evaluation metrics

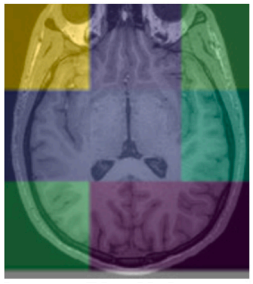
The efficacy of the MCBERT model is evaluated using the three primary metrics, namely sensitivity, accuracy, and specificity. Table 7 shows the metrics and their respective formulas.

4.4. Result analysis

The performance of our developed approach, which leverages BERT for meta-feature extraction and a multi-head CNN for image feature extraction, followed by the fusion of their outputs and passing them to a classification module, yielded promising results. Through the methodology outlined in Section 4.3, we aimed to enhance the diagnostic process by incorporating rich contextual embeddings from the BERT module and extracting spatial and channel-specific features through the multi-head CNN. We conducted experiments for approximately 100 epochs to both train and assess the performance of the developed MCBERT architecture. The use of BERT for meta-feature extraction allowed the model to capture complex semantic relationships between patient metadata and brain MRI data, which proved beneficial in enhancing diagnostic accuracy. The output from the multi-head CNN is illustrated in Figs. 4 and 5. Fig. 4 demonstrates the activation patterns captured by the initial convolutional layer of the multi-head CNN, which includes our Convolution Attention Block (CAB) module, during the processing of brain MRI images where each matrix represents the output from distinct filters of the initial convolutional layer, visualizing how different regions of the brain MRI images are processed by the CNN. The inclusion of multiple matrices serves to illustrate the model's ability to



(a)



(b)

Fig. 4. (a), (b). The output obtained via the activation pattern learned in the initial convolutional layer when the brain images are passed by block A.

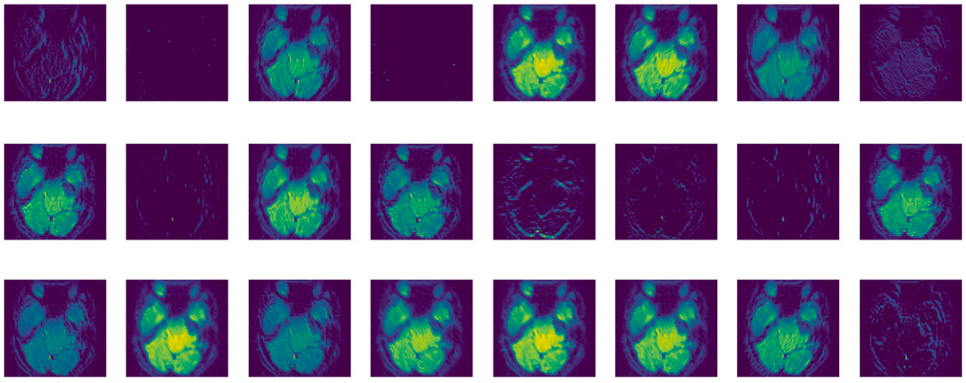
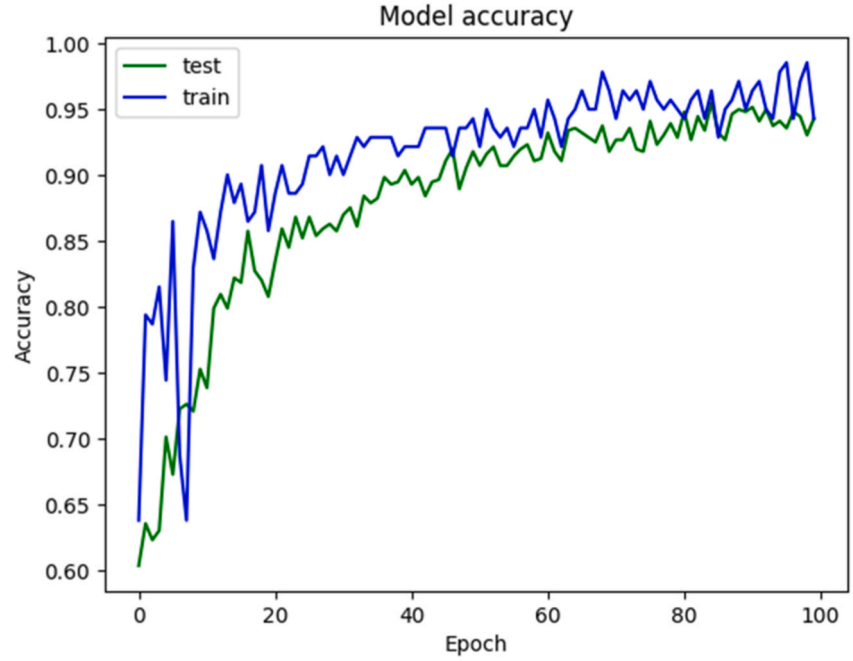


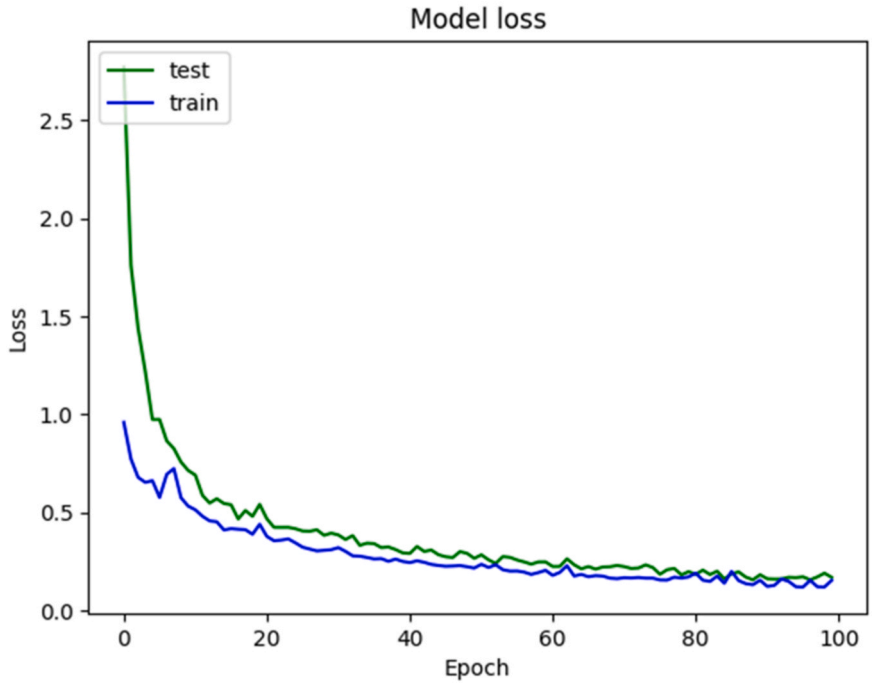
Fig. 5. Visual representation of the feature maps/feature extraction capability of multi-head CNN when applied to brain images.

focus on various diagnostic regions simultaneously, enabling it to capture complementary information for accurate classification. The activation maps are represented using a color gradient, where blue indicates low activation, green denotes medium activation, and yellow represents high activation levels. These color-coded maps show how the CNN's filters respond to specific regions of the input image. Early convolutional layers in CNNs generally focus on detecting fundamental image features, such as edges or anatomical structures. In Fig. 4(a) and (b), the high activations (yellow) in certain regions highlight key anatomical boundaries, edges, or structures that are likely to hold diagnostic significance, such as cortical boundaries or ventricles. These regions are critical for constructing hierarchical representations of the data as it passes through deeper layers of the network. Conversely, the medium (green) and low (blue) activation areas correspond to regions with less prominent or diagnostic features.

Over approximately 100 epochs, the model refined its feature extraction process. Early in the training, activations in the initial layers are more generalized, but as training progresses, these patterns become more focused, allowing the network to identify the most informative features for the diagnostic task. The activation maps presented in Fig. 5 provide visual evidence of this learning process, showing how the model transitions from emphasizing simple, low-level patterns, such as edges,



(a) Epoch vs Accuracy curve of MCBERT



(b) Epoch vs Loss curve of MCBERT

Fig. 6. (a) Epoch vs Accuracy curve of MCBERT. (b) Epoch vs Loss curve of MCBERT.

to capturing more complex features that contribute to the improved classification accuracy of the MCBERT architecture. This hierarchical feature extraction is typical of CNNs, where initial layers detect basic patterns, and deeper layers identify more abstract, high-level features. The observed activation patterns are crucial for understanding how the multi-head CNN processes spatial and channel-specific features from the MRI data.

By activating different filters in response to specific brain structures or abnormalities, the network demonstrates its ability to capture spatial relationships between anatomical structures, which directly supports the effectiveness of our approach in improving diagnostic accuracy. Additionally, the presence of distinct activations in key regions of the MRI images suggests that the model is effectively identifying features associated with autism. Fig. 5 further illustrates the contributions of each of the three heads in the multi-head CNN in the initial learning layers. Each head processes the MRI image through distinct convolutional paths, allowing the model to extract diverse, complementary features from different regions of the input, such as anatomical

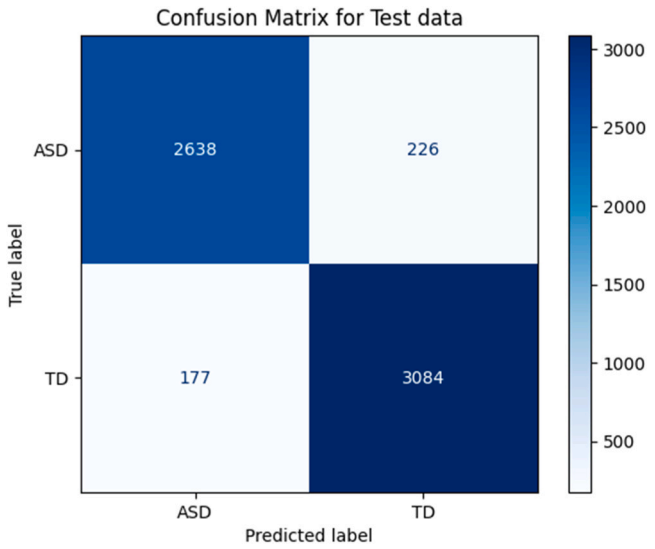


Fig. 7. Confusion matrix obtained on the test set.

structures, textures, and abnormalities. After passing through convolutional and global max pooling layers, the extracted feature maps are reduced to fixed-size vectors, preserving critical information for classification. The diverse activation patterns across the three heads indicate that the multi-head CNN is capable of focusing on various aspects of the MRI images, which is essential for identifying subtle differences in brain structures that could be associated with ASD. These visualizations effectively demonstrate the hierarchical nature of feature extraction in the multi-head CNN, where early layers detect simple patterns like edges, while deeper layers capture more complex, high-level features. In summary, the activation patterns and feature map displayed in Fig. 4 and Fig. 5 serve as a clear indication of the multi-head CNN's capability to identify and focus on relevant image features during the early stages of the feature extraction process. These early activations lay the groundwork for more sophisticated feature identification in deeper layers, which ultimately enhances the model's diagnostic accuracy. The figure provides valuable insights into how the convolutional layers of the CAB module interact with the MRI data and highlights the model's ability to focus on important diagnostic features. This contributes to the overall success of the MCBERT architecture in accurately classifying MRI images, demonstrating the robustness of the proposed approach in improving neuroimaging-based diagnosis.

Upon analyzing the experimental outcomes, we observed significant improvement in accuracy as compared to existing techniques (comparison with existing techniques is mentioned in the further sub-sections). Fig. 6(a) represents the accuracy vs epoch curve demonstrating a diagnosing accuracy of 93.4 %. Fig. 6(b) represents the loss curve of the developed architecture. Overall, our results demonstrate the effectiveness of the developed multi-modal MCBERT, a BERT, multi-head CNN, and its seamless integration with the classification module for the diagnosis of autism spectrum disorder. This approach not only showcases the power of utilizing pre-training architectures but also showcases the potential of deep learning in feature extraction and further advancement in the healthcare domain. Fig. 7 illustrates the performance of the developed model through the confusion matrix obtained from the test data. The test set consists of 6125 images (as mentioned in Table 6).

4.4.1. Comparison with existing works

This section highlights a comparison of various methods used for autism spectrum disorder (ASD) diagnosis on the ABIDE dataset, focusing on the performance of the proposed MCBERT model against other state-of-the-art techniques. The quantitative results of this comparison are summarized in Table 8. Mishra & Pati, (2023) proposed an sMRI-based ASD detection framework using an ensemble of deep convolutional neural networks (DCNN) combined with different optimizers (Adam, Nadam, and RMSProp). After filtering out noisy slices, the study utilized raw sMRI scans from the ABIDE dataset without advanced preprocessing. The ensemble of optimizers aimed to enhance model performance by improving robustness. They tested the model using three data splits (70:30, 80:20, and 90:10) and achieved accuracies of 77.58 %, 77.66 %, and 81.35 %, respectively. Y. Tang et al., (2023) proposed a two-stage adversarial learning model to address the challenges associated with multi-site ASD classification using resting-state functional magnetic resonance imaging (rs-fMRI). Their approach begins with the sliding window sampling technique, which preserves spatial and temporal information from rs-fMRI data. This is followed by an adversarial learning model that extracts site-shared features, effectively tackling the issue of site heterogeneity common in multi-site studies. The model is then fine-tuned to extract disease-related features, specific to ASD classification. The data used in this study was sourced from the ABIDE dataset. For model evaluation, the authors employed ten-fold cross-validation, with the dataset randomly split into training (81 %), validation (9 %), and test sets (10 %). Mingzhi Wang et al., (2023) developed a WL-DeepGCN framework that combines fMRI data and non-imaging demographic information for ASD diagnosis. The model uses a weight-learning network to define graph edge weights in the latent space, and residual connections in the GCN to avoid gradient issues. An edge-drop strategy reduces overfitting by sparsifying node connections. The study applied a nested 10-fold cross-validation on the ABIDE-I dataset to ensure robust evaluation, avoiding feature peeking and overfitting. Recursive feature elimination (RFE) was used for feature selection. Yang et al., (2022) conducted a comprehensive review of different brain networks and their functional connectivity to distinguish between individuals with ASD and TD participants. The study utilized 871 rs-fMRI samples from the ABIDE repository. The authors employed bootstrap analysis of stable clusters (BASC) as the most predictive brain parcellation technique, aiming to find the optimal method for classifying ASD. The methodology involved exploring eight different brain

Table 8
Comparison with existing works conducted for ASD on the ABIDE dataset.

Reference	Methodology	Dataset	Modalities Incorporated	Best accuracy	Sen	Spec
(Mishra & Pati, 2023)	Optimizer + Deep CNN	sMRI	1	77.58 %	78.16 %	76.99 %
(Rakhimberdina et al., 2020)	Graph NN + Ensemble technique	Phenotypic + HO	2	73.13 %	76.00 %	69.00 %
(Y. Tang et al., 2023)	Adversarial learning + LSTM	fMRI	1	80.00 %	81.00 %	80.00 %
(Mingzhi Wang et al., 2023)	Weight learning + Graph CNN + Deep CNN	Phenotypic + HO	2	77.27 %	80.96 %	-
(Herath et al., 2021)	Inception V3	fMRI	1	98.35 %	-	-
(Herath et al., 2024)	Inception V3 + ResNet50 + DenseNet + MobileNet	fMRI + Phenotypic	2	97.82 %	-	-
(Yang et al., 2022)	Feature extraction via function connectivity matrix	fMRI	1	69.43 %	64.57 %	73.61 %
(Rakić et al., 2020)	MLP + Autoencoder	sMRI + CC200	2	85.06 %	81.00 %	89.00 %
(Parui et al., 2023)	Correlation matrix + Graph Theory	fMRI	1	84.79 %	89.63 %	78.96 %
MCBERT (Proposed)	***Multi-Head CNN + BERT***	***Phenotypic + sMRI***	***2***	***93.4 %***	***92.1 %***	***94.5 %***

parcellation techniques, which included structural, functional, and data-driven approaches, to identify the best brain atlas for ASD classification. Additionally, three functional connectivity metrics, correlation, partial correlation, and tangent space, were evaluated to assess their stability and efficiency. The study found that the correlation metric was the most stable among the metrics. In terms of machine learning models, the paper compared four supervised learning algorithms: kernel Support Vector Machine (kSVM), which was identified as the optimal classifier for the task, outperforming others. The experiments used 5-fold cross-validation, repeated 10 times to ensure the reliability and stability of the results. Herath et al., (2021) worked on an unimodal ASD identification architecture incorporating Inception V3 model with CNN. They worked on fMRI employed from the ABIDE dataset. They took three imaging features into account i.e., epi images, glass brain images, and stat map images. In another work by Herath et al., (2024), they focused on developing multimodal ASD architecture incorporating transfer learning with deep ensemble learning. They developed a multimodal-multisite ensemble classifier to diagnose ASD from fMRI and phenotypic information from the ABIDE dataset. They tested their model on various parameters and presented a detailed analysis of their work. Rakić et al., (2020) focused on using a combination of functional and structural MRI data for the classification of ASD patients versus control participants. The key features used included functional connectivity patterns among brain regions from fMRI and volumetric correspondences of gray matter volumes from sMRI. Their classification network was built using stacked autoencoders trained in an unsupervised manner, combined with multilayer perceptrons (MLP) trained in a supervised manner. The study analyzed data from 817 cases in the ABIDE-I dataset, involving 368 ASD patients and 449 controls. The evaluation methodology involved 10-fold cross-validation, wherein each fold, 10 % of the data was used for testing, while 90 % was used for training and validation (split into 70 % for training and 30 % for validation). Additionally, they conducted leave-one-site-out cross--validation to assess the model's performance. This paradigm, alongside reporting of accuracy, sensitivity, and specificity, provided a thorough quantitative and qualitative comparison with other state-of-the-art methods.

Parui et al., (2023) utilized rs-fMRI data from the ABIDE-I dataset to propose an approach for diagnosing ASD. The study focuses on constructing functional connectivity networks from the rs-fMRI time-series data, calculating correlation matrices that represent interactions between brain regions. The ABIDE-I dataset, consisting of 1112 individuals (539 ASD and 573 typically developing controls), served as the basis for the experiments. The authors tested 11 classification algorithms, including linear support vector machines (SVM) and 2D CNN, and identified these as the best-performing methods across all atlases. Additionally, the authors also performed stratified 10-fold and 3-fold cross-validation on the best classifiers (Linear SVM and 2D CNN), observing consistent accuracy across these methods. Most of the existing works in the literature have utilized deep learning frameworks. Some approaches focus on single-modality data, such as structural MRI (sMRI) or functional MRI (fMRI), while others combine multiple data modalities, like phenotypic information and neuroimaging data. Our proposed MCBERT model, which integrates a Multi-Head CNN and BERT architecture, operates on multimodal inputs, specifically phenotypic data and sMRI. As shown in Table 8, the MCBERT model outperforms other methods in terms of accuracy, sensitivity, and specificity, which demonstrates its effectiveness in ASD diagnosis. The model's ability to handle both phenotypic and sMRI data contributes to its robust performance, leading to higher classification accuracy compared to the methods that rely on single-modality data.

4.4.2. Leave-one-site-out (LOSO) cross-validation test

In this study, the primary experimental paradigm utilized was leave-one-site-out (LOSO) cross-validation. This method was chosen to evaluate the generalization ability of the MCBERT model across different screening sites within the ABIDE-I dataset, which includes data from 17 different sites. For each LOSO iteration, one site was selected as the test set, while the remaining sites were split into training and validation sets. This setup allowed the model to be tested on unseen data from various sites, highlighting its adaptability to site-specific variations in the dataset. Each site/data was trained and tested under identical conditions, and performance metrics, including accuracy, specificity, sensitivity, and AUC, were recorded for each site. Table 9 presents the performance outcomes of the LOSO test, demonstrating the robustness of MCBERT in generalizing across different sites. The mean accuracy of MCBERT across all sites was determined to be 83.64 %. Notably, four sites UM, STANFORD, PITT, and MAX_MUN exhibited lower performance compared to the mean values of the evaluated metrics. This observation underscores the presence of site-specific variability and a lack of homogeneity in the dataset. Despite these variations, the high global mean values attest to the effectiveness of the MCBERT architecture.

4.4.3. Ablation study

This section presents an ablation study to validate the contribution of the proposed MCBERT architecture on the multimodal ABIDE-I dataset. In Case A (baseline model without attention mechanisms) both the channel attention component (CAC) and spatial attention component (SAC) are removed from the multi-head CNN architecture. The model relies solely on core convolutional layers without attention mechanisms to process the input data. This case aims to quantify the contribution of

Table 9
Quantitative performance analysis of the MCBERT model on the LOSO test using the ABIDE-I dataset.

Site	Accuracy	Specificity	Sensitivity	AUC
CMU	91.00	95.00	87.00	84.00
CALTECH	88.50	88.00	86.00	85.00
MAX_MUN	79.30	80.00	78.70	76.00
LEUVEN	85.00	86.00	84.01	87.00
KKI	86.08	84.03	91.03	90.03
OHSU	84.00	83.79	83.46	77.31
NYU	86.02	80.81	82.55	88.61
OLIN	89.50	87.62	91.52	89.01
SDSU	87.00	80.34	81.53	81.34
PITT	76.09	75.50	75.00	74.50
SBL	89.60	86.02	89.04	87.57
STANFORD	78.00	78.42	77.67	86.40
UCLA	80.50	82.09	78.52	79.66
TRINITY	81.07	82.41	80.54	82.24
USM	87.30	87.80	86.80	90.30
UM	75.40	76.03	75.51	76.06
YALE	87.50	82.47	82.47	79.84
Mean	**83.64**	**81.96**	**80.81**	**83.86**

Table 10
Ablation outcomes with the proposed MCBERT architecture.

Dataset	Case	Case description	Accuracy	Specificity	Sensitivity
ABIDE-I	A	Without attention mechanism i.e. CAC and SAC	72.9	81.2	81.8
	B	Without channel attention (CAC)	85.3	84.8	85.1
	C	Without spatial attention (SAC)	83.4	82.6	83.2
	D	Without BERT module (Image-only model)	78.3	78.2	78.3
	E	BERT module only	73.1	72.6	84.9
	F	Without global max pooling (GMP)	91.6	90.1	91.2
	G	Complete architecture (MCBERT)	***93.4 %***	***92.1 %***	***94.5 %***

attention mechanisms by comparing the model's performance with a standard CNN, allowing us to isolate the effect of the attention modules on classification performance. For case B (without channel attention) the channel attention component (CAC) is disabled while the spatial attention component (SAC) remains active. This setup focuses on analyzing the spatial features of the MRI data. This experiment aims to evaluate the importance of channel-specific attention. It provides insight into whether focusing on channel-specific features significantly impacts the model's ability to classify ASD. Similarly, for case C (without spatial attention) the spatial attention component (SAC) is removed, while the channel attention component (CAC) remains active. This setup assesses the model's performance when spatial patterns are not specifically highlighted. The focus here is to understand the role of spatial attention in identifying relevant spatial features from MRI images and determine its contribution to the model's overall performance. For the case D (without the BERT module) the BERT module, which processes meta-features, is removed entirely. The model uses only the multi-head CNN to process the MRI image data without leveraging any meta-feature information. The purpose of this experiment is to evaluate how much of the model's success is attributable to the BERT-processed meta-features. It will show whether the image data alone is sufficient to achieve high diagnostic accuracy or if meta-features play a crucial role.

In case E (BERT module only) the multi-head CNN is removed it explores the performance of the model when only meta-features are used for classification, without the additional information provided by the MRI images. It allows an assessment of the relative value of meta-feature data compared to image data in ASD classification. For case F (without global max pooling) the global max pooling (GMP) layer is removed from the architecture and replaced with the average pooling. The goal here is to determine the significance of the GMP layer in selecting the most salient features before the fully connected layers. It helps assess whether the GMP layer plays a critical role in the final classification performance by maximizing key features. The work evaluates the performance across the cases mentioned in Table 10. At last, case G refers to the performance of the complete architecture i.e., MCBERT.

4.5. Computational complexity

The computational complexity of the proposed model can be described in terms of the dominant operations in its architecture, including convolutional layers, attention mechanisms, and BERT. The convolutional layers, responsible for processing image data, contribute a complexity of $O(N^2)$, where N is the spatial dimension of the input (MRI images). This quadratic complexity arises from input size, number of channels, and filter sizes in the convolution operations. The channel and spatial attention mechanisms, which operate on feature maps, also scale linearly with the number of channels and spatial dimensions but remain dominated by the $O(N^2)$ behavior. Additionally, the BERT component, used for processing meta-features, introduces a complexity of $O(L^2)$, where L is the sequence length, reflecting the quadratic nature of the self-attention mechanism. As a result, the overall computational complexity of the model is approximately $O(N^2 + L^2)$, with the convolutional layers typically dominating for large image inputs, while the BERT module adds significant complexity depending on the length of the meta-feature sequences. This combined quadratic complexity is characteristic of deep learning models utilizing both convolution and attention mechanisms.

Generally, BERT requires high computational demands, but several strategies could be employed to reduce the computational needs. One approach is to use model compression techniques such as pruning and quantization, which can reduce the number of parameters without significantly impacting model performance. Additionally, lighter versions of BERT, such as DistilBERT or ALBERT, could be considered, as they retain most of the model's accuracy while offering reduced complexity. Furthermore, implementing mixed-precision training or utilizing distributed training frameworks may also help optimize computational resource usage. These techniques, in combination, can effectively reduce the overall processing load while maintaining the efficacy of the models.

4.6. Discussion

4.6.1. Study contributions

In this study, we proposed a novel multimodal architecture, MCBERT, for diagnosing autism spectrum disorder (ASD) by integrating brain MRI images and meta-features such as gender, behavioral characteristics, and patient history. Our model fuses a Multi-Head CNN (MCNN) with bidirectional encoder representations from transformers (BERT) to capture both spatial and channel attributes from the image modality, while efficiently handling the high dimensionality of meta-features. The results demonstrated that MCBERT achieves high diagnostic accuracy, with an overall accuracy of 93.4 %, surpassing other state-of-the-art systems.

The key contributions of this study include the development of a novel fusion technique that integrates multimodal data for ASD diagnosis. By combining the strengths of CNN for processing brain MRI images and BERT for extracting meaningful information from meta-features, we were able to achieve superior performance. Additionally, the incorporation of convolutional block attention components (CBAC) enhances the model's ability to capture spatial and channel attributes, further improving its diagnostic power. The use of leave-one-site-out (LOSO) cross-validation provided a rigorous assessment of the model's ability to generalize across different data sites, which is crucial for ensuring the robustness of ASD diagnostic models in real-world clinical settings.

4.6.2. Challenges and future directions

Despite these contributions, the proposed MCBERT architecture has several limitations that should be acknowledged. Firstly, the model only utilizes structural MRI (sMRI) data and does not include functional MRI (fMRI) data, which captures brain activity and could provide deeper insights into the neural mechanisms associated with ASD. By focusing solely on sMRI, the model may overlook important functional abnormalities that are often present in individuals with ASD. Furthermore, the study is limited to the ABIDE-I dataset, which constrains the ability to generalize the findings to other datasets or populations. The diversity of ASD manifestations across different groups means that relying on a single dataset could limit the model's applicability in broader clinical settings. Additionally, the reliance on pre-existing meta-features, which are not universally standardized across datasets, introduces potential variability in model performance when applied to new data.

Looking ahead, there are several promising directions for future work. One of the key areas for expansion is the inclusion of fMRI data in conjunction with sMRI, which would allow for a more comprehensive analysis of both structural and functional aspects of the brain. Exploring the combination of these two imaging modalities could enhance diagnostic accuracy and provide a more detailed understanding of ASD's underlying neural mechanisms. Additionally, extending the model to the ABIDE-II dataset and other large, multimodal datasets would enable further validation of the model's generalizability across different populations. The incorporation of advanced hybrid networks, combining convolutional and transformer-based architectures, could also lead to improved performance in early ASD detection. These advancements hold the potential to refine ASD diagnosis and contribute to the development of personalized treatment plans based on a more thorough understanding of each individual's neurodevelopmental profile.

4.6.3. Real-world applicability

In terms of real-world deployment, the MCBERT model shows strong potential for integration into clinical workflows, provided that certain developments are made. Future work should focus on adapting the

model to handle large-scale, real-time clinical data, ensuring that it meets regulatory standards and is interpretable by medical professionals. This could involve refining the model's output to provide clear, actionable insights that clinicians can easily integrate into their decision-making process. Additionally, integrating the model into existing hospital information systems or diagnostic software platforms would help streamline its adoption in clinical practice. By focusing on these areas, MCBERT could be positioned as a supportive diagnostic tool for healthcare professionals, facilitating more efficient and accurate ASD diagnoses. The model's multimodal approach, which combines brain imaging with meta-features, offers a robust framework that is aligned with the growing trend of precision medicine and personalized healthcare.

Another critical step toward real-world application is the validation of the model's results with input from medical experts. Collaboration with neurologists, radiologists, and other healthcare professionals specializing in ASD is essential for establishing the clinical credibility of the model. Future studies should focus on comparing the model's predictions with expert diagnoses to ensure its reliability in a clinical setting. Expert feedback could also be invaluable in refining the model further, particularly in cases where subtle patterns in the data might lead to misclassification. This validation process would not only enhance the model's accuracy but also foster trust among healthcare providers, increasing the likelihood of its integration into routine clinical practice.

In summary, the MCBERT architecture is well-positioned to be adopted as a real-world clinical tool. With further validation and refinement, particularly in the areas of regulatory compliance, scalability, and expert validation, the model could play a significant role in improving the early diagnosis of ASD. These advancements would ultimately contribute to better patient outcomes, supporting early interventions and more personalized treatment plans for individuals with ASD. By addressing these limitations and expanding the scope of the research, future work aims to push the boundaries of ASD diagnosis through more advanced multimodal deep learning techniques.

5. Conclusion

Healthcare analysis represents a complex domain of study, with a specific focus on intricate conditions like autism spectrum disorder (ASD), demanding meticulous attention due to their intricacy and the critical necessity for precise diagnostic support. Prior research has emphasized autism diagnosis biomarkers through various machine-learning strategies and deep-learning algorithms using single-modality data. The works focusing on multi-modality data for ASD are still limited. The amalgamation of multi-modality methodologies in the realm of autism spectrum disorder can be advantageous in enhancing diagnostic accuracy. This study introduces a novel methodology, MCBERT, a multi-head CNN (MCNN) and BERT-based network with brain MRI and meta-features dataset tailored for autism spectrum disorder diagnosis. The architecture further includes two attention components (CBAC) to learn spatial (SAC) and channel (CAC) features. The output of BERT and MCNN is then fused and fed to the classification module to get the final output. We employed the ABIDE-I dataset (multimodal dataset) to evaluate the architecture. Performance evaluation revolves around key parameters, including accuracy, sensitivity, and specificity, and we also conducted a leave-one-site-out-classification test offering a comprehensive assessment of the model's effectiveness in accurately identifying individuals with ASD. Experimental simulations reveal that MCBERT outperforms existing works, achieving a notable accuracy of 93.4 %. These results underscore the effectiveness of integrating multi-modal data with deep learning methodologies. The research aims to broaden its scope by exploring additional techniques in conjunction with other deep learning algorithms, anticipating future advancements in the efficacy and accuracy of ASD diagnosis.

Ethical consent

None

Funds

No funding was received

Declaration of Generative AI and AI-assisted technologies in the writing process

None.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

64dece7304f468474ed45342880865a4e7eb119d @ fcon_1000.projects.nitrc.org, 64dece7304f468474ed45342880865a4e7eb119d @ fcon_1000.projects.nitrc.org. (n.d.). https://fcon_1000.projects.nitrc.org/indi/abide/.

Abbas, S. Q., Chi, L., & Chen, Y. P. P. (2023). DeepMNF: Deep multimodal neuroimaging framework for diagnosing autism spectrum disorder. *Artificial Intelligence in Medicine, 136*(December 2022), Article 102475. https://doi.org/10.1016/j.artmed.2022.102475

Akter, T., Khan, M. I., Ali, M. H., Satu, M. S., Uddin, M. J., & Moni, M. A. (2021). Improved machine learning based classification model for early autism detection. *International Conference on Robotics, Electrical and Signal Processing Techniques*, 742–747. https://doi.org/10.1109/ICREST51555.2021.9331013

Alpar, O. (2023). A mathematical fuzzy fusion framework for whole tumor segmentation in multimodal MRI using Nakagami imaging. *Expert Systems with Applications, 216* (July 2022), Article 119462. https://doi.org/10.1016/j.eswa.2022.119462

Alsaade, F. W., & Alzahrani, M. S. (2022). Classification and detection of autism spectrum disorder based on deep learning algorithms. *Computational Intelligence and Neuroscience*, *2022*. https://doi.org/10.1155/2022/8709145

Çetintaş, D., Tuncer, T., & Ç ınar, A. (2023). Detection of autism spectrum disorder from changing of pupil diameter using multi-modal feature fusion based hybrid CNN model. *Journal of Ambient Intelligence and Humanized Computing, 14*(8), 11273–11284. https://doi.org/10.1007/s12652-023-04641-6

Chan, R. Y. Y., Wong, C. M. V., & Yum, Y. N. (2023). Predicting behavior change in students with special education needs using multimodal learning analytics. *IEEE Access, 11*(June), 63238–63251. https://doi.org/10.1109/ACCESS.2023.3288695

Chen, Y., Yan, J., Jiang, M., Zhang, T., Zhao, Z., Zhao, W., Zheng, J., Yao, D., Zhang, R., Kendrick, K. M., & Jiang, X. (2022). Adversarial learning based node-edge graph attention networks for autism spectrum disorder identification. *IEEE Transactions on Neural Networks and Learning Systems*, 1–12. https://doi.org/10.1109/TNNLS.2022.3154755

Ciceri, T., Squarcina, L., Giubergia, A., Bertoldo, A., Brambilla, P., & Peruzzo, D. (2023). Review on deep learning fetal brain segmentation from magnetic resonance images. *Artificial Intelligence in Medicine, 143*(December 2022), Article 102608. https://doi.org/10.1016/j.artmed.2023.102608

Dc, S., Gadgay, B., Farheen, S., & Waheed, M.A. (2022). A Machine Learning Approach for Early Detection and Diagnosis of Autism and Normal Controls and Estimating Severity Levels Based on Face Recognition. 2022 International Conference on Emerging Trends in Engineering and Medical Sciences, ICETEMS 2022, 35–40. https://doi.org/10.1109/ICETEMS56252.2022.10093412.

De Silva, S., Dayarathna, S., Ariyarathne, G., Meedeniya, D., Jayarathna, S., & Michalek, A. M. P. (2021). Computational Decision Support System for ADHD Identification. *International Journal of Automation and Computing, 18*(2), 233–255. https://doi.org/10.1007/s11633-020-1252-1

Du, Y., He, X., Kochunov, P., Pearlson, G., Hong, L. E., van Erp, T. G. M., Belger, A., & Calhoun, V. D. (2022). A new multimodality fusion classification approach to explore the uniqueness of schizophrenia and autism spectrum disorder. *Human Brain Mapping, 43*(12), 3887–3903. https://doi.org/10.1002/hbm.25890

Egger, J., Gsaxner, C., Pepe, A., Pomykala, K. L., Jonske, F., Kurz, M., Li, J., & Kleesiek, J. (2022). Medical deep learning—A systematic meta-review. *Computer Methods and Programs in Biomedicine, 221*, Article 106874. https://doi.org/10.1016/j.cmpb.2022.106874

El Mouatasim, A., & Ikermane, M. (2023). Control learning rate for autism facial detection via deep transfer learning. *Signal, Image and Video Processing, 17*(7), 3713–3720. https://doi.org/10.1007/s11760-023-02598-9

Elakkiya, M. K., & Dejey. (2024). Novel deep learning models with novel integrated activation functions for autism screening: AutiNet and MinAutiNet. *Expert Systems with Applications, 238*(PD), Article 122102. https://doi.org/10.1016/j.eswa.2023.122102

Emon, M.U., Keya, M.S., Sozib, A.R., Islam, S., Imran, F.A., & Zannat, R. (2021). *A Comparative Analysis of Autistic Spectrum Disorder (ASD) Disease for Children using ML Approaches*. *03*, 121–126.

Fu, X., Patrick, E., Yang, J. Y. H., Feng, D. D., & Kim, J. (2023). Deep multimodal graph-based network for survival prediction from highly multiplexed images and patient variables. *Computers in Biology and Medicine, 154*(February). https://doi.org/10.1016/j.compbiomed.2023.106576

Gracia, S. (2022). ScienceDirect ScienceDirect Algorithmic Algorithmic Approaches Approaches to to Classify Classify Autism Autism Spectrum Spectrum Disorders: Disorders: Perspective A Research Perspective. *Procedia Computer Science, 201*, 470–477. https://doi.org/10.1016/j.procs.2022.03.061

Han, J., Jiang, G., Ouyang, G., & Li, X. (2022). A Multimodal Approach for Identifying Autism Spectrum Disorders in Children. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 30*, 2003–2011. https://doi.org/10.1109/TNSRE.2022.3192431

Haputhanthri, D., Brihadiswaran, G., Gunathilaka, S., Meedeniya, D., Jayarathna, S., Jaime, M., & Harshaw, C. (2020). Integration of facial thermography in EEG-based classification of ASD. *International Journal of Automation and Computing, 17*(6), 837–854. https://doi.org/10.1007/s11633-020-1231-6

Hasan, M. M., Watling, C. N., & Larue, G. S. (2024). Validation and interpretation of a multimodal drowsiness detection system using explainable machine learning. *Computer Methods and Programs in Biomedicine, 243*(October 2023). https://doi.org/10.1016/j.cmpb.2023.107925

Herath, L., Meedeniya, D., Marasingha, J., & Weerasinghe, V. (2022). Optimize Transfer Learning for Autism Spectrum Disorder Classification with Neuroimaging: A Comparative Study. ICARC 2022 - 2nd International Conference on Advanced Research in Computing: Towards a Digitally Empowered Society, 171–176. https://doi.org/10.1109/ICARC54489.2022.9753949.

Herath, L., Meedeniya, D., Marasingha, M.A.J.C., & Weerasinghe, V. (2021). Autism spectrum disorder diagnosis support model using Inception V3. Proceedings - International Research Conference on Smart Computing and Systems Engineering, SCSE 2021, *4*, 1–7. https://doi.org/10.1109/SCSE53661.2021.9568314.

Herath, L., Meedeniya, D., Marasinghe, J., Weerasinghe, V., & Tan, T. (2024). Autism spectrum disorder identification using multi-model deep ensemble classifier with transfer learning (December) *Expert Systems, 2022*, 1–23. https://doi.org/10.1111/exsy.13623.

Jaafar, N., & Lachiri, Z. (2023). Multimodal fusion methods with deep neural networks and meta-information for aggression detection in surveillance. *Expert Systems with Applications, 211*(July 2022), Article 118523. https://doi.org/10.1016/j.eswa.2022.118523

Joudar, S. S., Albahri, A. S., & Hamid, R. A. (2023). Intelligent triage method for early diagnosis autism spectrum disorder (ASD) based on integrated fuzzy multi-criteria decision-making methods. *Informatics in Medicine Unlocked, 36*(ember 2022), Article 101131. https://doi.org/10.1016/j.imu.2022.101131

Kang, L., Chen, J., Huang, J., & Jiang, J. (2022). Autism spectrum disorder recognition based on multi-view ensemble learning with multi-site fMRI. *Cognitive Neurodynamics, 17*(2), 345–355. https://doi.org/10.1007/s11571-022-09828-9

Khan, K., & Katarya, R. (2023). Machine Learning Techniques for Autism Spectrum Disorder: Current trends and future directions. 2023 International Conference on Innovative Trends in Information Technology, ICITIIT *2023*, *1*, 1–7. https://doi.org/10.1109/ICITIIT57246.2023.10068658.

Khodatars, M., Shoeibi, A., Sadeghi, D., Ghaasemi, N., Jafari, M., Moridian, P., Khadem, A., Alizadehsani, R., Zare, A., Kong, Y., Khosravi, A., Nahavandi, S., Hussain, S., Acharya, U. R., & Berk, M. (2021). Deep learning for neuroimaging-based diagnosis and rehabilitation of Autism Spectrum Disorder: A review. *Computers in Biology and Medicine, 139*. https://doi.org/10.1016/j.compbiomed.2021.104949

Khor, S. W. H., Md Sabri, A. Q., & Othmani, A. (2023). Autism classification and monitoring from predicted categorical and dimensional emotions of video features. *Signal, Image and Video Processing*. https://doi.org/10.1007/s11760-023-02699-5

Kwon, H., Kim, J. I., Son, S. Y., Jang, Y. H., Kim, B. N., Lee, H. J., & Lee, J. M. (2022). Sparse hierarchical representation learning on functional brain networks for prediction of autism severity levels. *Frontiers in Neuroscience, 16*(July). https://doi.org/10.3389/fnins.2022.935431

Landowska, A., Karpus, A., Zawadzka, T., Robins, B., Barkana, D. E., Kose, H., Zorcec, T., & Cummins, N. (2022). Automatic emotion recognition in children with autism: A systematic literature review. *Sensors, 22*(4), 1–29. https://doi.org/10.3390/s22041649

Lasantha, D., Vidanagamachchi, S., & Nallaperuma, S. (2024). CRIECNN: Ensemble convolutional neural network and advanced feature extraction methods for the precise forecasting of circRNA-RBP binding sites. *Computers in Biology and Medicine, 174*(March), Article 108466. https://doi.org/10.1016/j.compbiomed.2024.108466

Le, H. D., Lee, G. S., Kim, S. H., Kim, S., & Yang, H. J. (2023). Multi-label multimodal emotion recognition with transformer-based fusion and emotion-level representation learning. *IEEE Access, 11*(February), 14742–14751. https://doi.org/10.1109/ACCESS.2023.3244390

Li, J., Chheang, V., Kullu, P., Brignac, E., Guo, Z., Bhat, A., Barner, K.E., & Barmaki, R.L. (2023). MMASD: A Multimodal Dataset for Autism Intervention Analysis. ACM International Conference Proceeding Series, 2, 397–405. https://doi.org/10.1145/3577190.3614117.

Li, N., Xiao, J., Mao, N., Cheng, D., Chen, X., Zhao, F., & Shi, Z. (2024). Joint learning of multi-level dynamic brain networks for autism spectrum disorder diagnosis. *Computers in Biology and Medicine, 171*(January). https://doi.org/10.1016/j.compbiomed.2024.108054

Li, Z., Li, H., Ralescu, A.L., Dillman, J.R., Altaye, M., Cecil, K.M., Parikh, N.A., & He, L. (2023). *Joint Self-Supervised and Supervised Contrastive Learning for Multimodal MRI Data: Towards Predicting Abnormal Neurodevelopment*. 157(August). http://arxiv.org/abs/2312.15064.

Lin, Y. S., Gau, S. S. F., & Lee, C. C. (2020). A Multimodal Interlocutor-Modulated Attentional BLSTM for Classifying Autism Subgroups during Clinical Interviews. *IEEE Journal on Selected Topics in Signal Processing, 14*(2), 299–311. https://doi.org/10.1109/JSTSP.2020.2970578

Loganathan, S., Geetha, C., Nazaren, A. R., & Harin Fernandez Fernandez, M. (2023). Autism spectrum disorder detection and classification using chaotic optimization based Bi-GRU network: An weighted average ensemble model. *Expert Systems with Applications, 230*(June), Article 120613. https://doi.org/10.1016/j.eswa.2023.120613

Luo, N., Zhong, X., Su, L., Cheng, Z., Ma, W., & Hao, P. (2023). Artificial intelligence-assisted dermatology diagnosis: From unimodal to multimodal. *Computers in Biology and Medicine, 165*(July), Article 107413. https://doi.org/10.1016/j.compbiomed.2023.107413

Lyra, L. O., Fabris, A. E., & Florindo, J. B. (2024). A multilevel pooling scheme in convolutional neural networks for texture image recognition. *Applied Soft Computing, 152*(December 2023). https://doi.org/10.1016/j.asoc.2024.111282

Min, L., Fan, Z., Dou, F., Sun, J., Luo, C., & Lv, Q. (2024). Adaption BERT for Medical Information Processing with ChatGPT and Contrastive Learning. *Electronics, 13*(13). https://doi.org/10.3390/electronics13132431

Mishra, M., & Pati, U. C. (2023). A classification framework for Autism Spectrum Disorder detection using sMRI: Optimizer based ensemble of deep convolution neural network with on-the-fly data augmentation. *Biomedical Signal Processing and Control, 84*(February), Article 104686. https://doi.org/10.1016/j.bspc.2023.104686

Moon, J. H., Lee, H., Shin, W., Kim, Y. H., & Choi, E. (2022). Multi-modal understanding and generation for medical images and text via vision-language pre-training. *IEEE Journal of Biomedical and Health Informatics, 26*(12), 6070–6080. https://doi.org/10.1109/JBHI.2022.3207502

Muizelaar, H., Haas, M., van Dortmont, K., van der Putten, P., & Spruit, M. (2024). Extracting patient lifestyle characteristics from Dutch clinical text with BERT models. *BMC Medical Informatics and Decision Making, 24*(1), 1–15. https://doi.org/10.1186/s12911-024-02557-5

Nogay, H. S., & Adeli, H. (2024). Multiple classification of brain MRI autism spectrum disorder by age and gender using deep learning. *Journal of Medical Systems, 48*(1). https://doi.org/10.1007/s10916-023-02032-0

Özbay, E., & Altunbey Özbay, F. (2023). Interpretable features fusion with precision MRI images deep hashing for brain tumor detection. *Computer Methods and Programs in Biomedicine, 231*. https://doi.org/10.1016/j.cmpb.2023.107387

Park, K. W., & Cho, S. B. (2023). A residual graph convolutional network with spatio-temporal features for autism classification from fMRI brain images. *Applied Soft Computing, 142*, Article 110363. https://doi.org/10.1016/j.asoc.2023.110363

Parlett-Pelleriti, C. M., Stevens, E., Dixon, D., & Linstead, E. J. (2022). Applications of unsupervised machine learning in autism spectrum disorder research: A review, 0123456789 *Review Journal of Autism and Developmental Disorders*. https://doi.org/10.1007/s40489-021-00299-y.

Parui, S., Samanta, D., Chakravorty, N., Ghosh, U., & Rodrigues, J. J. P. C. (2023). Artificial intelligence and sensor-based autism spectrum disorder diagnosis using brain connectivity analysis ☆, ☆☆. *Computers and Electrical Engineering, 108*(April), Article 108720. https://doi.org/10.1016/j.compeleceng.2023.108720

Passos, L. A., Papa, J. P., Del Ser, J., Hussain, A., & Adeel, A. (2023). Multimodal audio-visual information fusion using canonical-correlated Graph Neural Network for energy-efficient speech enhancement. *Information Fusion, 90*(September 2022), 1–11. https://doi.org/10.1016/j.inffus.2022.09.006

Qayyum, A., Razzak, I., Tanveer, M., & Mazher, M. (2022). Spontaneous facial behavior analysis using deep transformer based framework for child–computer interaction. *ACM Transactions on Multimedia Computing, Communications, and Applications*. https://doi.org/10.1145/3539577

Rakhimberdina, Z., Liu, X., & Murata, T. (2020). Population graph-based multi-model ensemble method for diagnosing autism spectrum disorder. *Sensors (Switzerland), 20*(21), 1–18. https://doi.org/10.3390/s20216001

Rakić, M., Cabezas, M., Kushibar, K., Oliver, A., & Lladó, X. (2020). Improving the detection of autism spectrum disorder by combining structural and functional MRI information. *NeuroImage: Clinical, 25*(January), Article 102181. https://doi.org/10.1016/j.nicl.2020.102181

Sheng, J., Zhang, Q., Zhang, Q., Wang, L., Yang, Z., Xin, Y., & Wang, B. (2024). A hybrid multimodal machine learning model for Detecting Alzheimer's disease. *Computers in Biology and Medicine, 170*(February), Article 108035. https://doi.org/10.1016/j.compbiomed.2024.108035

Shoeibi, A., Khodatars, M., Jafari, M., Moridian, P., Rezaei, M., Alizadehsani, R., Khozeimeh, F., Gorriz, J. M., Heras, J., Panahiazar, M., Nahavandi, S., & Acharya, U. R. (2021). Applications of deep learning techniques for automated multiple sclerosis detection using magnetic resonance imaging: A review. *Computers in Biology and Medicine, 136*(July). https://doi.org/10.1016/j.compbiomed.2021.104697

Song, C., Wang, S., Chen, M., Li, H., Jia, F., & Zhao, Y. (2023). A multimodal discrimination method for the response to name behavior of autistic children based on human pose tracking and head pose estimation. *Displays, 76*(ember 2022), Article 102360. https://doi.org/10.1016/j.displa.2022.102360

Sun, H., Lian, Z., Sun, L., Liu, B., & Tao, J. (2023). *RMNAS: A Multimodal Neural Architecture Search Framework For Robust Multimodal Sentiment Analysis*. ⟨http://arxiv.org/abs/2312.15583⟩.

Tan, Z., Madzin, H., Norafida, B., Rahmat, R. W. O., Khalid, F., & Sulaiman, P. S. (2024). SwinUNeLCsT: Global–local spatial representation learning with hybrid CNN–transformer for efficient tuberculosis lung cavity weakly supervised semantic segmentation. *Journal of King Saudade University - Computer and Information Sciences, 36*(4), Article 102012. https://doi.org/10.1016/j.jksuci.2024.102012

Tang, M., Kumar, P., Chen, H., & Shrivastava, A. (2020). Deep multimodal learning for the diagnosis of autism spectrum disorder. *Journal of Imaging, 6*(6). https://doi.org/10.3390/jimaging6060047

Tang, Y., Tong, G., Xiong, X., Zhang, C., Zhang, H., & Yang, Y. (2023). Multi-site diagnostic classification of Autism spectrum disorder using adversarial deep learning on resting-state fMRI. *Biomedical Signal Processing and Control, 85*(February), Article 104892. https://doi.org/10.1016/j.bspc.2023.104892

Timms, S., Lodhi, S., Bruce, J., & Stapleton, E. (2022). Auditory symptoms and autistic spectrum disorder: A scoping review and recommendations for future research (xxxx) *Journal of Otology*. https://doi.org/10.1016/j.joto.2022.08.004.

Tseng, Y. C., Kuo, C. W., Peng, W. C., & Hung, C. C. (2024). al-BERT: a semi-supervised denoising technique for disease prediction. *BMC Medical Informatics and Decision Making, 24*(1), 1–19. https://doi.org/10.1186/s12911-024-02528-w

Vasanthakumari, R. K., Nair, R. V., & Krishnappa, V. G. (2023). Improved learning by using a modified activation function of a Convolutional Neural Network in multi-spectral image classification. *Machine Learning with Applications, 14*(October), Article 100502. https://doi.org/10.1016/j.mlwa.2023.100502

Waizbard-Bartov, E., Fein, D., Lord, C., & Amaral, D. G. (2023). Autism severity and its relationship to disability. *Autism Research, 16*(4), 685–696. https://doi.org/10.1002/aur.2898

Wan, G., Deng, F., Jiang, Z., Song, S., Hu, D., Chen, L., Wang, H., Li, M., Chen, G., Yan, T., Su, J., & Zhang, J. (2022). FECTS: A facial emotion cognition and training system for chinese children with autism spectrum disorder. *Computational Intelligence and Neuroscience, 2022*. https://doi.org/10.1155/2022/9213526

Wang, Miaoyan, Xu, D., Zhang, L., & Jiang, H. (2023). Application of multimodal MRI in the early diagnosis of autism spectrum disorders: A review. *Diagnostics, 13*(19), 1–19. https://doi.org/10.3390/diagnostics13193027

Wang, Mingzhi, Guo, J., Wang, Y., Yu, M., & Guo, J. (2023). Multimodal autism spectrum disorder diagnosis method based on DeepGCN. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 31*, 3664–3674. https://doi.org/10.1109/TNSRE.2023.3314516

Wawer, A., Chojnicka, I., Okruszek, L., & Sarzynska-Wawer, J. (2022). Single and cross-disorder detection for autism and schizophrenia. *Cognitive Computation, 14*(1), 461–473. https://doi.org/10.1007/s12559-021-09834-9

Yang, X., Zhang, N., & Schrader, P. (2022). Machine Learning with Applications A study of brain networks for autism spectrum disorder classification using resting-state functional connectivity. *Machine Learning with Applications, 8*(March), Article 100290. https://doi.org/10.1016/j.mlwa.2022.100290

Yu, Q., Ma, Q., Da, L., Li, J., Wang, M., Xu, A., Li, Z., & Li, W. (2024). A transformer-based unified multimodal framework for Alzheimer's disease assessment. *Computers in Biology and Medicine, 180*(August). https://doi.org/10.1016/j.compbiomed.2024.108979

Zhang, Meiwei, Cui, Q., Lü, Y., Yu, W., & Li, W. (2024). A multimodal learning machine framework for Alzheimer's disease diagnosis based on neuropsychological and neuroimaging data. *Computers and Industrial Engineering, 197*(6), Article 110625. https://doi.org/10.1016/j.cie.2024.110625

Zhang, Mengyi, Sun, L., Kong, Z., Zhu, W., Yi, Y., & Yan, F. (2024). Pyramid-attentive GAN for multimodal brain image complementation in Alzheimer's disease classification. *Biomedical Signal Processing and Control, 89*(October 2023). https://doi.org/10.1016/j.bspc.2023.105652