RESEARCH

Open Access

Autism spectrum disorders detection based on multi-task transformer neural network

Le Gao[1,2], Zhimin Wang[2], Yun Long[3*], Xin Zhang[2], Hexing Su[2], Yong Yu[4] and Jin Hong[5*]

Abstract

Autism Spectrum Disorders (ASD) are neurodevelopmental disorders that cause people difficulties in social interaction and communication. Identifying ASD patients based on resting-state functional magnetic resonance imaging (rs-fMRI) data is a promising diagnostic tool, but challenging due to the complex and unclear etiology of autism. And it is difficult to effectively identify ASD patients with a single data source (single task). Therefore, to address this challenge, we propose a novel multi-task learning framework for ASD identification based on rs-fMRI data, which can leverage useful information from multiple related tasks to improve the generalization performance of the model. Meanwhile, we adopt an attention mechanism to extract ASD-related features from each rs-fMRI dataset, which can enhance the feature representation and interpretability of the model. The results show that our method outperforms state-of-the-art methods in terms of accuracy, sensitivity and specificity. This work provides a new perspective and solution for ASD identification based on rs-fMRI data using multi-task learning. It also demonstrates the potential and value of machine learning for advancing neuroscience research and clinical practice.

Keywords Autism Spectrum Disorders, Artificial intelligence, Biological information, Multi-task learning, Transformer network

Introduction

ASD (Autism Spectrum Disorders) is a heterogeneous condition that affects communication, behavior, and social interactions in various ways and degrees [1]. According to the latest Diagnostic and Statistical Manual of Mental Disorders (DSM-5), ASD encompasses a spectrum of disorders that were previously diagnosed separately, such as autism, Asperger's syndrome, and other pervasive developmental disorders. The global prevalence of ASD has increased dramatically over the years, reaching 1 in 59 children in the United States in 2014 [2]. ASD poses a major public health challenge, as it impacts not only the individuals with ASD, but also their families and society [3]. Early diagnosis and intervention are crucial for improving the outcomes and reducing the costs of ASD [4], but the current standard diagnosis relies on subjective and time-consuming assessments by multidisciplinary teams using standardized tools [5]. These assessments require highly specialized knowledge and experience from the evaluators, and are often inaccessible or unavailable to many patients [6]. Therefore, there is an urgent need for objective and efficient diagnostic methods based on biological markers.

With the rapid development of AI (Artificial Intelligence) technology, machine learning as a subfield of AI, it has largely enhanced the role of computational methods

*Correspondence:
Yun Long
Long.Y1990@outlook.com
Jin Hong
hongjin280@gmail.com
[1] School of Computer Engineering, Guangzhou Huali College, Guangzhou 511325, China
[2] Faculty of Intelligent Manufacturing, Wuyi University, Jiangmen 529000, China
[3] State Key Laboratory of Public Big Data, Guizhou University, Guizhou 550025, China
[4] School of Computer Science, Shaanxi Normal University, Xi'an 710062, China
[5] School of Information Engineering, Nanchang University, Nanchang, 330031, China

in neuroscience [7]. Machine learning has been successfully applied in Alzheimer's disease, mild cognitive impairment [8, 9], temporal lobe epilepsy, schizophrenia, Parkinson's [10], dementia [11, 12], ADHD [13, 14], ASD [15, 16] and major depressive disorder [17]. In particular, the identification of ASD has made great progress and a series of effective methods have been developed [18]. These methods can be briefly divided into two categories as follows: (1) Based on traditional machine learning methods, it models ASD data as a binary classification problem using traditional machine learning techniques. Crippa et al. [19] used support vector machine (SVM) algorithm to segment ASD patient samples and normal controls (NC) samples by fitting a hyperplane. Rane et al. [20] used logistic regression method to predict ASD diagnosis by transforming fMRI data into probabilities of specific binary values through linear operations. Abbas et al. [21] used an integrated learning approach to construct an ASD screening tool by combining a parent questionnaire-based classifier and a behavioral video-based classifier; (2) Based on the deep learning approach, it uses deep neural networks to extract hidden features in ASD data for ASD identification. Heinsfeld et al. [22] used autoencoders to downscale rs-fMRI data and then used deep neural networks for ASD prediction. Alsaade et al. [23] performed prediction of ASD disease by constructing a functional brain connectivity matrix and projecting it to a deep feature space. Pavăl [24] used convolutional neural networks for facial abnormality identification in ASD patients.

Despite the success of these methods, the identification of ASD remains a challenge due to the complex causes of autism formation and unclear pathogenesis [25, 26]. Moreover, most existing methods are based on single-task learning, which ignores the potential correlations and complementarities among different ASD recognition tasks [27, 28]. To address these issues, we propose a novel multi-task learning framework for ASD identification based on resting-state functional magnetic resonance imaging (rs-fMRI) data. Figure 1 is the multitasking transformer framework diagram. Rs-fMRI is increasingly used to study neural connectivity and identify biomarkers of psychiatric disorders. It performs imaging based on blood oxygen level-dependent (BOLD) signal changes in brain regions in a non-invasive manner. Thus rs-fMRI-based ASD identification can provide more accurate, stable and interpretable predictions.

The main contributions and novelties of our work are as follows:

This paper proposes a novel multi-task learning framework for ASD identification based on rs-fMRI data, which can leverage useful information from multiple related tasks to improve the generalization performance of the model. We introduce a temporal encoding module to encode the rs-fMRI data, which can capture the sequential information embedded in the temporal nodes. Meanwhile, we adopt an attention mechanism to extract ASD-related features from each rs-fMRI dataset, which can enhance the feature representation and interpretability of the model.

We design a feature sharing module to share the ASD features learned from each dataset, which can exploit the correlations and complementarities among different tasks.

We conduct extensive experiments on two public rs-fMRI datasets to evaluate the effectiveness of our proposed method. The results show that our method outperforms state-of-the-art methods in terms of accuracy, sensitivity and specificity. This work provides a

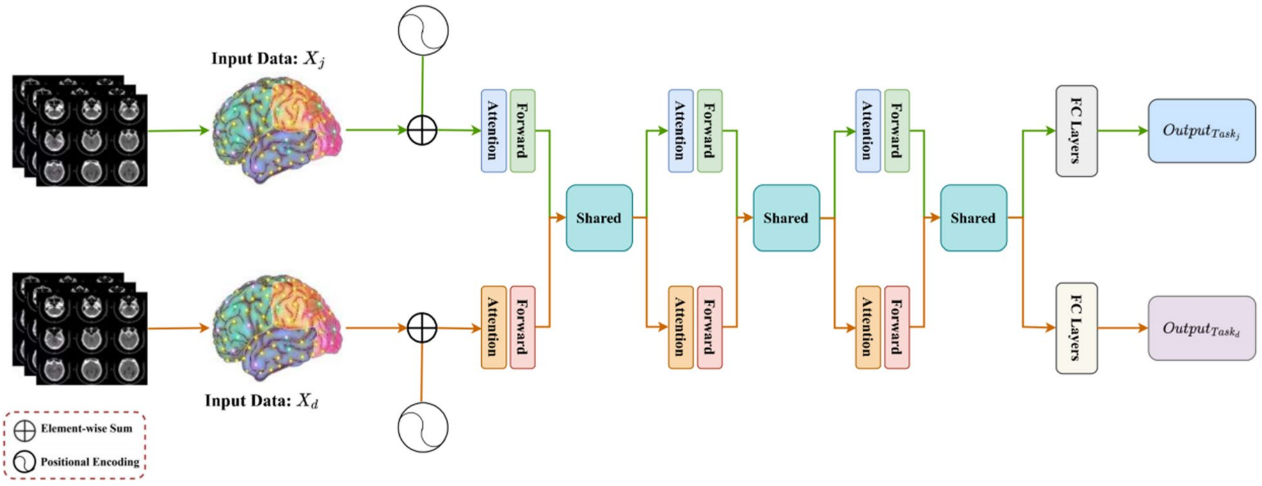

Fig. 1 Multitasking transformer framework diagram

new perspective and solution for ASD identification based on rs-fMRI data using multi-task learning. It also demonstrates the potential and value of machine learning for advancing neuroscience research and clinical practice.

Materials and methods

Materials

In the present study, we used rs-fMRI data from the Autism Imaging Data Exchange (ABIDE). Due to the limited number of subjects at the site, we selected 2 different sites (number of subjects > 100) from a large number of sites, including UM and NYU. Also, data and detailed information are available at https://fcon_1000.projects.nitrc.org/indi/abide/, where Table 1 shows the demographic information of the subjects aggregated.

Data preprocessing

There is no consensus on the best methods for preprocessing resting state fMRI data. Rather than being prescriptive and favoring a single processing strategy, we have preprocessed the data using Connectome Computation System (CCS), Configurable Pipeline for the Analysis of Connectomes (CPAC), Data Processing Assistant for Resting-State fMRI (DPARSF), Neuroimaging Analysis Kit (NIAK), each of which was implemented using the chosen parameters and settings of the pipeline developers.

The preprocessing steps implemented by the different pipelines are very similar. The largest changes are for the specific algorithms used for each step, their software implementations, and the parameters used. The following sections outline the different preprocessing steps and their differences in the pipeline.

Basic processing

Step	CCS	C-PAC	DPARSF	NIAK
Drop first "N" volumes	4	0	4	0
Slice timing correction	Yes	Yes	Yes	No
Motion realignment	Yes	Yes	Yes	Yes
Intensity normalization	4D Global mean = 1000	4D Global mean = 1000	No	Non-uniformity correction using median volume

Table 1 Demographic information of subjects

Site	ASD		NC		Total
	Age Avg(SD)	Count	Age Avg(SD)	Count	
NYU	14.76(7.12)	M 64, F 10	15.75(6.18)	M 72, F 26	172
UM	13.71(2.37)	M 38, F 9	14.84(3.62)	M 55, F 18	120

Nuisance signal removal

Each pipeline implemented some form of nuisance variable regression to clean confounding variation due to physiological processes (heart beat and respiration), head motion, and low frequency scanner drifts, from the fMRI signal.

Regressor	CCS	C-PAC	DPARSF	NIAK
Motion	24-param	24-param	24-param	Scrubbing and 1st principal component of 6 motion parameters and their squares
Tissue signals	Mean WM and CSF signals	CompCor (5 PCs)	Mean WM and CSF signals	Mean WM and CSF signals
Motion realignment	Yes	Yes	Yes	Yes
Low-frequency drifts	Linear and quadratic trends	Linear and quadratic trends	Linear and quadratic trends	Discrete cosine basis with a 0.01 Hz high-pass cut-off

Processing strategies

Each pipeline was used to calculate four different preprocessing strategies:

Strategy	Band-pass filtering	Global signal regression
filt_global	Yes	Yes
filt_noglobal	Yes	No
nofilt_global	No	Yes
nofilt_noglobal	No	No

For strategies that include global signal correction, the global mean signal was included with nuisance variable regression. Band-pass filtering (0.01–0.1 Hz) was applied after nuisance variable regression.

Registration

A transform from original to template (MNI152) space was calculated for each dataset from a combination of functional-to-anatomical and anatomical-to-template transforms. The anatomical-to-template transforms were calculated using a two step procedure that involves (one or more) linear transform that is later refined with a very high dimensional non-linear transform. When data are written into template space (typically after the calculation of derivatives, except for NIAK) all transforms are used simultaneously to avoid multiple interpolations.

Registration	CCS	C-PAC	DPARSF	NIAK
Functional to Anatomical	Boundary-based rigid body (BBR)	Boundary-based rigid body (BBR)	Rigid body	Rigid body
Anatomical to Standard	FLIRT + FNIRT	ANTs	DARTEL	CIVET

Methods

In this section, we design the multitask Transformer framework to improve the ASD prediction performance by sharing the knowledge learned from multiple tasks. Specifically, Section“Problem definition” formally defines the problem. Section“Location coding” describes how to encode positions according to the order of time nodes. Section“Attention module” defines the way the attention mechanism in the Transformer captures useful features. Section“Feature sharing” describes the process of feature sharing among different tasks. Section“Objective function” defines the objective function for Optimization of the objective function.

Problem definition

In this section, we describe the proposed multi-task Transformer learning framework. Suppose we have D tasks and the rs-fmri dataset as follows (1), An instance as follows (2) in X_d contains T time nodes and N brain regions, and the corresponding label $y_d \in \{0, 1\}$ is a binary classification task, In the experiment, label 1 represents illness, label 0 represents no disease, and the label and label of both tasks have the same significance. We further assume that there are D different Transformer networks, and each Transformer network consists of L-layer feedforward networks, where the lth layer network extracts the features of task d through $f_d^l \in \mathbb{R}^{T \times N}$. Specifically, our goal is to improve the generalization performance of task d by sharing features learned from other tasks as follows (3).

$$D = \{X_d, Y_d\}_{d=1}^{D} \tag{1}$$

$$\{T_d(\cdot)\}_{d=1}^{D} \tag{2}$$

$$\{f_j^l\}_{l=1}^{L}, \forall j \neq d \tag{3}$$

Location coding

Temporal order information in time series data helps to improve model prediction accuracy [29]. To take full advantage of the sequential information embedded in the time nodes in the rs-fMRI data, we inject information about the position in the time node sequence for each input data. Specifically, we obtain a position-encoded PE with the same dimensionality as x_d using the sine and cosine function, which is calculated as follows (4, 5).

$$PE_{(t,2n)} = \sin(t/10000^{2n/N}) \tag{4}$$

$$PE_{(t,2n+1)} = \cos(t/10000^{2n/N} \tag{5}$$

where t denotes the position of the time node in T time nodes, 2n denotes the brain region of even number, and $2n + 1$ denotes the brain region of base number. In our model, x_d represents the input features of the task d. It is a two-dimensional matrix, where each row corresponds to a time node and each column corresponds to a brain region. Thus, the dimension of x_d is (T, R), where T is the number of time nodes and R is the number of brain regions. Then, the location information embedding, which we implement by summing the location encoding PE and x_d, is calculated as follows (6).

$$\widetilde{x}_d = PE + x_d \tag{6}$$

Attention module

The Transformer network consists of several attention modules to improve the ASD prediction performance, as shown in Fig. 2. In our model, Q, K, and V represent the query (Query), the Key (Key), and the value (Value), respectively. These manipulations are central parts of the attention mechanism and are used to compute correlations between input features. f_d^l represents the l-layer feature of the task d. These features are extracted through feed-forward networks and attention modules and can capture important information about the input data. Our goal is to improve the generalization performance of task d by sharing features learned from other tasks.

Specifically, the features of f_d^l are first extracted by three different linear operations of Q, K and V, and the number of channels is reduced to half of the original one to reduce the computational effort, as follows (7–9).

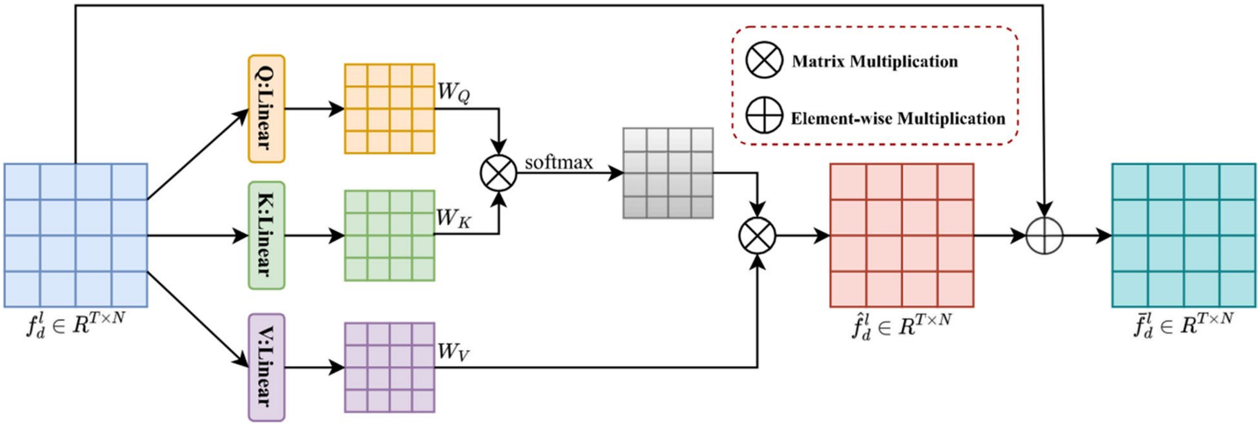
Q:Linear
K:Linear
V:Linear
W_Q
W_K
W_V
softmax
Matrix Multiplication
Element-wise Multiplication
$f_d^l \in R^{T\times N}$
$\hat{f}_d^l \in R^{T\times N}$
$\bar{f}_d^l \in R^{T\times N}$

Fig. 2 Attention module

$$W_Q = Q(f_d^l) \tag{7}$$

$$W_K = K(f_d^l) \tag{8}$$

$$W_V = V(f_d^l) \tag{9}$$

where $W_Q \in \mathbb{R}^{T\times N}$, $W_K \in \mathbb{R}^{T\times N}$ and $W_V \in \mathbb{R}^{T\times N}$ denote the output feature vectors. Then, we matrix multiply W_Q and W_K to calculate the correlation weights between time nodes and score them using softmax operation. Finally, we weight sum the correlation weights and W_V to obtain the attention feature vector, which is calculated as follows (10).

$$\widehat{f}_d^l = \mathrm{softmax}(W_Q W_K{}^T) W_V \tag{10}$$

where $\bar{f}_d^l \in \mathbb{R}^{T\times N}$ denotes the attention feature vector. Finally, we fuse the attentional feature vector $\widehat{f}_d^l$ and the feature f_d^l, which aims to compensate for the information lost when the attentional mechanism captures the features, calculated as follows (11).

$$\bar{f}_d^l = \widehat{f}_d^l + f_d^l \tag{11}$$

where $\bar{f}_d^l \in \mathbb{R}^{T\times N}$ denotes the output fused features. In addition, each layer of the feedforward network consists of an attention module and a Forward network, and the fused features are transformed into the task-specific feature space by a fully connected Forward network, calculated as follows (12).

$$\widetilde{f}_d^l = \mathrm{Relu}(W_f \bar{f}_d^l + b_f) \tag{12}$$

where $\widetilde{f}_d^l \in \mathbb{R}^{T\times N}$ denotes the output, Relu(•) denotes the activation function, and W_f and b_f denote the corresponding parameters.

Feature sharing

To realize the interaction of features between tasks, we build a feature sharing module, as shown in Fig. 3. Each layer of the network defines D learnable activation mappings $M_D = \{M_d\}_{d=1}^D$, where $M_d = \{M_{1d}, ..., M_{Dd}\}$. We use M_D to linearly combine the feature vectors of different task networks and use them as inputs for the next layer of feedforward networks. Specifically, we matrix the activation mapping $M_D = \begin{vmatrix} M_{11} & ... & M_{D1} \\ ... & M_{dd} & ... \\ M_{1D} & ... & M_{DD} \end{vmatrix}$ and use it to linearly combine multiple feature vectors, which are computed as follows (13).

$$\begin{vmatrix} f_1^{l+1} \\ \vdots \\ f_d^{l+1} \\ \vdots \\ f_D^{l+1} \end{vmatrix} = \begin{vmatrix} M_{11} & ... & M_{D1} \\ ... & M_{dd} & ... \\ M_{1D} & ... & M_{DD} \end{vmatrix} \begin{vmatrix} \widetilde{f}_1^l \\ \vdots \\ \widetilde{f}_d^l \\ \vdots \\ \widetilde{f}_D^l \end{vmatrix} \tag{13}$$

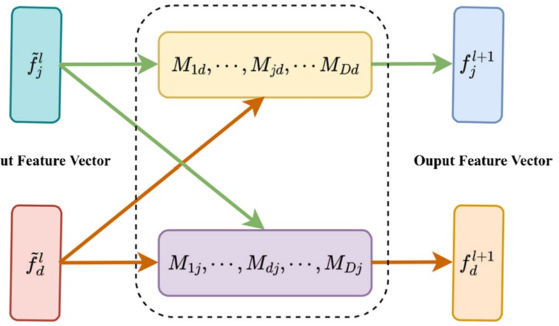

Fig. 3 Feature sharing module, in this graph, M represents the activation map for linearly combining the feature vectors of different task networks

where f_d^{l+1} denotes the output of the l+1-layer network. We can identify specific layer tasks by setting $M_{i,k}$, $i < D, k < D$ to zero, or share more features by assigning them higher values.

Objective function

Feedforward networks do not change the dimensionality of the feature vectors; however, high-dimensional and high-noise data have a negative impact on the prediction performance. To solve this problem, we reduce the dimensionality of the feature vector $f_d^L \in \mathbb{R}^{T\times N}$ by FC Layers and perform the prediction. FC Layers consists of three layers of fully connected operations, the first two layers are used to reduce the dimensionality and the last layer gets the prediction output, which is calculated as follows (14):

$$\hat{y}_d = \mathrm{softmax}(W_3(\mathrm{Relu}(W_2(\mathrm{Relu}(W_1 f_d^L + b_1)) + b_2)) + b_3) \tag{14}$$

where $\hat{y}_d$ denotes the output, and $W_{i=1,2,3}$ and $b_{i=,1,2,3}$ denote the corresponding parameters. Then, we use the binary cross-entropy as the loss and the objective function is calculated as follows (15).

$$\mathcal{L} = \sum_{d=1}^{D}\left[\sum\nolimits_{\substack{x_d\in X_d\\ y_d\in Y_d}} -y_d\log\hat{y}_d\right] \tag{15}$$

Results and discussion

In this section, we conduct extensive experiments to verify the effectiveness of our approach. Specifically, Section"Experimental setup" describes the experimental setting and setup. Section"Evaluation metrics" gives the evaluation metrics to evaluate the experimental results. Section"Experimental results and discussion" presents the comparison of our method with the current popular methods on two ASD datasets and the analysis of the experimental results.

Experimental setup

The experiments are programmed and implemented as follows: PyTorch 1.9, Python 3.8, using a GeForce RTX 3090 GPU for training. With grid search method for tuning hyperparameters, we use Adam as the training optimizer with 120 iterations, an initial learning rate of 1×10^{-5}, 50% decay every 30 iterations, and a Batch size of 16. The number of feedforward network layers L is 5, and the three fully-connected layers in FC Layers have output dimensions of 4096, 2048, and 2. In addition, we divide the ASD data in Section"Materials and methods" randomly into a training set and a test set in the ratio of 8:2 ratio randomly into training set and test set for subsequent experiments.

Evaluation metrics

We used Accuracy, Sensitivity and Specificity as metrics to evaluate the ASD identification results. All methods are tested using these metrics and calculated as follows (16–18):

$$\mathrm{Accuracy} = \frac{\mathrm{TruePostive} + \mathrm{TrueNegative}}{\mathrm{TruePostive} + \mathrm{FalsePostive} + \mathrm{TrueNegative} + \mathrm{FalseNegative}} \tag{16}$$

$$\mathrm{Sensitivity} = \frac{\mathrm{TruePostive}}{\mathrm{TruePostive} + \mathrm{FalseNegative}} \tag{17}$$

$$\mathrm{Specificity} = \frac{\mathrm{TrueNegative}}{\mathrm{TrueNegative} + \mathrm{FalsePostive}} \tag{18}$$

where True Positive indicates the number of ASD-positive patients correctly classified, True Negative indicates the number of ASD-false-negative patients, False Positive indicates the number of ASD-false-positive patients, and False Negative indicates the number of ASD-negative patients correctly classified.

Experimental results and discussion

Effects of loss function

This experiment established two datasets. Figure 4 shows the loss plot lines during the training of the experiments. For the Fig. 4a and b loss function, the loss value curve has fluctuated several times in a large range during the training process, which may indicate the occurrence of gradient explosion, resulting in excessive weight update of the model, thus causing instability of the model. Therefore, we need multiple training to improve the stability of training. From the graphs, the following conclusions can be drawn: (1) The experiments have converged for both datasets and the experimental results are reliable; (2) The experiments both have the fastest rate of decline until 90 iterations. This indicates that the model was able to effectively learn how to classify ASD patients and NC patients during this time; and (3) The experiments both reached convergence at 105–120 iterations, and the model was able to fit the training data. In summary, our model can fit the ASD dataset well and the experimental results are reliable and valid.

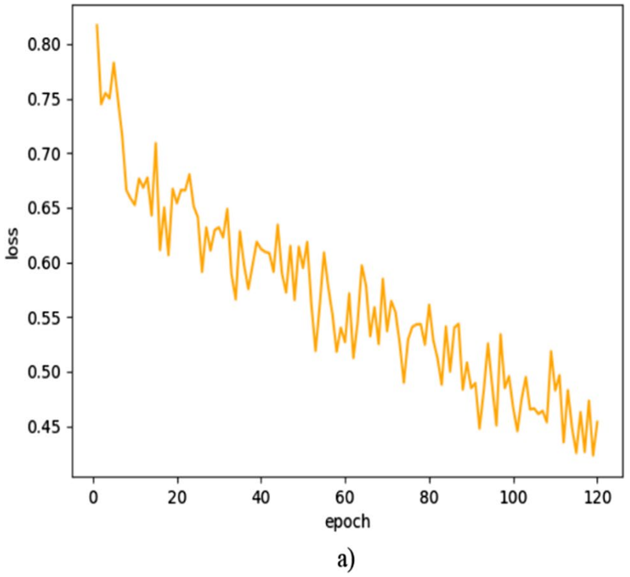

Fig. 4 Training loss plots. **a** NYU corresponding loss plot. **b** NYU corresponding loss plot

Confusion matrix

Figure 5 is a confusion matrix showing the number of true negative (TN), false positive (FP), false negative (FN) and true positive (TP) samples. From the figure, we can observe that (1) the TN value is the largest among the four values, i.e., the number of correctly predicted NC samples is the largest. Meanwhile, FP is the smallest, i.e., the number of incorrectly predicted NC samples is the least. This again validates that our method has a low misdiagnosis rate; (2) TP indicates the number of samples that correctly identified ASD patients. The difference between the TP and FN values is not significant. The reason for this result is that the number of NC samples in the training sample is high, which leads to category imbalance and thus affects the ability of the model to identify ASD patients; and (3) In the confusion matrix corresponding to the two datasets, the proportions of TN, FP, FN and TP are similar, which proves that

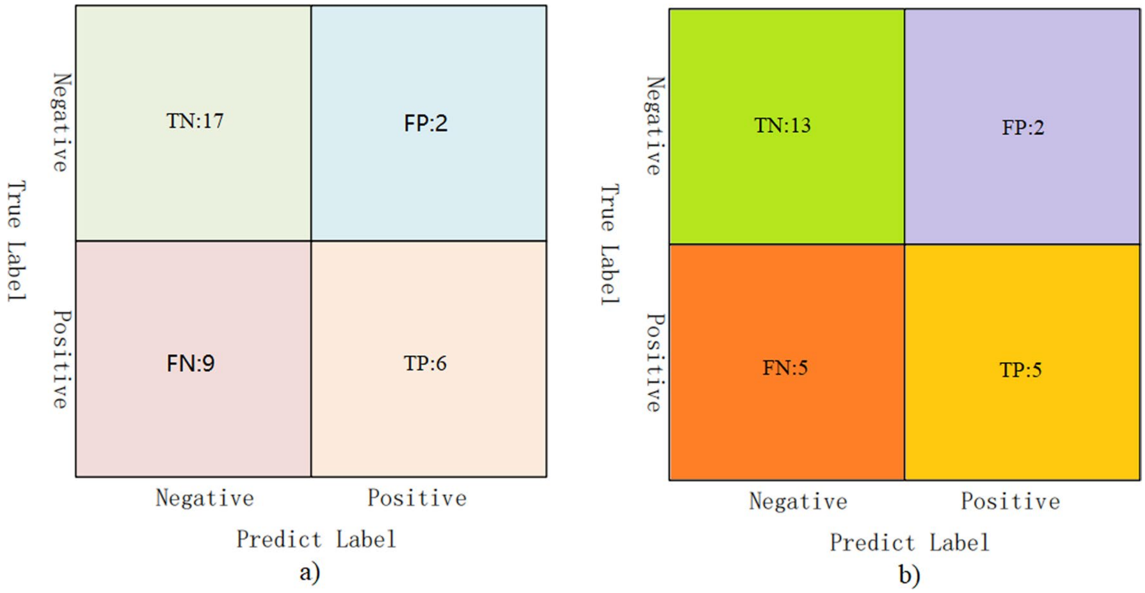

Fig. 5 Confusion matrix. **a** Confusion matrix corresponding to NYU dataset. **b** Confusion matrix corresponding to UM dataset

the model has some generalizability. In summary, our method can identify NC patients well and has some ability to identify ASD patients.

Ablation studies

As shown in Fig. 1, Multitasking transformer framework diagram can be regarded as a federated network composed of multiple features share modules. In this section, we conduct ablation studies to verify the effectiveness of the crucial components in multi-task learning framework and evaluate the impact of each single task network on the results. The transformer network consists of several attention modules to improve the ASD prediction performance. Based on the transformer network, we built a single network and a feature sharing module respectively. All experiments are performed with the same hyperparameter configuration. Table 2 shows the ablation studies with different network configurations.

From index in Table 2, we can see that when we simply add a single task network to the transformer network, Accuracy and Specificity all suffer a decline, but the sensitivity suffers a rise. This shows that adding the multitask will bring better results. When we applied Single task network and Multitask network to the NYU dataset, the accuracy, sensitivity, and specificity indicators of Single task network were 63.15%, 52.63%, and 73.68%, respectively. The accuracy, sensitivity, and specificity indicators of Multitask network were 67.64%, 40.00%, and 89.47%, respectively. When we applied Single task network and Multitask network to the UM dataset, the accuracy, sensitivity, and specificity indicators of Single task network were 70.68%, 66.00%, and 73.68%, respectively, while the accuracy, sensitivity, and specificity indicators of Multitask network were 72.00%, 55.00%, and 86.66%, respectively. In summary, adding the feature sharing module to the transformer network has the best recognition and prediction performance for rs-fMRI data, indicating the necessity of the feature sharing module in deep learning networks.

Comparison with the state-of-the-art methods

In this section, we compare the proposed method with some popular machine learning and deep learning methods, including support vector machines [30], random forests [31], multilayer perceptron [32], SAENet [33], MLwSGSU [34] and MCNNet [35]. To test the results of these methods, we used their public codes on the NYU and UM datasets for training and evaluation. The experimental results of the seven models on the two datasets are shown in Figs. 6 and 7. The figures show that compared with other methods, the accuracy values obtained by us have better results. Figures 6 and 7 show that compared with the suboptimal method, the accuracy values obtained by us have increased by 4.54% and 5.88% on the two data sets respectively.

Table 3 shows the results of the different methods on the NYU and UM datasets. On the NYU dataset, our proposed model achieves 67.64%, 40% and 89.47% in accuracy, sensitivity and specificity respectively, which are the best results among all compared methods. Unlike the case on the NYU dataset, on the UM dataset, our proposed model achieves 72%, 50% and 86.66% in accuracy, sensitivity and specificity respectively. Through qualitative comparisons on the two datasets, we find that both our model can guarantee the improvement of comprehensive performance and maintain a high specificity without introducing too many false positives. Therefore, compared to other methods, we believe that Multitasking Transformer framework can better cope with the ASD prediction.

To sum up, (1) Multi-task learning methods are competitive with traditional machine learning methods and deep learning methods in ASD recognition; (2) Our methods are significantly better than other methods in both accuracy and specificity; and (3) Our methods are not as sensitive as other methods. We hypothesize that since the number of NC patients in the dataset is slightly more than that of ASD patients, the attention mechanism when training the model is more biased towards learning to capture NC features, thus negatively influencing the extraction of ASD features, and therefore less sensitive. In addition, the method is effective for ASD identification. In conclusion, our method can better identify ASD patients with a lower probability of misdiagnosis of NC patients.

Table 2 Ablation studies with different network configurations

Site	Method	Accuracy (%)	Sensitivity (%)	Specificity (%)
NYU	Single task	63.15	52.63	73.68
	Ours	67.64	40.00	89.47
UM	Single task	70.68	66.00	73.68
	Ours	72.00	50.00	86.66

Limitations

Although our method performs very well compared to other methods, several limitations exist. Although our method has higher accuracy and specificity, there is still lower sensitivity. And the accuracies are only 67.64% and 72%. This is attributed to the amount of training data being too small, which leads too poor generalization on the rs-fMRI data. We plan to explore more effective data augmentation techniques in future work.

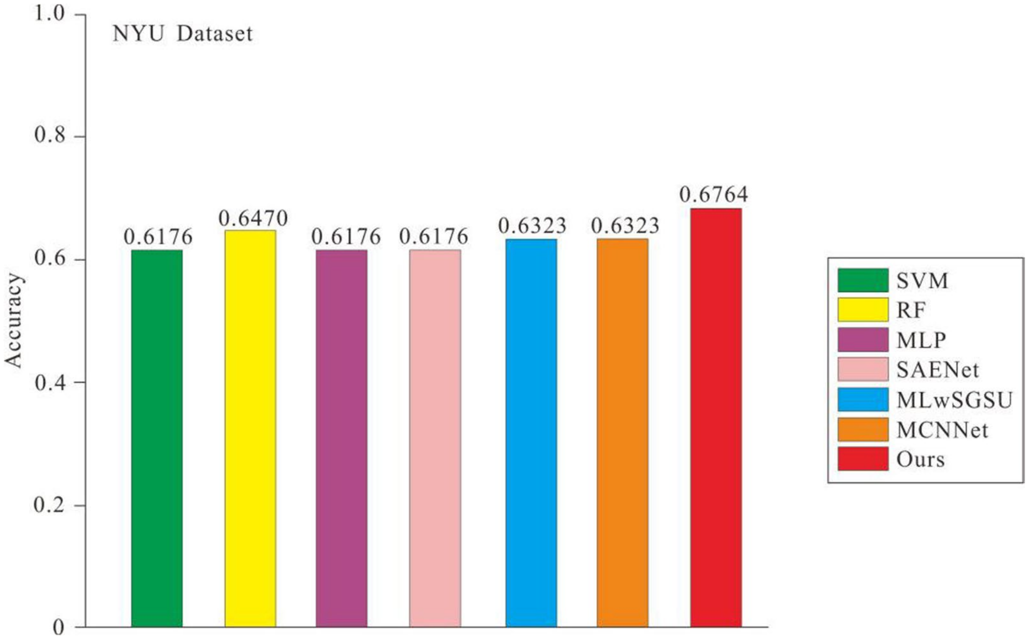

Fig. 6 Accuracy comparison of different network models in NYU dataset

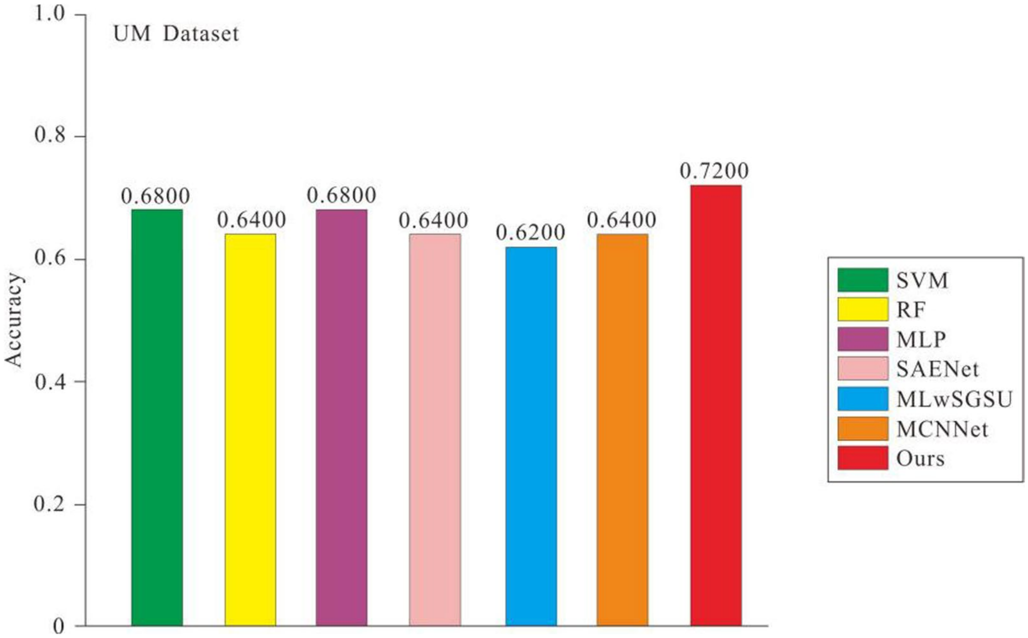

Fig. 7 Accuracy comparison of different network models in UM dataset

Conclusion

In this study, we propose the multi-task Transformer network, which are essential for predicting and diagnosing ASD diseases. The proposed network utilizes multi-task learning and attention mechanisms for ASD recognition and achieves excellent classification performance on NYU and UM ASD datasets. In addition, the attention mechanism enhances the model's attention to ASD-related features. Multi-task learning enhances the model generalization performance by fusing knowledge learned from different ASD datasets. We evaluated our method

Table 3 Experimental results of different methods

Site	Method	Accuracy (%)	Sensitivity (%)	Specificity (%)
NYU	SVM	61.76	46.66	73.68
	RF	64.70	46.66	78.94
	MLP	61.76	46.66	73.68
	SAENet	61.76	40.00	78.94
	MLwSGSU	63.23	53.33	71.05
	MCNNet	63.23	49.99	73.68
	Ours	67.64	40.00	89.47
UM	SVM	68.00	40.00	86.66
	RF	64.00	40.00	80.00
	MLP	68.00	70.00	66.66
	SAENet	64.00	60.00	66.66
	MLwSGSU	62.00	65.00	60.00
	MCNNet	64.00	50.00	73.33
	Ours	72.00	50.00	86.66

on two public datasets and found that it outperformed several state-of-the-art methods with high performance. The results show that combining multitask learning and attention mechanism can better classify ASD patients and NC patients.

Acknowledgements
The authors have no additional acknowledgments to make.

Author contributions
Conceptualization, L.G. and J.H.; methodology, J.H.; software, Z.W., X. Z.; validation, L.G., Y.L. and J.H.; formal analysis, J.H.; investigation, L.G, X. Z.; resources, H.S.; data curation, Y.L, Y. Y.; writing—original draft preparation, L.G.; writing—review and editing, J.H.; visualization, Z.W.; supervision, J.H.; project administration, L. G. and Y.L.; funding acquisition, L.G. All authors have read and agreed to the published version of the manuscript.

Funding
Thanks for the research funding support from Guangzhou Huali College.

Availability of data and materials
Supplementary material for this article is available on network data set. http://fcon_1000.projects.nitrc.org/indi/abide/.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interest.

Received: 6 September 2023 Accepted: 1 May 2024
Published online: 13 June 2024

References

1. Rapin I, Tuchman RF. Autism: definition, neurobiology, screening, diagnosis. Pediatr Clin North Am. 2008;55(5):1129–46.
2. Maenner MJ, Shaw KA, Bakian AV, Bilder DA, Durkin MS, Esler A, Hughes MS, Lee L, Williams S, Cogswell ME. Prevalence and characteristics of autism spectrum disorder among children aged 8 years—autism and develop-mental disabilities monitoring network, 11 sites, United States, 2018. Morb Mortal Wkly Rep. 2011;70(11):1–16.
3. Hayes SA, Watson SL. The impact of parenting stress: a meta-analysis of studies comparing the experience of parenting stress in parents of children with and without autism spectrum disorder. J Autism Dev Disord. 2013;43(3):629–42.
4. Webb SJ, Jones EJH, Kelly J, Dawson G. The motivation for very early intervention for infants at high risk for autism spectrum disorders. Int J Speech Lang Pathol. 2014;16(1):36–42.
5. Rutter M, Le CA, Lord C. Autism diagnostic interview-revised, vol. 29. Los Angeles: Western Psychological Services; 2003. p. 30.
6. Falkmer T, Anderson K, Falkmer M, Horlin C. Diagnostic procedures in autism spectrum disorders: a systematic literature review. Eur Child Adolesc Psychiatry. 2013;22(6):329–40.
7. Vieira S, Pinaya WH, Mechelli A. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: methods and applications. Neurosci Biobehav Rev. 2017;74:58–75.
8. Duchesnay E, Cachia A, Boddaert N, Chabane N, Mangin JF, Martinot JL. Feature selection and classification of imbalanced datasets: application to pet images of children with autistic spectrum disorders. Neuroimage. 2011;57(3):1003–14.
9. Feng Q, Huang YJ, Long Y, Gao L, Gao X. A deep spatiotemporal attention network for mild cognitive impairment identification. Front Aging Neurosci. 2022;14: 925468.
10. Bind S, Tiwari AK, Sahani AK, Koulibaly P, Nobili F, Pagani M. A survey of machine learning based approaches for parkinson disease prediction. Int J Comput Sci Inform Technol. 2015;6(2):1648–55.
11. Ye J, Wu T, Li J, Chen K. Machine learning approaches for the neuroimaging study of Alzheimer's disease. Computer. 2011;44(4):99–101.
12. Pellegrini E, Ballerini L, Hernandez MDCV, Chappell FM, González-Castro V, Anblagan D, Trucco E, Wardlaw JM. Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: a systematic re-view. Alzheimers Dement. 2018;10:519–35.
13. Eslami T, Saeed F. Similarity based classification of ADHD using singular value decomposition. In: Eslami T, Saeed F, editors. ACM international conference on computing frontiers (Ischia). New York: Association for Computing Machinery; 2018. p. 19–25.
14. Itani S, Lecron F, Fortemps P. A multi-level classification framework for multi-site medical data: application to the ADHD-200 collection. Exp Syst Appl. 2018;91:36–45.
15. Pagnozzi AM, Conti E, Calderoni S, Fripp J, Rose SE. A systematic review of structural MRI biomarkers in autism spectrum disorder: a machine learning perspective. Int J Dev Neurosci. 2018;71:68–82.
16. Hyde KK, Novack MN, LaHaye N, Parlett-Pelleriti C, Anden R, Dixon DR, Linstead E. Applications of super-vised machine learning in autism spectrum disorder research: a review. Rev J Autism Dev Disord. 2019;6(2):128–46.
17. Gao S, Calhoun VD, Sui J. Machine learning in major depression: from classification to treatment outcome prediction. CNS Neurosci Therap. 2018;24(11):1037–52.
18. Hosseinzadeh M, Koohpayehzadeh J, Bali AO, Rad FA, Souri A, Mazaherinezhad A, Bohlouli M. A review on diagnostic autism spectrum disorder approaches based on the Internet of things and machine learning. J Supercomput. 2021;77(3):2590–608.
19. Crippa A, Salvatore C, Perego P, Forti S, Nobile M, Molteni M. Use of machine learning to identify children with autism and their motor abnormalities. J Autism Dev Disord. 2015;45(7):2146–56.
20. Rane S, Jolly E, Park A, Jang H, Craddock C. Developing predictive imaging biomarkers using whole-brain classifiers: application to the ABIDE I dataset. Res Ideas Outcomes. 2017;3(3):1–5.
21. Abbas H, Garberson F, Glover E, Wall DP. Machine learning approach for early detection of autism by combining questionnaire and home video screening. J Am Med Inform Assoc. 2018;25(8):1000–7.
22. Heinsfeld AS, Franco AR, Craddock RC, Buchweitz A, Meneguzzi F. Identification of autism spectrum disorder using deep learning and the ABIDE dataset. NeuroImage Clin. 2018;17:16–23.

23. Alsaade FW, Alzahrani MS. Classification and detection of autism spectrum disorder based on deep learning algorithms. Comput Intell Neurosci. 2022;2022:8709145.
24. Pavăl D. A dopamine hypothesis of autism spectrum disorder. Dev Neurosci. 2017;39(5):355–60.
25. Ebrahimi MS, Rostam-Abadi Y, Rezaei N. Autism spectrum disorders and natural killer cells: a review on pathogenesis and treatment. Exp Rev Clin Immunol. 2021;17(1):27–35.
26. Santana CP, Carvalho EA, Rodrigues ID, Bastos GS, Brito LL. rs-fMRI and machine learning for ASD diagnosis: a systematic review and meta-analysis. Sci Rep. 2022;12(1):1–20.
27. Tabarestani S, Aghili M, Eslami M, Cabrerizo M, Barreto A, Rishe N, Curiel RE, Loewenstern D, Duara R, Adjouadi M. A distributed multitask multimodal approach for the prediction of Alzheimer's disease in a longitudinal study. Neuroimage. 2020;206(1):1–14.
28. Chelaramani S, Gupta M, Agarwal V, Gupta P, Habash R. Multi-task knowledge distillation for eye disease prediction. In: Chelaramani S, editor. Proceedings of the IEEE/CVF winter conference on applications of computer vision. Waikoloa: WACV; 2021. p. 3983–93.
29. Vaswani A, Shazeer N, Parmar N, Uzskoreit J, Jones L, Gomez AN, Kaiser L, Polosuhin I. Attention is all you need. Adv Neural Inform Process Syst. 2017;2017:6000–10.
30. Song H, Chen L, Gao RQ, Bogdan IM, Yang J, Wang SL, Dong WT, Quan WX, Dang WM, Yu X. Automatic schizophrenic discrimination on fNIRS by using complex brain network analysis and SVM. BMC Med Inform Decis Mak. 2017;17(3):1–9.
31. Fredo AJ, Jahedi A, Reiter M, Muller R. Diagnostic classification of autism using resting-state fMRI data and conditional random forest. Age (years). 2018;12(2):6–41.
32. Shanmuganathan S. Artificial neural network modelling: an introduction. Stud Comput Intell. 2016;628(1):1–14.
33. Almuqhim F, Saeed F. ASD-SAENet: a sparse autoencoder, and deep-neural network model for detecting autism spec-trum disorder (ASD) using fMRI data. Front Comput Neurosci. 2021;15:1–10.
34. Lee S, Son Y. Multitask learning with single gradient step update for task balancing. Neurocomputing. 2022;467:442–53.
35. Heuer F, Mantowsky S, Bukhari S, Schneider G. Multitask-centernet (mcn): Efficient and diverse multitask learning using an anchor free approach. In: Heuer F, editor. Proceedings of the IEEE/CVF International conference on computer vision. Montreal: IEEE/CVF; 2021. p. 997–1005.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.