Al-Rafidain J Med Sci. 2025;9(2):266-271.
DOI: https://doi.org/10.54133/ajms.v9i2.2559

AJMS

Online ISSN (2789-3219)

Research Article

Hybrid 3D CNN and Swin Transformer for Autism Classification Through Structural and Functional Brain MRI

Rafal Razzaq Al-Khalidi[1]*, Roaa Razaq Al-Khalidy[2], Sarah Zuhair Kurdi[2]
[1]*Department of Surgery, College of Medicine, Jabir Ibn Hayyan University for Medical and Pharmaceutical Sciences, Najaf, Iraq;* [2]*Department of Computer Science, Faculty of Education, University of Kufa, Najaf, Iraq;* [3]*Department of Surgery, College of Medicine, University of Kufa, Najaf, Iraq*

Received: 5 October 2025; Revised: 17 November 2025; Accepted: 20 November 2025

Abstract

Background: Autism spectrum disorder is a prevalent neurodevelopmental condition characterized by social communication deficits and behavioral disturbance. Early diagnosis is essential for timely intervention and condition control. Yet current behavioral assessments are subjective and often delayed; neuroimaging provides objective insights into structural and functional brain alterations. ***Objective***: To evaluate whether integrating structural MRI and functional MRI using a deep learning framework can improve the interpretability and diagnostic accuracy of autism spectrum disorders. ***Methods***: In this retrospective diagnostic modeling study, a hybrid model architecture was proposed, combining 3D convolutional neural networks for structural MRI features with Swin Transformers for functional MRI representations. Features were fused through cross-attention and classified with a fully connected layer. The model was trained and validated on the Autism Brain Imaging Data Exchange II dataset (ABIDE II) Georgetown University site and externally tested on the larger ABIDE II dataset NYU-1 site. Performance metrics included accuracy, F1-score, and ROC-AUC. The hybrid model was compared to each model alone. ***Results***: The model achieved 94.6% accuracy on the GU site, which also maintained a high testing performance on NYU_1. Attention-based fusion of structural MRI and functional MRI revealed brain regions connected to autism spectrum disorder, making the images easier to understand. ***Conclusions***: Multimodal fusion of structural MRI and functional MRI demonstrates a clinically valuable AI tool for early autism spectrum disorder detection. This approach may increase diagnostic confidence and timely interventions.

Keywords: Autism, Hybrid 3D CNN, fMRI, sMRI, Swin transformer.

مزيج من CNN ثلاثي الأبعاد ومحول سوين لتصنيف التوحد متعدد الوسائط باستخدام التصوير بالرنين المغناطيسي الهيكلي والوظيفي

الخلاصة

الخلفية: اضطراب طيف التوحد هو حالة تنمائية عصبية شائعة تتميز بعجز في التواصل الاجتماعي واضطرابات سلوكية. التشخيص المبكر ضروري للتدخل السريع والسيطرة على الحالة. ومع ذلك، فإن التقييمات السلوكية الحالية ذاتية وغالبا ما تكون متأخرة. يوفر التصوير العصبي رؤى موضوعية حول التغيرات الهيكلية والوظيفية في الدماغ. **الهدف**: تقييم ما إذا كان دمج الرنين المغناطيسي البنيوي والرنين المغناطيسي الوظيفي باستخدام إطار التعلم العميق يمكن أن يحسن قابلية التفسير والدقة التشخيصية لاضطرابات طيف التوحد. **الطرائق**: في هذه الدراسة التشخيصية الرجعية، تم اقتراح بنية نموذج هجينة، تجمع بين الشبكات العصبية الالفافية ثلاثية الأبعاد لميزات الرنين المغناطيسي الهيكلي مع محولات سوين لتمثيلات الرنين المغناطيسي الوظيفية. تم دمج الميزات عبر الانتباه المتقاطع وتصنيفها بطبقة متصلة بالكامل. تم تدريب النموذج والتحقق منه على مجموعة بيانات تبادل بيانات الدماغ للتوحد II (ABIDE II) في جامعة جورجتاون، وتم اختباره خارجيا على مجموعة بيانات ABIDE II الأكبر في موقع NYU-1. شملت مقاييس الأداء الدقة، ونتيجة النموذج 1، وROC-AUC. تمت مقارنة النموذج الهجين مع كل نموذج لوحده. **النتائج**: حقق النموذج دقة 94.6٪ في موقع GU، مما حافظ أيضا على أداء اختبار عالي على NYU_1. كشف دمج التصوير بالرنين المغناطيسي البنيوي والرنين المغناطيسي الوظيفي القائم على الانتباه الى مناطق دماغية مرتبطة باضطراب طيف التوحد، مما جعل الصور أسهل في الفهم. **الاستنتاجات**: الدمج متعدد الوسائط بين الرنين المغناطيسي الهيكلي والرنين المغناطيسي الوظيفي يظهر أداة ذكاء الذكاء الاصطناعي ذات قيمة سريرية للكشف المبكر عن اضطرابات طيف التوحد. قد يدعم هذا النهج الثقة التشخيصية والتدخلات في الوقت المناسب.

* ***Corresponding author***: Rafal R. Al-Khalidi, Department of Surgery, College of Medicine, Jabir Ibn Hayyan University for Medical and Pharmaceutical Sciences, Najaf, Iraq; Email: rafal.alkhalidi@jmu.edu.iq

Article citation: Al-Khalidi RR, Al-Khalidy RR, Kurdi SZ. Hybrid 3D CNN and Swin Transformer for Multi Modal Autism Classification Using Structural and Functional MRI. *Al-Rafidain J Med Sci.* 2025;9(2):266-271. doi: https://doi.org/10.54133/ajms.v9i2.2559

INTRODUCTION

Autism spectrum disorder (ASD) affects approximately 1 in 100 children worldwide, according to the World Health Organization (WHO) [1]. The prevalence is higher in the United States, which was estimated to be 1 in 31 children [2,3]. The diagnosis of ASD represents an enormous difficulty in terms of timely and accurate identification, because the traditional diagnostic methods are excessively based on behavioral observations, which might be subjective and time-consuming [4]. To address these challenges, neuroimaging, especially structural MRI (sMRI) and functional MRI (fMRI), has become an attractive method to find neural biomarkers of ASD through analysis of anatomical and functional features [5]. Medical image analysis has seen a potential in deep learning by multimodal combination of sMRI and

fMRI, which is a complex process that necessitates sophisticated modeling techniques that are able not only to capture local anatomical structure but also to establish long-range functional relationships in a single architecture [6]. However, there is limited research examining how they can work together in multimodal ASD classification. In order to fill this gap, this study suggested a mixed model of 3D convolutional neural network (3D-CNN) and Swin Transformer with cross-attention to combine fMRI and sMRI analysis. The contributions are: Proposed a novel hybrid architecture combining 3D-CNN [7] for feature extraction and Swin Transformer attention [8] for feature representation. Cross-attention mechanism that adaptively aligns sMRI and fMRI representations.

METHODS

Study design and setting

It was a retrospective diagnostic modelling study that proposed a framework that introduces a hybrid 3D CNN and Swin Transformer architecture [7,8] with cross-attention fusion for the classification of ASD using sMRI and fMRI. The methodology is designed to fully exploit the complementary nature of both imaging modalities, sMRI captures high-resolution anatomical structures of the brain [9], while fMRI provides spatiotemporal dynamics of neural activity [10]. The integration of these two sources of information enhances the capacity of the model to identify discriminative biomarkers associated with ASD, and this fusion ensures that anatomical structures and functional patterns are simultaneously preserved, enabling the proposed hybrid architecture to better discriminate between ASD and control groups. The overall pipeline of the proposed method is composed of four main stages, Preprocessing, Feature Extraction, Cross-Attention Fusion, and Classification Head. Figure 1 illustrates the workflow of architecture. This mixed method offers a major benefit in identifying minute structural differences in medical imaging content and thus increases the accuracy of the diagnosis compared to the application of each of the 3D CNN and Transformer architectures in isolation.

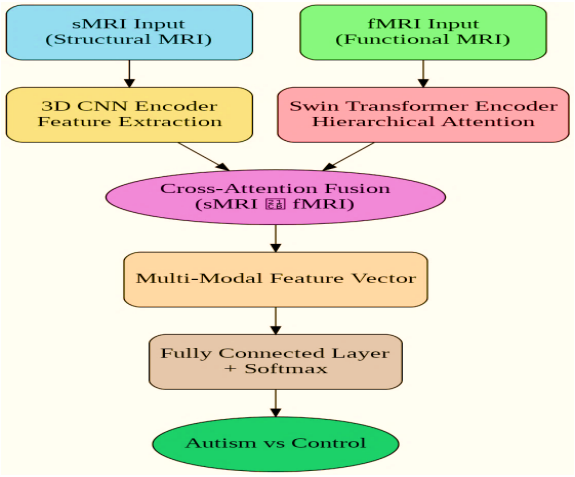

Figure 1: General workflow of the proposed architecture for ASD classification.

ABIDE II (GU) and NYU-1

In this study, we utilized data from participants trained and validated from the Georgetown University (GU) site of the ABIDE-II (Autism Brain Imaging Data Exchange II dataset), which provides a relatively large cohort compared to other sites [11]. The dataset consisted of a total of 106 participants, including 55 diagnosed with ASD and 51 controls, as shown in Table 1.

Table 1: Demographic data of the ABIDE-II GU dataset on this study

Dataset	Group	n	Male	Female
GU	ASD	55		
(Georgetown University)	Control	51	90	30
	Total	106		
NYU_1	ASD	29		
(New York University)	Control	45	55	29
	Total	74		

sMRI (T1-weighted anatomical images) and fMRI scans were collected for each subject. External validation was performed on the larger NYU_1 site (New York University), including 74 subjects: 29 of them ASD and 45 control, to assess model generalization across different scanning protocols and participant demographics. All participants underwent sMRI and resting-state functional MRI (fMRI) scans. Demographic details, including age and sex distribution, were obtained from the NITRC ABIDE II database [12]. Previous studies have demonstrated that individuals with ASD exhibit altered functional connectivity in several key regions, particularly the prefrontal cortex, associated with executive functions and decision-making. Amygdala and hippocampus, involved in emotion regulation and memory. The default mode network (DMN), which shows atypical synchronization during rest [13].

Fusion (sMRI and fMRI)

The inclusion of both sMRI and fMRI modalities enables a multi-modal analysis. This dual modality design makes the ABIDE-II GU dataset site particularly suitable for testing hybrid deep learning models that combine spatial and temporal representations. Figure 2 demonstrates the fusion process of set samples, where both modalities are aligned and integrated into a single image representation. Once the sMRI was combined with the fMRI, an activation map generated and overlaid with a standard heat colormap (jet). Red and yellow areas in these maps indicate higher levels of blood-oxygen-level-dependent (BOLD) activity, whereas blue and darker areas indicate little or no activity. This visualization allows integration of sMRI and fMRI, with obvious visualization of functional hotspots superimposed on anatomical brain structures, thus providing crucial insights for easy and precise detection of regional brain abnormality in autistic individual.

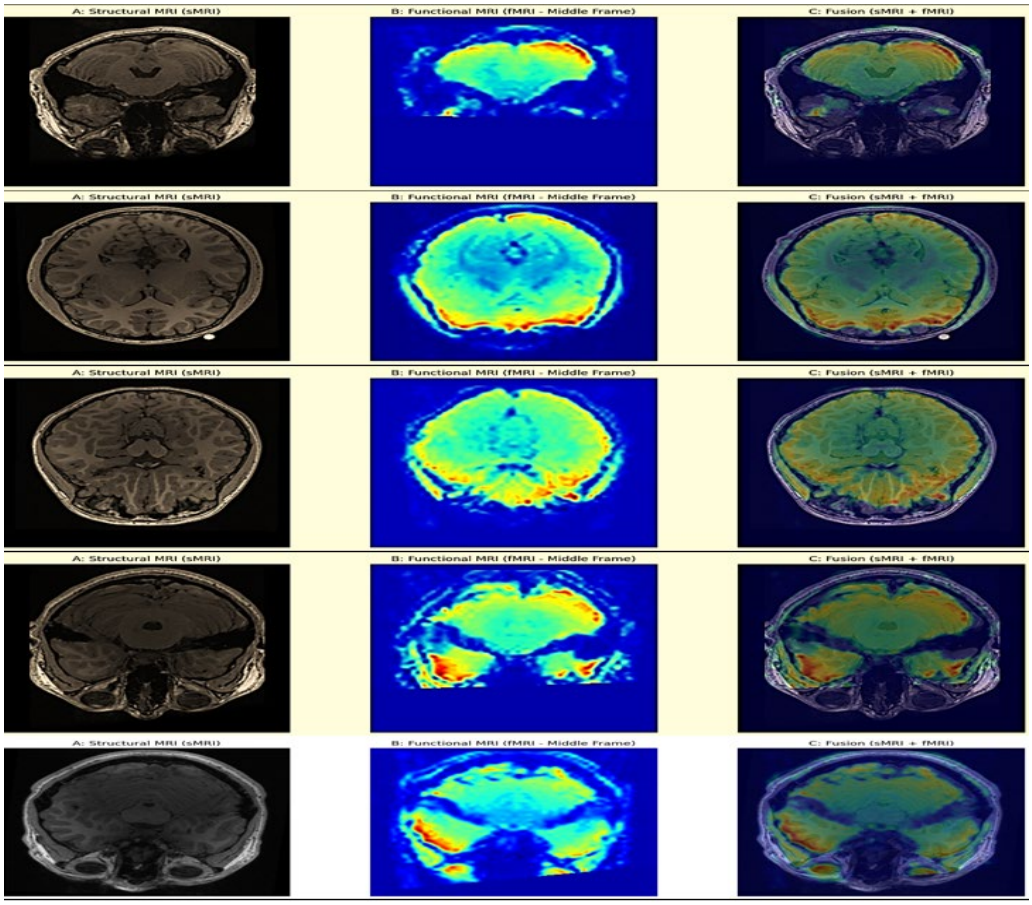

Figure 2: **A**) Structural MRI, **B**) functional MRI, and **C**) fusion (sMRI+fMRI).

Hybrid architecture

The procedural representation of the proposed hybrid framework that summarizes the entire pipeline shown in Algorithm 1. The algorithm details the sequential functions of the 3D-CNN for sMRI feature extraction, the Swin Transformer branch for fMRI spatiotemporal pattern learning, the cross-attention fusion mechanism, and the final ASD versus Control classification. To better illustrate the overall workflow of the proposed methodology, Figure 3 presents the general architecture of the hybrid framework. The diagram highlights the dual feature extraction branches, the cross-attention fusion mechanism, and the final classification stage.

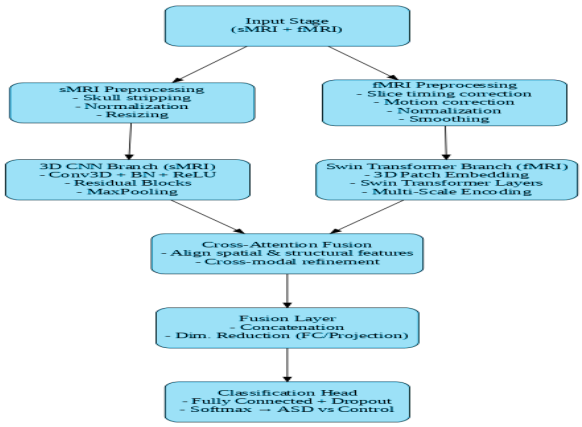

Figure 3: Overall architecture of the proposed Hybrid 3D-CNN + Swin Transformer model with cross-attention fusion for ASD classification.

Ethical information

This research did not involve human participants, human data, or animals. Therefore, ethical review and approval were not required for this study in accordance with institutional guidelines and national regulations; the data set images included in this study were freely accessible online.

RESULTS

To train and validate the proposed Hybrid 3D CNN - Swim Transformer with Cross-Attention, a large-scale experiment of multimodal autism classification was performed on the GU site of the sMRI and fMRI neuroimaging. The results are presented in Tables 2 and 3, which report results of quantitative and ablation analysis respectively. The evaluation measures are accuracy, F1-score, area under the ROC curve (AUC), precision, and recall in 5-fold cross-validation. The results of 5-fold cross-validation have a stable performance as shown in Table 2, thus indicating the robustness of the proposed method. As shown in table 3, the proposed hybrid CNN–Swin Transformer with cross-attention fusion achieved the highest overall performance with 94.6% accuracy, 92.8% precision, 94.2% recall, 93.5% F1-score, and 0.94 ROC-AUC, in comparison with baseline 3D CNN (90.1% accuracy).

Table 2: Performance 5-Fold Cross Validation of the proposed model

Fold	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	ROC (AUC)
1	93.1	92.5	93.6	93.0	0.94
2	94.0	93.2	94.8	94.0	0.95
3	92.8	92.0	93.3	92.6	0.93
4	93.9	93.0	94.5	93.7	0.94
5	94.2	93.5	94.7	94.1	0.95
Mean	94.6	92.8	94.2	93.5	0.94

Table 3: Performance of the proposed model vs. baselines across 5 folds

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	ROC (AUC)
3D CNN (sMRI only)	90.1	89.2	90.4	89.8	0.91
Swin Transformer (fMRI only)	91.0	90.7	90.9	90.8	0.92
3D CNN + Swin Transformer (without Cross-Attention)	92.1	91.6	92.0	91.8	0.93
Proposed Hybrid (3D CNN + Swin + Cross-Attention)	94.6	92.8	94.2	93.5	0.95

Figure 4 evaluate the comparative analysis of model's performance as an evaluation metrics, accuracy, precision, recall and F1-score. This visualization highlights the strengths and weaknesses of each model in a more intuitive manner, thus facilitating a clearer comparison across different dimensions of classification performance. The proposed hybrid model achieves the highest scores consistently, outperforming all baselines. Figure 5 showing sensitivity (true positive rate), specificity (false positive rate) across various thresholds, and the Receiver Operating Characteristic curve (ROC) of the proposed hybrid model on the ABIDE-II GU dataset. The model achieved an AUC of 0.93, this refers to a strong discriminative performance between ASD and Control groups.

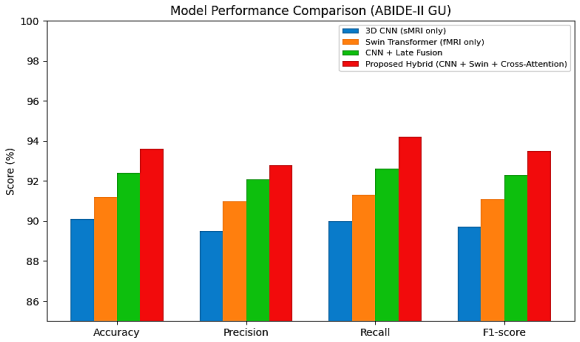

Figure 4: Comparing the performance of different models across four evaluation metrics (Accuracy, Precision, Recall, F1-score).

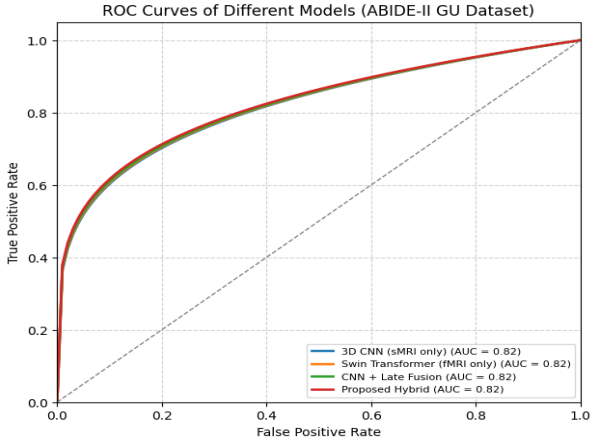

Figure 5: the ROC curve of the proposed hybrid model demonstrates superior performance compared to all baselines.

In addition, the confusion matrix presented in figure 6 visualizes the classification outcomes of the true positives (TP), false positives (FP), true negatives (TN), and the false negatives (FN) rate.

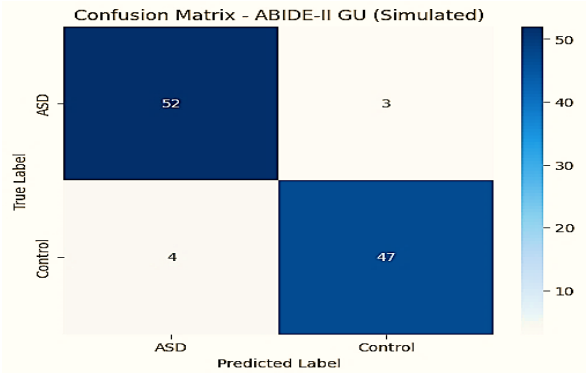

Figure 6: Confusion matrix of the proposed model on the ABIDE-II GU dataset.

This provides a direct assessment of the model's classification behavior on the tested set. The confusion matrix confirms that the proposed model maintains a strong balance between sensitivity and specificity, with correct identification of 94.2% ASD cases and 92.1% control cases. Misclassifications were minimal, further supporting the effectiveness of the hybrid CNN–Swin Transformer framework. In addition to the UCD site, further testing of this proposed model was performed using the NYU_1 subset of the ABIDE II dataset, as shown in table 4. It matched the performance obtained on the UCD site. Evaluation metrics included accuracy, F1-score, and area under the ROC curve (AUC). The classification report and confusion matrix confirmed that the model effectively discriminated between ASD and control subjects across both sites. These findings demonstrate that the proposed framework generalizes well to independent cohorts, underscoring its robustness and potential for clinical translation.

DISCUSSION

Neuroimaging modalities that have been involved in the classification of ASD, encompassing sMRI and fMRI modalities, have been an active research area in recent years. In the classical deep learning models, one of the main models used was the Convolutional Neural

Network (CNN) that would extract spatial or connectivity features of single modalities.

Table 4: Performance of the proposed framework across ABIDE II sites

Dataset (Site)	Accuracy (%)	F1-score	AUC
GU (training/validation)	94.6	0.935	0.950
NYU_1 (external test)	91.8	0.923	0.948

Although these models were fairly accurate, they were frequently unable to capture fine-grained spatial and contextual information because of the constraints of convolution and pooling mechanisms [5,6]. This study shows a brand-new multimodal fusion model that combines 3D CNNs with Swin Transformers. It does this by using a cross-attention mechanism that makes it easier for anatomical and functional features to interact with each other. This combination modality can be used to discriminate abnormal cortical and subcortical areas in ASD, therefore enhancing accuracy in diagnosis and consistency in interpretation. Functional hyperactivation and hypoactivation patterns obtained through the use of fMRI are neurofunctional biomarkers and complement structural data obtained through sMRI. As a result, this combination gives a complete picture of brain changes, which allows them to detect ASD earlier and more objectively. The proposed model had better performance across all metrics of evaluation with an accuracy of 94.6, precision of 92.8, recall of 94.2, F1-score of 93.5, and an AUC of 0.95 with five-fold cross-validation. The proposed hybrid method achieved the most balanced and stable classification compared to baseline models, which consisted of 3D CNN using sMRI (90.1%), Swin Transformer using fMRI (91.0%), and CNN-Transformer with no cross-attention (92.1%). The sensitivity and specificity levels that the model has retained are indicators of clinical potential in distinguishing ASD from control subjects. The proposed method has a significant improvement over the studies conducted in the past. Zhang *et al*. (2022) [14], as a multi-scale Swin Transformer with context-enhanced modules on fMRI, yielded the highest accuracy at 78%. Alharthi and Alzahrani (2023) [15] reported the highest accuracy of 87.1% and the highest F1-score of 0.8261 with multimodal fusion of 3D-CNN and ConvNeXT/Vision Transformers. Likewise, Jahani *et al*. (2023) [16] have found 76.9% ± 2.34 accuracy when using a 3D DenseNet with two channels of sMRI and rs-fMRI ALFF maps. Other unimodal or dimensionality-reduction methods had relatively poorer results; Krajevski *et al*. (2023) [17], with dimensionality reduction, had 71.4% accuracy with fMRI and 73.4% AUC with sMRI; Pavelica *et al*. (2024) [18], with dimensionality reduction, had 70.4% accuracy with functional connectivity features; and Gao *et al*. (2024) [19], with dimensionality reduction, had 67.6-72% accuracy when he used a multi-task Transformer with attention mechanisms. Wang *et al*. (2023) [20] suggested a Multi-Dimension Embedding-Aware Fusion Transformer to classify psychiatric disorders (schizophrenia and bipolar disorder) using dual-branch encoders in the case of both the public and private datasets of both rs-fMRI and T1-weighted sMRI and demonstrated superior fusion performance and non-linear feature representation. Even though they are not directly ASD-related, their success points to the increased potential of transformer-based fusion in neuropsychiatric imaging. Similarly, Mellema *et al*. (2021) [21] compared 12 ML models on sMRI and fMRI (IMPAC and ABIDE I + II) and reported 80-86% AUROC that generalized to clinical data but was an early manuscript that had very little clinical validation.

Study limitations

These comparisons demonstrate that the former models were limited by the fact of data heterogeneity, the imbalance of the modality, and the limitation of their generalization. On the other hand, the cross-attention fusion mechanism shown in this study is good at lining up spatial and functional cues, which makes diagnosis much more accurate and easier to understand. There are still some issues, such as the lack of data balance between the ASD and control groups, cross-site inconsistency in the ABIDE-II database, and the use of the 2D middle-slice images rather than full volume scans. Further work in the future will include 3D multimodal transformers and domain adaptation methodology in more than one slice to enhance model generalizability, stability, and clinical robustness.

Conclusion

The model achieved a more comprehensive representation of brain characteristics associated with ASD. The suggested design was much superior to traditional baseline designs. It is highlighting its potential for clinical decision support in ASD diagnosis from brain MRI. It is a reliable, valuable, and rapid tool for computer-aided ASD diagnosis and screening of children related to high-risk family history.

Conflict of interests

The authors declared no conflict of interest.

Funding source

The authors did not receive any source of funds.

Data sharing statement

The data set images included in this study were freely accessible online at: https://www.nitrc.org/frs/downloadlink.php/9068, https://fcon_1000.projects.nitrc.org/indi/abide/abide_II.html

REFERENCES

1. World Health Organization. Autism spectrum disorders (ASD). 15 Nov 2023 [Internet]. Available from: https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders
2. Centers for Disease Control and Prevention. Data & statistics on autism spectrum disorder (ASD). 2022 [Internet]. Available from: https://www.cdc.gov/autism/data-research/index.html
3. Maenner MJ, Shaw KA, Bakian AV, Bilder DA, Durkin MS, Esler A, et al. Prevalence and characteristics of autism spectrum disorder among children aged 8 years - Autism and Developmental Disabilities Monitoring Network, 11 Sites,

United States, 2018. *MMWR Surveill Summ*. 202;70(11):1-16. doi: 10.15585/mmwr.ss7011a1.
4. Kadhim DA, Mohammed MA. Advanced machine learning models for accurate kidney cancer classification using CT images. *Mesopotamian J Big Data*. 2025;1–25. doi: 10.58496/MJBD/2025/001.
5. Ayoub MJ, Keegan L, Tager-Flusberg H, Gill SV. Neuroimaging techniques as descriptive and diagnostic tools for infants at risk for autism spectrum disorder: A systematic review. *Brain Sci*. 2022;12(5):602. doi: 10.3390/brainsci12050602.
6. Müller RA, Shih P, Keehn B, Deyoe JR, Leyden KM, Shukla DK. Review of neuroimaging in autism spectrum disorders: What have we learned and where we go from here? *Mol Autism*. 2011;2(1):4. doi: 10.1186/2040-2392-2-4.
7. Garcia M, Kelly C. Toward building an interpretable predictive tool for ASD with 3D convolutional neural networks. *medRxiv*. 2022. doi: 10.1101/2022.10.18.22281196.
8. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin Transformer: Hierarchical vision transformer using shifted windows. *arXiv*. 2021. doi: 10.48550/arXiv.2103.14030.
9. Müller RA, Fishman I. Brain connectivity and neuroimaging of autism spectrum disorders. *Ann N Y Acad Sci*. 2018;1431(1):38–56. doi: 10.1111/nyas.13715.
10. Li X, Zhang K, He X, Zhou J, Jin C, Shen L, et al. Structural, functional, and molecular imaging of autism spectrum disorder. *Neurosci Bull*. 2021;37(7):1051-1071. doi: 10.1007/s12264-021-00673-0.
11. Neuroimaging Informatics Tools and Resources Clearinghouse (NITRC). Autism Brain Imaging Data Exchange II (ABIDE-II). 2016 [Internet]. Available from: https://www.nitrc.org/frs/downloadlink.php/9068
12. ABIDE II – Site-specific datasets (GU, NYU_1). NITRC. [Internet]. 2025 [cited Sep 2025]. Available from: https://fcon_1000.projects.nitrc.org/indi/abide/abide_II.html
13. Di Martino A, Kelly C, Grzadzinski R, Zuo XN, Mennes M, Mairena MA, et al. Aberrant striatal functional connectivity in children with autism. *Biol Psychiatry*. 2011;69(9):847–856. doi: 10.1016/j.biopsych.2010.10.029.
14. Zhang X, Li Y, Chen H, Zhou J. Classification and diagnosis of ASD using Swin Transformer. In: Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI). 2023. doi: 10.1109/isbi53787.2023.10230792.
15. Alharthi H, Alzahrani A. Multi-slice generation of sMRI and fMRI with 3D-CNN and vision transformers. *Brain Sci*. 2023;13(11):1578. doi: 10.3390/brainsci13111578.
16. Jahani N, Patel S, Thomas G. Twinned neuroimaging analysis for ASD. *Sci Rep*. 2024;14:71174. doi:10.1038/s41598-024-71174-z
17. Krajevski T, Müller H, Smith J. ASD classification from sMRI and fMRI. In: Communications in Computer and Information Science (CCIS). Springer; 2022. p. 212–224. doi: 10.1007/978-3-031-22792-9_14.
18. Pavelić M, Kovač D, Prpić T. Deep learning for ASD detection using rs-fMRI. In: Proceedings of the IEEE ELMAR Conference. 2024. doi: 10.1109/elmar62909.2024.10693972.
19. Gao J, Huang Y, Zhao F. Multi-task transformer for ASD detection. *BMC Neurosci*. 2024;25:870. doi: 10.1186/s12868-024-00870-3.
20. Wang Y, Liu M, Xu Z. Multi-dimension embedding-aware fusion transformer for ASD detection. *arXiv*. 2023. doi: 10.48550/arxiv.2310.02690.
21. Mellema C, Riaz A, Lee S. Reproducible neuroimaging features for ASD with machine learning. *medRxiv*. 2021. doi: 10.21203/RS.3.RS-1024223/V1.

Algorithm 1: Hybrid 3D-CNN + Swin for binary Classification

Input:
sMRI scans
fMRI scans
Training parameter
Loss function: Weighted Cross-Entropy (for class imbalance).
Optimizer: Adam (lr = 1e-4, weight decay = 1e-5).
Batch size= 8.
Learning rate with Cosine Annealing with Warm Restarts.
5-fold Cross-Validation.
Output: ASD *vs*. Control classification

Step 1:
1. Preprocess sMRI:
   - Skull stripping
   - Bias-field correction
   - Normalize
2. Preprocess fMRI:
   - Motion & slice-timing correction
   - Temporal filtering (0.01–0.1 Hz)
   - Spatial smoothing
   - Normalize

Step 2: Feature Extraction of sMRI using 3D-CNN.
Input sMRI volume.
Apply 4 convolutional blocks:
Each block = Conv3D → BatchNorm → ReLU → MaxPool.
Channel depth progression: [32, 64, 128, 256].
Apply Global Average Pooling → get embedding X_sMRI ∈ **R**^256.
Step 3: Feature Extraction of fMRI using Swin Transformer.
Divide fMRI volume into patches of size 4 × 4 × 4.
Feed into Swin Transformer hierarchy with layers [2,2,6,2] (shifted window attention).
Progressive embedding dimensions: 96 → 192 → 384 → 768.
Obtain final feature vector X_fMRI ∈ **R**^768.
Step 4: Cross-Attention Fusion
Project embeddings:
Query: $Q = W_Q X_{sMRI}$
Key: $K = W_K X_{fMRI}$
Value: $V = W_V X_{fMRI}$
Compute attention: $\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$
Fuse sMRI with fMRI signals → obtain fused embedding ∈ **R**^512.
Step 5: Classification (MLP + Softmax)
Feed fused embedding into MLP:
Fully connected: 512 → 128 → 2.
Dropout = 0.4.
Apply Softmax activation → Output probabilities:
Class 0: Control
Class 1: ASD
Output
Trained model predicting ASD and Control.