AI in education

Article

Multi-Architecture Convolutional Neural Networks with Attention Mechanisms for Autism Spectrum Disorder Classification

Ioana Diana Moldovanu [1], Ecaterina Popa [1] and Simona Moldovanu [2,3,*]

[1] Department of Psychology, Danubius International University, Bulevardul Galați 3, 800654 Galati, Romania; m.ioana31@yahoo.com (I.D.M.); ecaterina.popa@univ-danubius.ro (E.P.)
[2] Department of Computer Science and Information Technology, Faculty of Automation, Computers, Electrical Engineering and Electronics, "Dunarea de Jos" University of Galati, 800146 Galati, Romania
[3] The Modelling & Simulation Laboratory, "Dunarea de Jos" University of Galati, 47 Domneasca Str., 800008 Galati, Romania
* Correspondence: simona.moldovanu@ugal.ro

Abstract

Background: The early identification of individuals with autism spectrum disorder (ASD) is crucial for their proper integration into the educational system and society. AI methods introduce novel approaches for the detection and classification of individuals diagnosed with ASD. Methods: We employed three custom-built Convolutional Neural Networks (CNNs) alongside two pretrained CNNs, specifically YOLO8 and ResNet18. The integrated Convolutional Block Attention Module (CBAM) was utilized to enhance feature representations for classifying individuals with ASD and non-ASD. Results: The results from the binary classification using the YOLO8-CBAM model demonstrated notable performance metrics: an accuracy of 77.6%, an F1-score of 72.6%, a Matthews Correlation Coefficient (MCC) of 60%, and an area under the curve (AUC) of 0.912. Conclusion: The backbone of the pretrained YOLO8-CBAM, enhanced by the integration of the CBAM after selected convolutional blocks, improved the feature refinement utilized in the classification process. Additionally, the gradient-weighted Class Activation Mapping (Grad-CAM) model provides interpretability by highlighting the regions that are most influential in distinguishing between individuals with ASD and those without.

Keywords: convolutional neural networks; autism; Convolutional Block Attention Module

Check for updates

Academic Editor: Savvas A. Chatzichristofis

Received: 14 January 2026
Revised: 12 March 2026
Accepted: 24 April 2026
Published: 5 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Recent surveillance data from the Autism and Developmental Disabilities Monitoring (ADDM) Network indicate a continuous increase in ASD prevalence among children in the United States. Shaw et al. (2025) determined the highest estimate of 32.2 per 1000 (3.22%) ASD prevalence among 8-year-olds in 2022 by utilizing population-based statistics from 16 monitoring sites. In 2022, early identification improved significantly. By 48 months, children were more likely to receive an ASD diagnosis than children born in earlier years. These results indicate that enhanced screening protocols, increased awareness, and improved access to diagnostic services have facilitated earlier detection; however, prevalence and age at identification differ across regions, races, ethnicities, and levels of intelligence.

Early diagnosis in childhood is important, as it can enhance the social skills and communication abilities of children with ASD, ultimately improving their quality of life.

Identifying autism is difficult before the age of about 12 months, but diagnosis is generally possible by the age of 2 years (Sherkatghanad et al., 2020). Autism spectrum disorder (ASD) is a neurodevelopmental condition characterized by persistent challenges in social communication and interaction, as well as restricted and repetitive behaviours. Early diagnosis is critical for enabling timely intervention and improving long-term outcomes (Okoye et al., 2023), helping to develop language and social interaction (Oesch, 2024).

Educational and training interventions for autism spectrum disorder (ASD) are diverse and often effective; however, they lack a unified classification and well-defined criteria that align them with the specific educational needs of individuals with ASD and their parents (Morsa et al., 2022). Differentiating ASD cases using deep learning models remains a significant challenge. Convolutional neural networks (CNNs) have demonstrated strong performance across a range of face detection applications. Applying CNNs to image datasets of individuals diagnosed with ASD may extend their utility and improve classification accuracy.

Regarding the use of AI methods in research into ASD, most prior studies have relied on supervised learning techniques such machine learning classifiers (Heinsfeld et al., 2018), deep learning (DL) (Sherkatghanad et al., 2020), graph-based methods (Yu et al., 2026), multi-level feature fusion networks (EMLF-Net) (Shi et al., 2026), the CNN–Genetic Algorithm (Ramani et al., 2025), multimodal AI (Pang et al., 2026), the Dynamic Functional Graph Metrics Stacked Convolutional Neural Network (DFGM-SCNN) (Smitha et al., 2026), the Fractional Hiking Optimization Algorithm combined with Dense WideResNet (FHOA + DenWide-Net) (Arunkumar & Ilango, 2026), a pretrained CNN (S. Das et al., 2026), deep learning-based computer-aided diagnosis (Wen et al., 2026), and NeuroNet57 (Ashraf et al., 2025).

Sherkatghanad et al. (2020) achieved 70.22% accuracy in ASD identification using speech recognition and image classification, with deep learning methods. Yu et al. (2026) proposed an adaptive adversarial graph learning framework for ASD identification from Functional Magnetic Resonance Imaging (fMRI) data; this study improves discriminative brain connectivity representations by jointly using contrastive learning, adversarial training, and cross-entropy supervision, leading to better classification performance and more meaningful neurobiological biomarkers than existing graph-based methods.

In the paper published by Shi et al. (2026), an EMLF-Net was introduced. The design includes a two-part electroencephalogram (EEG) network that combines time and frequency information using special blocks to keep both basic and detailed data. This model achieved high ASD recognition accuracy and enabled insight into brain channel importance.

Ramani et al. (2025) developed a hybrid convolutional neural network–genetic algorithm (CNN–GA) model for autism spectrum disorder (ASD) diagnosis, in which CNNs were used to extract complex patterns from high-dimensional data and genetic algorithms optimized the network architecture and hyperparameters. The results demonstrated superior classification accuracy (97%) compared with other evolutionary and traditional methods, indicating that the proposed approach is a promising tool for early ASD detection.

Pang et al. (2026) proposed a review summarizing recent studies on pediatric ASD that utilize multimodal fusion strategies. Their findings indicate that integrating multiple data sources enhances classification, severity estimation, and subtyping, while advancing precision medicine and personalized intervention.

Smitha et al. (2026) published a dynamic functional graph metric–stacked convolutional neural network (DFGM-SCNN) model that integrates dynamic functional connectivity metrics derived from functional magnetic resonance imaging (fMRI). Using multiscale sliding window-based time–frequency dynamic functional connectivity (MSW-TFDFC), the model captured temporal and spatial patterns of brain activity. The proposed methodology

demonstrated high accuracy and robustness in classifying neurological disorders, including epilepsy, stroke, and autism spectrum disorder (ASD).

In a recent study, Arunkumar and Ilango (2026) applied a deep learning-based model (FHOA + DenWide-Net) for ASD detection from brain MRI, where FHOA selects optimal regions of interest and DenWide-Net classifies them, achieving improved accuracy, sensitivity, and specificity over existing methods.

S. Das et al. (2026) presented a DL framework using facial images for early ASD detection, leveraging EfficientNet, transfer learning (TL), and Grad-CAM interpretability. This proposal achieved 82% accuracy and an AUC of 0.88; these methods demonstrate the potential of AI for scalable, accurate, and explainable ASD screening.

Wen et al. (2026) proposed deep learning-based computer-aided diagnosis, a pseudo-label consistency-driven framework that reconstructs and refines functional connectivity features from fMRI data. The authors used temporal and spatial pseudo-labelling, combined with a Feature Optimization Autoencoder and a Depthwise Separable CNN, to improve the discriminative power and classification accuracy for neuropsychiatric disorders such as ASD, Attention Deficit Hyperactivity Disorder (ADHD), and Major Depressive Disorder (MDD) across multi-site datasets.

Ashraf et al. (2025) involved NeuroNet57, a 57-layer CNN combined with Ant Colony Optimization (ACO) for feature selection, to identify and classify autistic females from fMRI data. The results depicted high accuracy (92–93%) and demonstrate the effectiveness of deep learning for gender-specific ASD diagnosis.

In previous studies, CNNs were used; these produce deep features that are classified by SoftMax. To enhance salient deep features and effectively attenuate non-informative components, a Convolutional Block Attention Module (CBAM) is incorporated into the CNN layers. The CBAM is composed of two sequential sub-modules—channel attention followed by spatial attention—which jointly enable adaptive feature recalibration. This process is performed by assigning higher weights to informative features, with the CBAM improving the discriminative power and representational capability of the network (Woo et al., 2018).

Despite advances in neuroimaging and deep learning models, accurate ASD diagnosis remains challenging due to the disorder's heterogeneity and the limitations of traditional behaviour-based assessments. This work addresses these challenges by leveraging facial analysis with CNNs and developing a more reliable method for classification, contributing to improved diagnostic accuracy and understanding of ASD's underlying characteristics. In this sense, a rigorous selection of pretrained CNNs used in this paper was done, finding on Google Scholar the frequency of occurrence of papers in accordance with the keywords "autism" and "CNN", "autism" and "CBAM", "autism" and "ResNet", "autism" and "YOLO". We identified the salient pre-trained CNNs (ResNet and YOLO) after the initial search. Figure 1 depicts the number of papers identified that were published in the past five years.

The study has yielded the following contributions:

Advanced Preprocessing Technique: We utilized a Multi-task Cascaded Convolutional Neural Network (MTCNN) deep learning model for face detection and alignment. Additionally, data augmentation techniques, including horizontal flipping and rotation, increased the size of the dataset.

Models Development and Application: We implemented three custom-built CNN-CBAM models and utilized two pre-trained CNN-CBAM models. In each CNN, the CBAM was integrated after a convolutional layer to reweight the feature maps through channel attention, followed by spatial attention, before proceeding to the next stage.

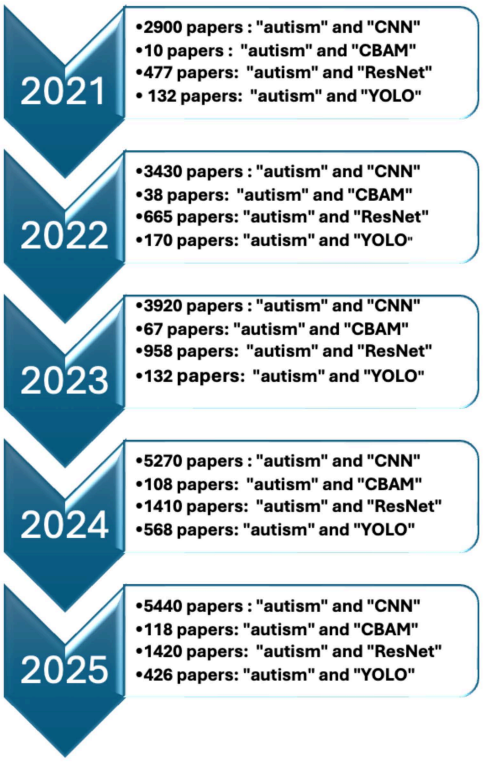


Figure 1. The frequency of papers that align with the specified keywords.

The comparative analysis involved evaluating the performance of all CNN-CBAM models based on metrics such as accuracy, F1-score, AUC, and MCC.

Interpretability with Grad-CAM: We utilized Grad-CAM visualization to improve interpretability by emphasizing key features that differentiate autistic from non-autistic subjects, thereby providing important knowledge about the classification process.

The rest of this paper is structured as follows: Section 2 discusses the framework of the methodology. Section 3 presents the material and methods in detail. Section 4 presents the results and discussions; the performance of the proposed deep learning models is compared with findings of contemporary research. Finally, Section 5 concludes the paper with the contributions of this research and future work.

2. Methodology

In this work we propose a new computational model based on CNN-CBAM to classify individuals with Autism Spectrum Disorder (ASD). First, we utilized MTCNN to detect and align the original images. Next, the pre-processed images were fed into five CNN-CBAM models. In this approach, the standard CNN block was replaced with three custom-built CNNs, along with pretrained models YOLO8 and ResNet18.

To overcome the prediction and processing challenges associated with ASD, we utilized a large dataset and train deep learning models to achieve optimal performance.

The deep learning models were built around a CBAM block that integrates both channel and spatial attention mechanisms, enabling the network to focus on the most informative facial regions.

The model-interpretability technique utilizing Grad-CAM (Gradient-weighted Class Activation Mapping) was applied to understand which areas of the input are emphasized by the convolutional neural network during its prediction process.

The evaluation and comparison of model performance were conducted using fivefold cross-validation, with qualitative metrics such as accuracy, F1-score, MCC, and the area

under the receiver operating characteristic curve (ROC-AUC). A flowchart illustrating this process is presented in Figure 2.

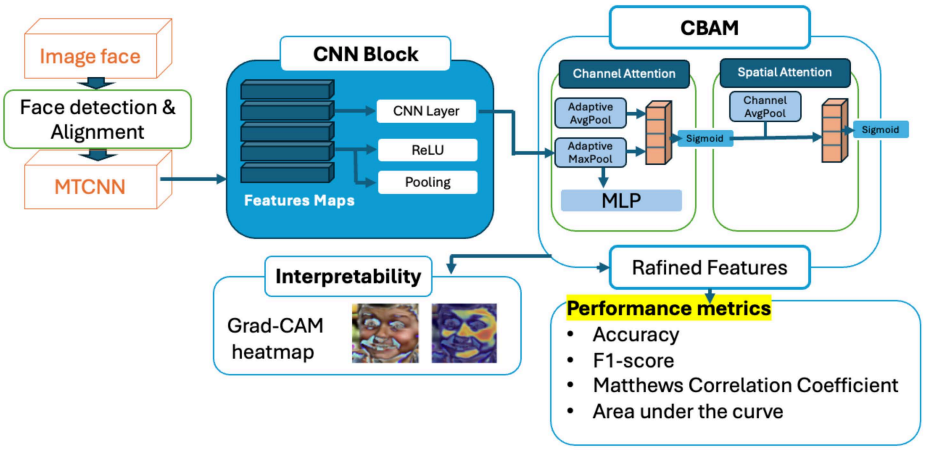


Figure 2. The framework of the proposed CNN-CBAM model.

3. Materials and Methods

3.1. Image Dataset

A large dataset is required to train CNN-CBAM models for optimal performance. We utilized the autistic children dataset from the Zenodo repository (Piosenka, 2021) to develop our proposed models (https://zenodo.org/records/14890016, accessed on 18 December 2025). This dataset contains 2D RGB images of ASD children aged 2 to 14 years, mostly between 2 and 8 years old. Each group maintains an equal ratio of ASD to non-ASD classes. The male-to-female ratio is about 3:1, while the autistic and control groups are roughly equal at 1:1. The images are divided into three groups: a training set, a testing set, and a validation set, comprising 2536, 300 and 100 images, respectively.

In training process, we included the random horizontal flip and random rotation with 15 degrees as data augmentation.

3.2. Hardware, Software and Experimental Settings

All research tasks took place on a PC equipped with the Apple Mac Studio (2023) architecture, with an Apple M2 Ultra CPU, 64 GB of RAM, a 1 TB SSD, and a 32-core GPU.

The CNN-CBAM models were designed in Python 3.10.19 with the following primary libraries: Pandas 2.3.3, Torch 2.2.2, Scikit-learn 1.7.2, Pillow 10.2.0, Matplotlib 3.10.8, Seaborn 0.13.2, TensorFlow 2.19.1, and Tqdm 4.67.1.

The training procedure's hyperparameters were established with a batch size of 16, a learning rate of 0.0001, and a total of 30 epochs, image size 324 × 324. The outcomes were evaluated using five-fold cross-validation to ensure the precision and consistency of the classifications.

3.3. Deep Learning Models

3.3.1. Multi-Task Cascaded Convolutional Neural Network

The MTCNN contains three blocks, the Proposal Network (P-Net), Refine Network (R-Net), and Output Network (O-Net), and is proposed to improve face detection and landmark localization (K. Zhang, 2016).

The P-Net block is a fully convolutional net on a multiscale input image pyramid. It must swiftly generate face bounding boxes of numerous candidates and a rough facial landmark assessment. It gives candidates confidence scores and combines highly overlapping detections using Non-Maximum Suppression (NMS).

The R-Net block enhanced CNN processes each P-Net candidate proposal at a predetermined size. R-Net filters out most P-Net candidate false positives and refines bounding box coordinates through bounding box regression.

The O-Net block calculates the final bounding box, confidence measure, and five face landmarks (the four corners of both eyes and the nose tip). For normalizing input to the FaceNet model, these coordinates are utilized to position and crop the face using a similarity transformation (Karamizadeh et al., 2025).

3.3.2. Pretrained CNN

YOLOv8 is a cutting-edge real-time object detection model, distinguished by its improved detection robustness, all while maintaining a high processing speed. This is accomplished through an improved network architecture, an optimized training strategy, and more effective feature extraction techniques. These include multi-scale prediction and adaptive anchor box optimization, which facilitate the precise detection of objects across a range of sizes. The model's architecture is structured into a backbone, neck, and head; the backbone, in particular, reduces the number of parameters while simultaneously enhancing feature efficiency when compared to previous YOLO iterations. The backbone incorporates convolutional layers, residual connections, and a Spatial Pyramid Pooling-Fast (SPPF) module, thereby enabling the capture of both low-level details and high-level semantic information (J. Zhang et al., 2025).

The ResNet18 architecture was selected for the second pretraining phase due to its residual (skip-connection) backbone network design, which effectively addresses the issue of vanishing gradients while simultaneously retaining crucial input data without incurring additional computational expense. ResNet18, constructed from stacked Basic Blocks, offers a hierarchical feature extraction mechanism that strikes a balance between comprehensive representation learning and operational efficiency. Consequently, it is especially well-suited for multimodal sheep face recognition applications; its residual blocks facilitate the effective integration of RGB and depth features, thereby enhancing robustness, recognition accuracy, and overall computational efficiency (Liao et al., 2025).

3.3.3. Convolutional Block Attention Module

The Convolutional Block Attention Module (CBAM) improves feature representation through the sequential application of channel and spatial attention mechanisms to an input feature map. Initially, channel-wise features are reweighted to prioritize the most informative channels. Subsequently, spatial locations are reweighted to concentrate on significant regions within the feature map. This two-stage attention process enables the network to selectively emphasize pertinent information across both channel and spatial dimensions, thereby enhancing overall feature discrimination.

Assume the convolutional layer outputs a F feature map, with H number of rows, W number of columns, and C number of channels:

$$F \in \mathbb{R}^{C \times H \times W}. \tag{1}$$

The channel attention for deep learning is important. The convolution output F to generate the channel attention map (M_c) can be written as

$$F' = M_c(F) \otimes F. \tag{2}$$

Spatial attention determines where to focus within the image. For channel attention, this involves applying average and max pooling across channels, followed by a 7×7

convolution to generate a spatial attention map (M_s), which is defined as follows (Woo et al., 2018):

$$F'' = M_s(F') \otimes F'. \tag{3}$$

3.3.4. Gradient-Weighted Class Activation Mapping

The Grad-CAM is a post hoc explainability method for CNNs. It visualizes which regions of an input image contribute most to a specific class prediction, producing a class-discriminative heatmap. CNNs learn spatial feature hierarchies, but their decision process is not directly interpretable. Grad-CAM utilizes two key components: gradients, which indicate how sensitive the output is to changes in feature map gradients, and feature maps, which convey spatial semantic information. The importance weights (α_k^c) depend on the score for the class (y^c) and the feature map from the last convolutional layer (A^k). For Z number of pixels in the feature map, α_k^c is given by

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}. \tag{4}$$

The feature map $L^c_{\text{Grad-CAM}}$ needs to be calculated according to Equation (5), where the elements of the original ASD and non-ASD are scaled to [0, 1] by the ReLU function (Zhu et al., 2025):

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right). \tag{5}$$

3.3.5. Custom-Built CNNs

Typically, pretrained CNN models are utilized for classifying images. In this paper, we propose three custom-built architectures in addition to the YOLO8 and ResNet18 architectures for ASD and non-ASD classification. Each model incorporates a CBAM layer to enhance classification performance. Figure 3 illustrates the CNN models, along with the proposed layers, output shapes, and parameters.

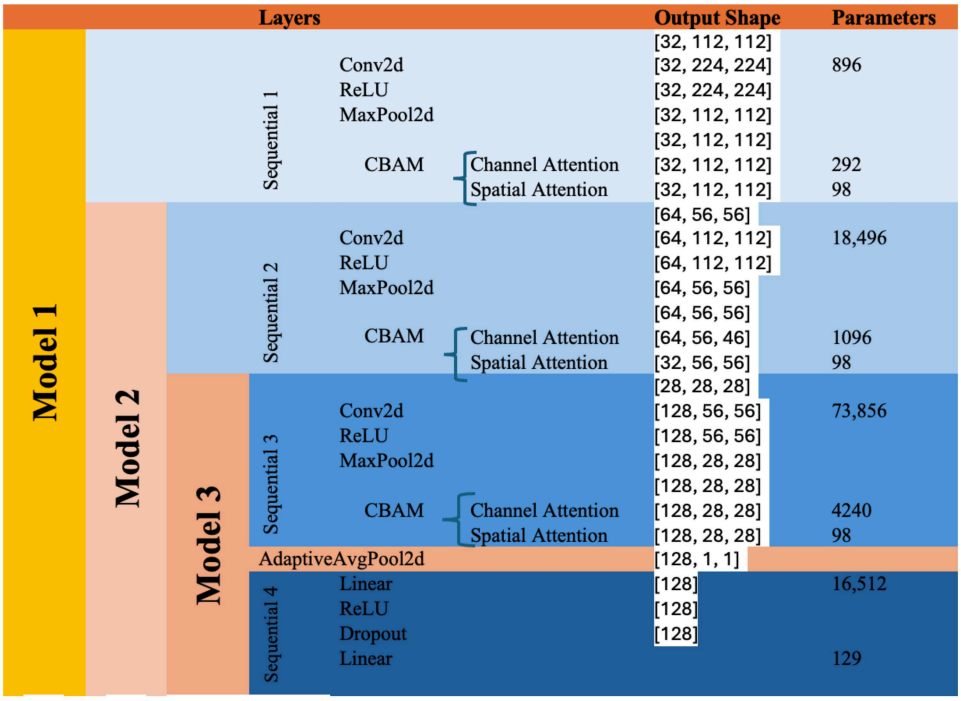


Figure 3. Architectures of custom-built CNNs.

Model 1: three Conv2D layers, three MaxPool2D layers, three CBAM layers, and one AdaptiveAvGPool2D layer.

Model 2: two Conv2D layers, two MaxPool2D layers, two CBAM layers, and one AdaptiveAvGPool2D layer.

Model 3: one Conv2D layer, one MaxPool2D, one CBAM layer, and one AdaptiveAvG-Pool2D layer.

4. Results and Discussions

The initial strategy for this research focused on enhancing the classification of face features related to ASD by utilizing MTCNN for face detection and image recalibration. We processed an improved dataset through various image-processing techniques, including resizing and augmentation, before feeding it into pre-trained models. The incorporation of CBAM layers in these models enabled the extraction of crucial facial characteristics from the images. To achieve this, we developed three custom-built convolutional neural networks (CNNs) and proposed three additional pre-trained models.

The deep learning model performance was evaluated using a confusion matrix, which is central to understanding how well it classifies images. The confusion matrix categorizes all test images into four groups: true positives (TPs) represent correctly identified positive cases, true negatives (TNs) represent correctly identified negative cases, false positives (FPs) are negative cases that are incorrectly labelled as positive, and false negatives (FNs) are positive cases that are incorrectly labelled as negative. These values form the basis for calculating various performance metrics such as accuracy, F1-score, MCC, AUC, False Positive Rate (FPR), and False Negative Rate (FNR). Each metric provides a distinct perspective on the system's classification ability, allowing researchers to analyze strengths and weaknesses and compare the effectiveness of different systems. The evaluation equations referenced in the study use combinations of TP, TN, FP, and FN to derive these metrics.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

$$\text{F1-Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{9}$$

$$FPR = \frac{FP}{FP + TN} \tag{10}$$

$$FNR = \frac{FN}{TP + FN} \tag{11}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{12}$$

$$AUC = \frac{TP\ Rate}{FP\ Rate} \tag{13}$$

Table 1 presents the findings of this study, which explored custom-built CNNs—CBAM, YOLO8-CBAM, and ResNet18-CBAM—to classify facial feature images from the ASD dataset. The accuracy, F1-score, MCC, and AUC quality metrics are calculated for each model. The evolution of these metrics is illustrated in Figure 4, which includes a radar plot that highlights the performance of each model.

Table 1. The quality metrics computed for each model.

Model	Accuracy	F1-Score	MCC	FNR	FPR	AUC
Model 1	0.683	0.692	0.367	0.304	0.327	0.753
Model 2	0.706	0.639	0.445	0.349	0.170	0.793
Model 3	0.746	0.714	0.506	0.298	0.181	0.841
Yolo8-CABM	0.776	0.726	0.6	0.297	0.063	0.912
ResNet18-CBAM	0.773	0.792	0.556	0.163	0.269	0.878

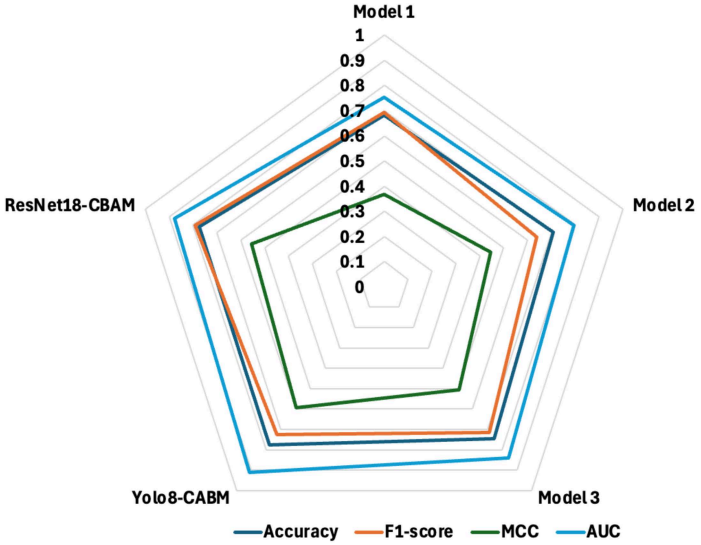
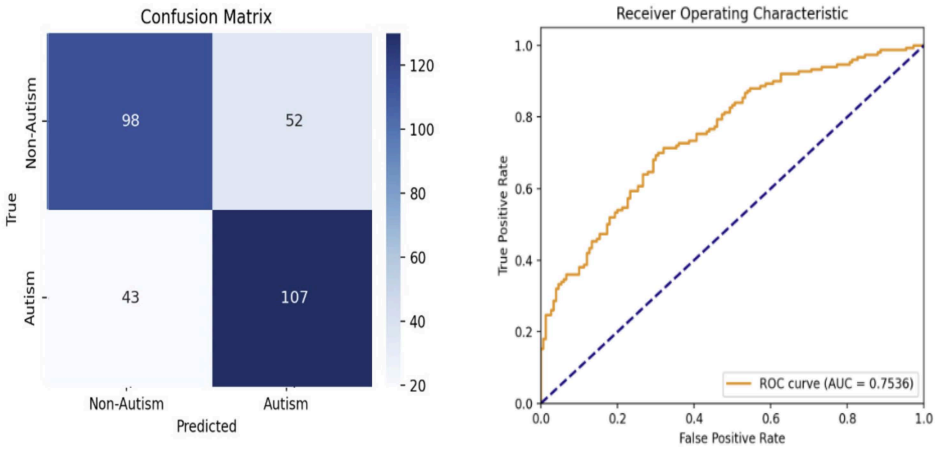


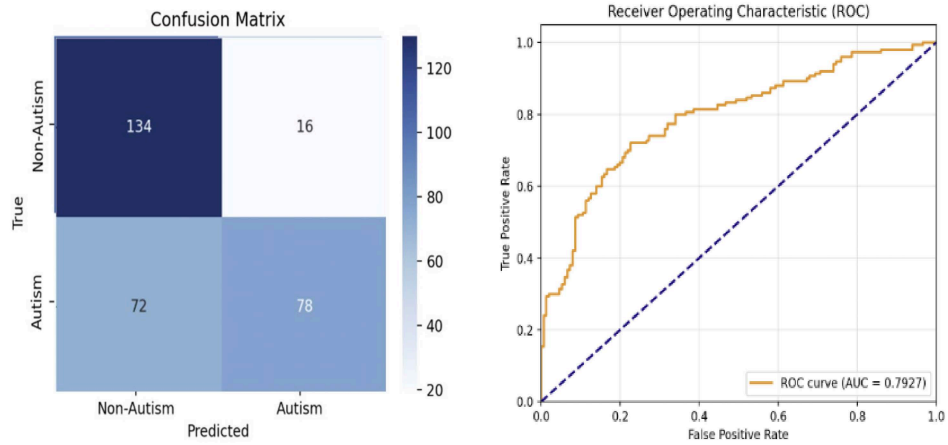
Figure 4. Evolution of proposed models in accord with quality metrics.

The results are represented in the confusion matrices (left column) and receiver operating characteristic (right column) depicted for each model in Figure 5. These visualizations provide a clear overview of the models' performance, highlighting areas of strength and potential improvement. By analyzing the metrics presented, we can get helpful information about the predictive capabilities of each model.

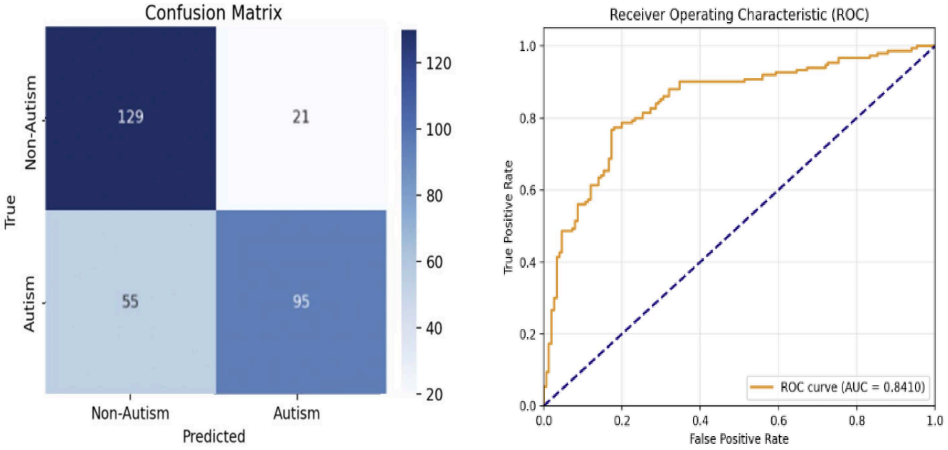


(**a1**) (**a2**)

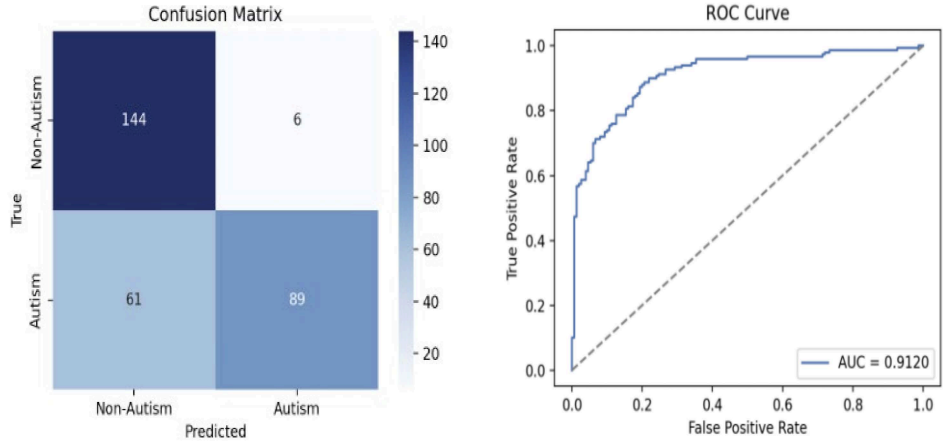
Figure 5. *Cont.*



(**b1**) (**b2**)

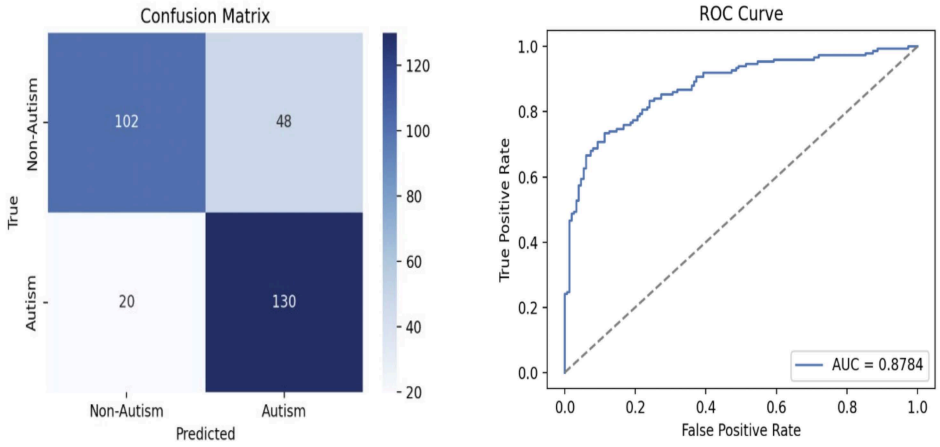


(**c1**) (**c2**)



(**d1**) (**d2**)

Figure 5. *Cont.*



(**e1**) (**e2**)

Figure 5. Confusion matrices (CMs) and ROCs (ROC); CMs: (**a1**) Model 1; (**b1**) Model 2; (**c1**) Model 3; (**d1**) ResNet18-CBAM; (**e1**) Yolo8-CABM; ROC: (**a2**) Model 1; (**b2**) Model 2; (**c2**) Model 3; (**d2**) ResNet18-CBAM; (**e2**) Yolo8-CABM.

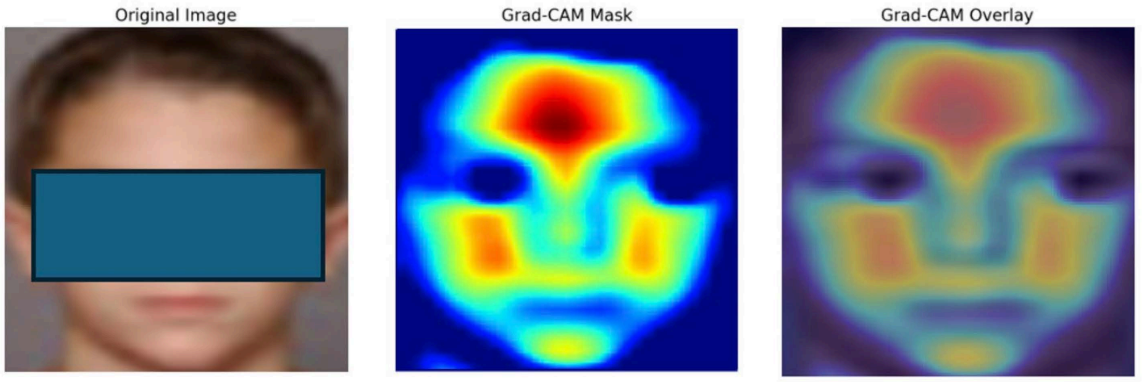
First, we discuss the results of combining CNNs with CBAM. Model 3, which includes a single Conv2D layer followed by a CBAM layer, achieved an accuracy of 74.6% and an AUC of 0.841. However, as the number of layers increases, the classification quality tends to decrease. The increasing number of layers in custom-built CNNs reflects difficulty in adapting to the dataset.

The results for YOLO8-CBAM and ResNet18-CBAM show promising performance in the classification of individuals with ASD and non-ASD. Specifically, YOLO8-CBAM achieved an accuracy of 77.6%, an AUC of 0.912, an FNR of 0.297, and an FPR of 0.063; this means a sensitivity of 70.03% and a specificity of 93.7%, respectively. The Res-Net18-CBAM model provided an accuracy of 77.3%, an AUC of 0.912, an FNR of 0.163, and an FPR of 0.063; this means a sensitivity of 83.7% and a specificity of 73.1%, respectively.

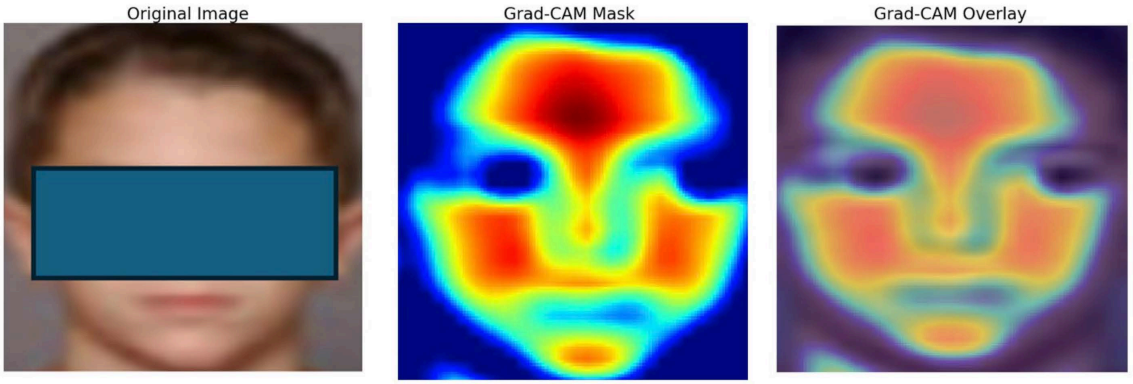
The models that used pretrained CNNs were able to effectively distinguish the different classes, even though the quality measurements were asymmetric. Additionally, the AUC indicates how effectively the model differentiates between ASD and non-ASD across all thresholds. The Yolo8-CABM, with an AUC exceeding 0.9, demonstrates the model's ability to accurately distinguish between positive and negative cases. These results suggest that the Yolo8-CABM can capture non-random structures within the data.

Grad-CAM was used to create attention maps for autism and non-autism facial images, providing qualitative insights into feature localization. We evaluated three backbone models: Model 1, Model 2, Model 3, YOLO8-CBAM, and ResNet18-CBAM. Heatmaps demonstrate changes in attention patterns between models and classes, highlighting visual regions with the greatest influence on model predictions.

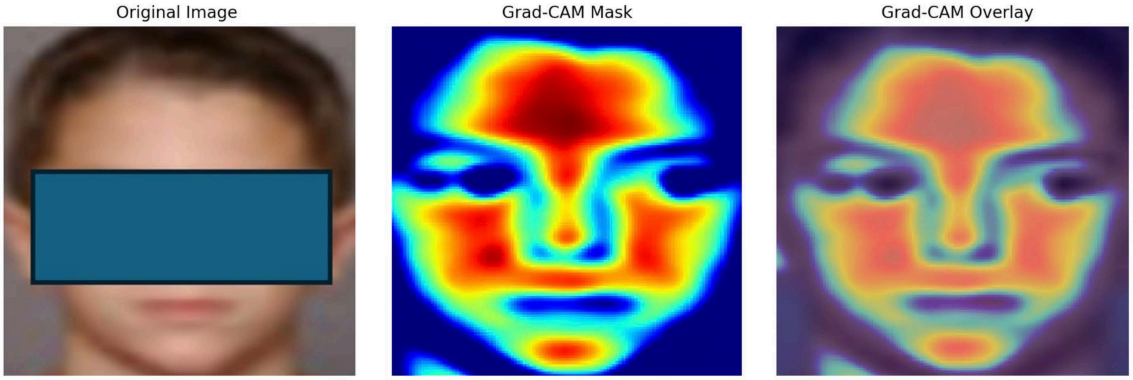
The Grad-CAM was applied after MTCNN highlighted and aligned the face. Model 1, Model 2, and Model 3 mostly focused on a spread area, the cheeks, nose, chin, and forehead; the red colour highlights these facial parts. Figure 6 expresses the original image, the Grad-CAM mask, and the overlay between the first two.



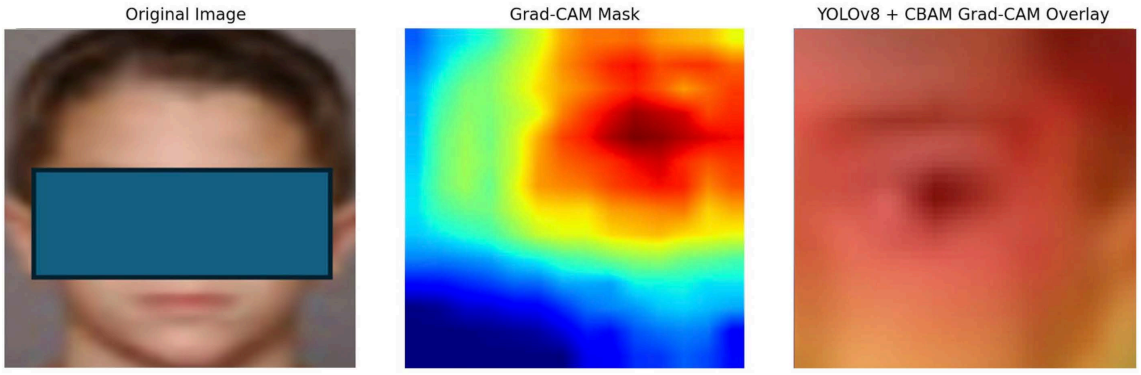
(**a**)



(**b**)

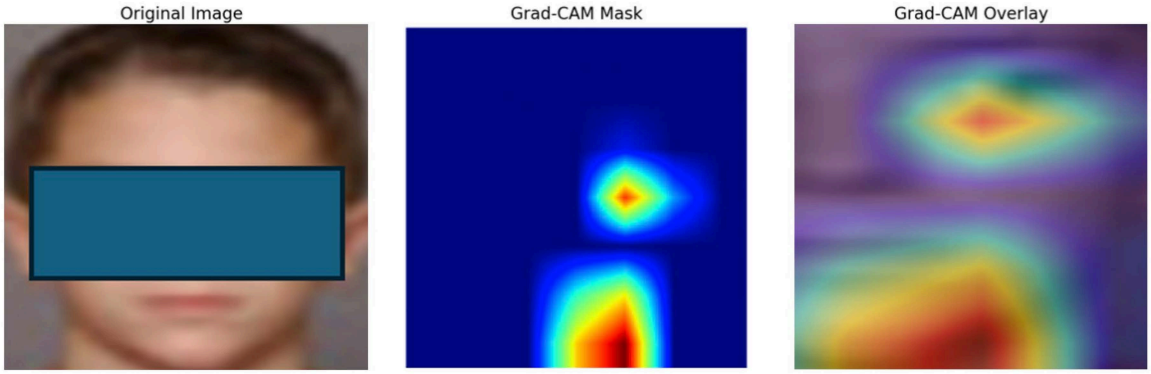


(**c**)



(**d**)

Figure 6. *Cont.*



(**e**)

Figure 6. Original image, Grad-CAM mask and overlay: (**a**) Model 1; (**b**) Model 2; (**c**) Model 3; (**d**) Yolo8-CABM; (**e**) ResNet18-CBAM. The image belongs to the National Institute of Mental Health Child Emotional Faces Picture Set (NIMH-ChEFS).

Pretrained YOLO8-CBAM focuses on eyes and mouth, while ResNet18-CBAM focuses only on the mouth area.

The results obtained were compared with those from state-of-the-art papers, which included details such as the reference, year of publication, first author, deep learning model used, and the quality metric applied. All mentioned elements are stored in Table 2. This comprehensive analysis highlighted the advancements made in the field and identified areas where further research could enhance performance.

Table 2. Comparisons with state-of-the-art-methods.

Reference	Deep Learning Model	Metrics
M. Das et al. (2025)	EfficientNet	accuracy of 84%, AUC-ROC of 0.89
Negin et al. (2021)	LSTM, ConvLSTM, and 3DCNN	accuracy of 72%
Khalaji et al. (2022)	ML classifier	accuracy of 75%
Liu et al. (2025)	3D-CNN + Recurrent neural network + Random Forest	accuracy of 77%
Proposed method	YOLO8-CBAM	accuracy of 77.6% AUC of 0.912

This study is limited by its single dataset, which limits generalizability. A single dataset may train a model to learn dataset-specific patterns rather than robust features that work across data distributions. Thus, the model's performance may not reflect its efficacy when applied to unknown data or real-world settings. In the absence of a well-structured dataset, the generalization of models can be obtained under the following conditions: (i) applying recent advances in computer vision, such as transformer-based models, can improve facial feature detection, particularly in the eyes, mouth, and mid-face regions; (ii) using computer vision methods like adjusting shapes, changing angles, and aligning key facial points can help AI models handle differences in head position, size, and facial expressions, which improves their detection accuracy; and (iii) clinical psychologists specializing in early autism detection can assist in the collection of a new dataset for testing the YOLOV8-CBAM model. This collaboration will ensure that the dataset comprehensive and accurately represents the various indicators of autism, facilitating effective model training. By incorporating insights from experts, our goal is to enhance the model's ability to identify early signs of autism in children.

The classification method relies on a CNN, which is both computationally expensive and time-consuming. The use of many convolutional layers, a lot of parameters, and the

need for repeated training mean that CNN-based models require a lot of computing power. This prolongs training and inference times, especially when handling high-resolution images or large datasets. This poses challenges for deploying models in real-time or resource-constrained environments.

5. Conclusions

This article explored the performance of three custom-built CNNs and two pretrained CNN models for ASD prediction. We focused mainly on the integration of the CBAM block in all models. The performance was evaluated by metrics such as accuracy, F1-score, AUC, and MCC. After analyzing the developed models, it is evident that the YOLO8_CBAM model was the optimum performer, producing an accuracy of 77.6%, an F1-score of 72.6%, an MCC of 60%, and an AUC of 0.912. Corroborating the results provided by deep learning models and Grad-CAM, the salient classification of ASD is obtained when the CNNs focus on the eyes and mouth. In contrast, the M3 model, custom-built with a single convolutional layer and CBAM, struggled to adapt to the dataset and produced an accuracy of 74.6%.

The limitations suggest that future research should use numerous different datasets to improve model robustness and computationally efficient architectures or optimization methodologies to reduce processing time and classification accuracy.

Author Contributions: Conceptualization, I.D.M. and S.M. Methodology, E.P.; Software, S.M. Validation, I.D.M.; Formal analysis, E.P. Investigation, I.D.M.; Resources, S.M.; Data curation, I.D.M.; Writing—original draft, E.P.; Writing—review & editing, S.M.; Visualization, E.P.; Supervision, S.M.; Project administration, S.M.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in classification process is visible at https://zenodo.org/records/14890016, accessed on 18 December 2025. The processing of images was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Center for Developmental Epidemiology (CDE), Duke University Medical Center (Approval No. R01 MH081025), https://psychiatry.duke.edu/research/research-programs-areas/assessment-intervention/developmental-epidemiology-instruments-6 (accessed on 18 December 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ASD	Autism Spectrum Disorder
CNN	Convolutional Neural Networks
CBAM	Convolutional Block Attention Module
Grad-CAM	Class Activation Mapping
DFGM-SCNN	Dynamic Functional Graph Metrics Stacked Convolutional Neural Network
FHOA	Fractional Hiking Optimization Algorithm
EMLF-Net	Multi-Level Feature Fusion Network
DL	Deep Learning
EEG	Electroencephalogram
GA	Genetic Algorithm
fMRI	Functional Magnetic Resonance Imaging

AUC	Area Under the Curve
ROC	Receiver Operating Characteristic Curve
RGB	Red Green Blue
MTCNN	Multi-task Cascaded Convolutional Neural Network
NMS	Non-Maximum Suppression
ACO	Ant Colony Optimization
MDD	Major Depressive Disorder
SPPF	Spatial Pyramid Pooling-Fast
ADHD	Attention Deficit Hyperactivity Disorder

References

Arunkumar, A., & Ilango, S. S. (2026). DenWide-Net: Dense WideresNet for autism spectrum disorder detection using MRI image. *Biomedical Signal Processing and Control*, *113*(A), 108976. [CrossRef]

Ashraf, A., Zhao, Q., Bangyal, W. H., Raza, M., & Iqbal, M. (2025). Female autism categorization using CNN based NeuroNet57 and ant colony optimization. *Computers in Biology and Medicine*, *189*, 109926. [CrossRef]

Das, M., Halder, A., & Mahadevappa, M. (2025, July 14–18). *Deep-fusion of scalogram and spatio-temporal EEG features with Attention mechanism for autism spectrum disorder identification* [Conference session]. 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Copenhagen, Denmark.

Das, S., Biswas, S., Sharma, A., & Vijay, R. V. (2026). Leveraging transfer learning with efficient net for early detection of autism spectrum disorder using facial imagery. In *Artificial intelligence and sustainable innovation*. CRC Press.

Heinsfeld, A. S., Franco, A. R., Craddock, R. C., Buchweitz, A., & Meneguzzi, F. (2018). Identification of autism spectrum disorder using deep learning and the abide dataset. *NeuroImage Clinical*, *17*, 16–23. [CrossRef]

Karamizadeh, S., Chaeikar, S. S., & Salarian, H. (2025). Combining MTCNN and enhanced FaceNet with adaptive feature fusion for robust face recognition. *Technologies*, *13*(10), 450. [CrossRef]

Khalaji, E., Eraslan, S., Yesilada, Y., & Yaneva, V. (2022). Effects of data preprocessing on detecting autism in adults using web-based eye-tracking data. *Behaviour & Information Technology*, *42*, 2476–2484.

Liao, S., Shu, Y., Tian, F., Zhou, Y., Li, G., Zhang, C., Yao, C., Wang, Z., & Che, L. (2025). Multi-modality sheep face recognition based on deep learning. *Animals*, *15*(8), 1111. [CrossRef] [PubMed]

Liu, Z., Chen, Y., Dong, X., & Liu, J. (2025). An autism spectrum disorder identification method based on 3D-CNN and segmented temporal decision network. *Brain Sciences*, *15*(6), 569. [CrossRef]

Morsa, M., Andrade, V., Alcaraz, C., Tribonnière, X., Rattaz, C., & Baghdadli, A. (2022). A scoping review of education and training interventions in autism spectrum disorder. *Patient Education and Counselling*, *105*, 2850–2859. [CrossRef] [PubMed]

Negin, F., Ozyer, B., Agahian, S., Kacdioglu, S., & Ozyer, G. T. (2021). Vision-assisted recognition of stereotype behaviors for early diagnosis of Autism Spectrum Disorders. *Neurocomputing*, *446*, 145–155. [CrossRef]

Oesch, N. (2024). Social brain perspectives on the social and evolutionary neuroscience of human language. *Brain Sciences*, *14*(2), 166. [CrossRef]

Okoye, C., Obialo-Ibeawuchi, C. M., Obajeun, O. A., Sarwar, S., Tawfik, C., Waleed, M. S., Wasim, A. U., Mohamoud, I., Afolayan, A. Y., & Mbaezue, R. N. (2023). Early diagnosis of autism spectrum disorder: A review and analysis of the risks and benefits. *Cureus*, *15*, e43226. [CrossRef]

Pang, L., Zhao, X., Zhao, L., Li, J., Kuo, F., Wang, H., & Liu, C. (2026). Multi-modal data analysis for autism spectrum disorder in children: State of the art and trends. *Eng Medicine*, *3*(1), 100117. [CrossRef]

Piosenka, G. (2021). *Detect autism from a facial image*. Available online: https://cutt.ly/ibIXt5a (accessed on 18 December 2025).

Ramani, D. R., Devadas, T. J., & Prasanth, N. (2025). Enhancing autism prediction using a hybrid approach of neural networks and evolutionary algorithms. In *Advanced computational intelligence techniques for engineers* (p. 14). Chapman and Hall/CRC.

Shaw, K. A., Williams, S., Patrick, M. E., Valencia-Prado, M., Durkin, M. S., Howerton, E. M., Ladd-Acosta, C. M., Pas, E. T., Bakian, A. V., Bartholomew, P., Nieves-Muñoz, N., Sidwell, K., Alford, A., Bilder, D. A., DiRienzo, M., Fitzgerald, R. T., Furnier, S. M., Hudson, A. E., Pokoski, O. M., . . . Maenner, M. J. (2025). Prevalence and early identification of autism spectrum disorder among children aged 4 and 8 years—Autism and developmental disabilities monitoring network, 16 sites, United States, 2022. *MMWR Surveillance Summaries*, *74*, 1–22. [CrossRef]

Sherkatghanad, Z., Akhondzadeh, M., Salari, S., Zomorodi-Moghadam, M., Abdar, M., Acharya, U. R., Khosrowabadi, R., & Salari, V. (2020). Automated detection of autism spectrum disorder using a convolutional neural network. *Frontiers in Neuroscience*, *13*, 1325. [CrossRef]

Shi, L., Ma, L., Zhao, J., Wang, H., Kuang, Z., Han, Q., & Sun, L. (2026). EMLF-Net: A multi-level feature fusion network based on autism EEG signals for autism spectrum disorder classification. *Biomedical Signal Processing and Control*, *113*(D), 109180. [CrossRef]

Smitha, C. S., Sathiya, S., & Soumya, T. (2026). Efficient dynamic graph metrics stacked convolutional neural network for classifying the neurological disorder. *Biomedical Signal Processing and Control*, *112*(B), 108620. [CrossRef]

Wen, X., Guo, S., Dong, L., Liu, X., Ning, W., & Shi, J. (2026). A consistency-driven pseudo-labeling framework for robust functional connectivity modeling in neuropsychiatric disorder diagnosis. *Engineering Applications of Artificial Intelligence*, *166*(A), 113556. [CrossRef]

Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018, September 8–14). *Cbam: Convolutional block attention module* [Conference session]. European Conference on Computer Vision (ECCV), Munich, Germany.

Yu, R., Zhu, J., Peng, C., Zhao, Z., Chen, M., & Zhang, R. (2026). Adversarial training based adaptive weighted hybrid network for autism spectrum disorder diagnosis. *Biomedical Signal Processing and Control*, *114*, 109313. [CrossRef]

Zhang, J., Guo, L., & Wang, X. (2025). Student classroom behavior recognition based on YOLOv8 and attention mechanism. *Information*, *16*(11), 934. [CrossRef]

Zhang, K. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, *23*, 1499–1503. [CrossRef]

Zhu, J., Zhang, Y., Ma, C., Wu, J., Wang, X., & Kong, D. (2025). GMCBAM-ResNet: A lightweight deep learning network for diagnosis of COVID-19. *Journal of Imaging*, *11*(3), 76. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.