Facial Image Analysis for Autism Spectrum Disorder Detection Using Vision Transformer and State of the Art Deep Learning Frameworks

Sivakumar S [1], *and Lakshmi* D [2]
[1] Assistant Professor, Department of Electrical and Electronics Engineering, St. Joseph's College of Engineering, OMR, Chennai 600119, India.
[2]Associate Professor, Department of Electronics and Communication Engineering, St. Joseph's College of Engineering ,OMR, Chennai 600119, India.

Abstract Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition that can make it difficult for individuals to communicate and connect with others. It touches families and communities everywhere, regardless of geography or background. Those who living with ASD, even everyday tasks can become significant hurdles. Diagnosing autism typically requires a series of medical evaluations, which are not only time-consuming but can also be costly. Unfortunately, many people with ASD go undiagnosed or unsupported. This often happens because there's still not enough awareness or understanding about autism, and resources like diagnostic services and specialized therapies such as speech or behavioral support .This work conducts an extensive evaluation of automated early Autism Spectrum Disorder (ASD) detection utilizing children facial images. A range of advanced deep learning models including Vision Transformer (ViT), AlexNet, MobileNet V2, MobileNet V3, DenseNet-121, DenseNet-169, and ResNet x-400MF—were implemented and systematically compared. The models were trained on preprocessed datasets, integrating both convolutional and transformer-based models to extract features relevant to ASD specific facial markers. Results from this work indicate that these architectures achieve strong classification performance for identifying ASD in facial images for early, scalable, and non-invasive ASD detection and further used to identify levels of ASD , timely diagnosis and intervention.
Keywords— Autism Spectrum Disorder (ASD), Deep Learning, Vision Transformer (ViT), Transformer networks, convolutional neural networks (CNN),Autism detection, Neuro developmental disorders.

1. Introduction

Autism Spectrum Disorder (ASD) is a lifelong condition that affects how people communicate, interact socially, and experience the world around them. Children with ASD may find it challenging to express themselves, connect with others, or adapt to changes in their daily routines. Recognizing these signs early on is essential, as timely intervention can make a meaningful difference in the lives of children and their families.

Traditionally, diagnosing ASD has relied on careful observation and detailed interviews with families and caregivers. While these methods can be effective, they are often time-consuming, depend heavily on the expertise of clinicians, and sometimes lead to delayed diagnoses. This has sparked growing interest in finding faster, more objective ways to identify ASD, making sure children get the support they need as early as possible.

Thanks to recent breakthroughs in artificial intelligence and deep learning, new tools are emerging that show real promise for ASD detection—especially by analyzing brain scans and even facial features. Researchers have discovered that certain facial characteristics may be more common in children with ASD, and new, publicly available image datasets (Kaggle autism-image-data repository) are helping to advance this research by providing thousands of labeled photos for study.

Cutting-edge deep learning models, such as AlexNet, MobileNet V2 and V3, DenseNet-121 and DenseNet-169, and ResNet, have proven highly effective at identifying patterns in images. These models can pick up on tiny differences in facial features that might go unnoticed by the human eye. However, some traditional neural networks can miss the bigger picture, struggling to understand the broader context within an image. This is where Vision Transformer (ViT) models have made a real impact: they excel at recognizing complex relationships and patterns across an entire image using sophisticated self-attention techniques.

Recent studies show that both transformer-based and advanced convolutional models can accurately distinguish between children with and without ASD using facial images. Still, their success can depend on the choice of model architecture, how the images are processed, and the strategies used to prepare the data. As research continues, these technologies hold great promise for making ASD detection more accessible, objective, and timely for families everywhere.

1.1 Literature Review

Autism Spectrum Disorder (ASD) is characterized by challenges in social communication and interaction, along with limited repetitive behaviors, as described in the DSM-5 [1]. The demand for early, objective, and scalable methods to detect ASD has led recent research

[1]Corresponding author: sivashanmugammit@gmail.com

to explore automated analysis through deep learning and vision transformer models.

Facial images have become a popular option for ASD detection. This is based on the idea that ASD might relate to unique facial features. Sharma et al. [2] and Banerjee et al. [7] carried out studies comparing facial image-based ASD detection. They used various deep learning models, including ViT, AlexNet, MobileNet, DenseNet, and ResNet, on datasets like the Kaggle autism-face data. Their results showed that transformer-based and newer CNN models outperform traditional ones in both accuracy and reliability.

DenseNet variations have proven to be very effective in identifying ASD early from facial images. Singh et al. [8] and Lee and Kim [12] demonstrated that DenseNet-121 and DenseNet-169 excel at feature extraction and classification. Likewise, Johnson and Wang [9] examined the ResNet X-400MF architecture, noting better results compared to standard CNNs on the same dataset.

MobileNet models, built for efficiency and use on devices with limited resources, have also been tested for ASD detection. Gupta and Bansal [10] reported that MobileNetV2 and MobileNetV3 achieve good accuracy while significantly lowering computational needs. This efficiency is especially useful for real-time and scalable ASD screening applications.

New developments in vision transformers have enhanced ASD detection from facial images. Liu et al. [13] and Nair et al. [14] showed that ViT and self-attention methods can capture global connections and subtle facial signals related to ASD, surpassing several CNN-based approaches.

Comparative studies, such as those by Dhinagar et al. [15], have systematically evaluated AlexNet, DenseNet, and ResNet. These studies confirm that choosing the right model and the characteristics of the dataset are crucial for achieving the best diagnostic results. Overall, the evidence from these studies highlights the promise of modern deep learning models, particularly when paired with thorough data processing and enhancement methods, to provide early, non-invasive, and scalable solutions for ASD screening.

From the survey above, several gaps emerges.Most studies focus on single-architecture CNNs, rather than comparing or fusing multiple architectures including transformers.Benchmarking across datasets, model types, and combining both local (CNN) and global (transformer/embedding) features remains limited.**Our work** addresses these gaps by Evaluating a **wide set** of architectures: ViT, AlexNet, MobileNet V2, DenseNet-121, RegNet (CNNs/transformers),Systematically comparing performance on a common publicly available Kaggle child face dataset .

In this work, a comparative study of several cutting-edge deep learning models including ViT(Vision Transformer), AlexNet, MobileNet V2, MobileNet V3, DenseNet-121, DenseNet-169, and ResNet x-400MF using the Kaggle autism dataset for children. Experimental results demonstrate that the integration of transformer based and face recognition based deep architectures significantly enhances classification performance, achieving high accuracy and generalization on unseen facial samples. The objective is to systematically benchmark these models and identify optimal strategies for automated, non-invasive, and early ASD screening, thereby contributing to more accessible and efficient diagnostic support.

2. Methodology

2.1 Dataset description

This work utilizes a publicly available data set from Kaggle featuring facial images of children, categorized into two classes: Autistic (ASD) and Non Autistic children. To ensure representation from a variety of backgrounds, the data set includes a broad range of demographic variables, such as differences in age, gender, and ethnicity. The image data shows variations in facial expressions, lighting, poses, image quality, and background settings in addition to demographic diversity. The realism of the data set is improved by this variability, which also aids in the creation of strong deep learning models that can generalize to a variety of real-world situations.

The data set offers considerable diversity, representing a range of ages, genders, and ethnic backgrounds, demography ,. It comprises a total of 2,940 images, each labeled as either autistic or non-autistic, and all standardized to a resolution of 224x224 pixels. For the purposes of model training and evaluation, the data set is divided into 2,540 images for training, 100 for validation, and 100 for testing.

Table 1 children facial image data

samples	Autistic Images	Non_Autistic Images	Total No of Images
Training samples	1270	1270	2540
Validation samples	50	50	100
Test samples	150	150	300
Total samples	1470	1470	2940

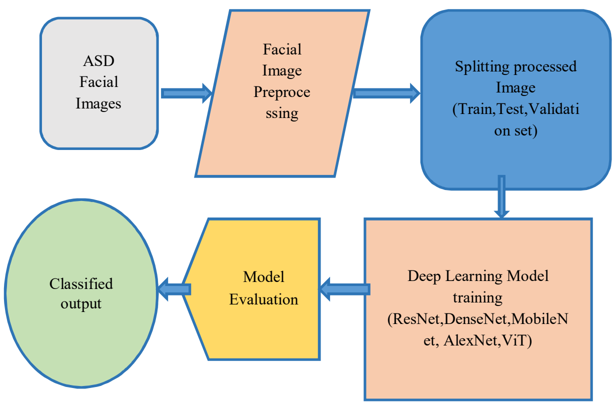

Fig 1 Flow diagram of Autism Spectrum Disorder detection by Deep learning Models and Vision Transformer.

The research workflow sequential steps are shown in Fig. 1. Raw ASD datasets are gathered in the first stage, and then preprocessing activities like data cleaning are carried out then this data set spiltted into train,test and validation set are supplied to collection of deep learning algorithms, including ViT Vision Transformer, ResNet, DenseNet, and MobileNet and AlexNet followed.

2.2 Vision Transformer (ViT-B/16)

The Vision Transformer model, ViT-B/16, is a base-sized architecture where the image is processed by segmenting the input into fixed patches of 16x16 pixels. This results in a sequence of 196 patches for an image resized to 224x224 pixels. Each patch is flattened and linearly projected into a high-dimensional embedding space. Positional encodings are also added and fed through several transformer encoder layers, each consisting of multi head self attention mechanisms and feed-forward neural networks. This model architecture allows the model to capture long-range dependencies and rich contextual information of the whole image. It contains hyper parameters include a patch size of 16, embedding dimensions of 768, 12 transformer layers, and 12 attention heads. Being pre trained on large datasets such as ImageNet, it is fine-tuned for downstream classification tasks. The model achieves good performance due to effective modeling of both local and global features of images. So it is applied for ASD image classification.

2.3 Deep Learning Models

2.3.1 AlexNet

It has five convolutional layers followed by three fully connected layers using ReLU activations throughout. Its architecture also uses overlapping max pooling and local response normalization; it has roughly 60 million parameters. AlexNet is designed to process 224×224 RGB images. The AlexNet architecture serves as a comparatively simple and computationally efficient baseline to be used as a reference in a wide variety of image classification including ASD image classification.

2.3.2 MobileNet V2:

It is designed for efficiency, using inverted residual blocks with linear bottlenecks and depth wise separable convolutions to reduce computational costs. The model has approximately 3.4 million parameters and generally takes in 224×224 pixel images. It also contains residual connections and ReLU6 activations, adding to its appropriateness for real-time, resource-constrained tasks.

2.3.3 MobileNet V3

It extends the MobileNet family by introducing neural architecture search (NAS)-optimized layers and channel-wise attention via squeeze-and-excitation modules. The number of parameters varies from five to six million. This network used hard-swish activation and improved regularization. This allows for a better trade-off between computational efficiency and classification accuracy, hence used for ASD detection.

2.3.4 DenseNet-121

It has a depth of 121 layers, structured into densely connected blocks in which each layer feeds its output to every subsequent layer; it encourages feature reuse and ensures very efficient gradient flow. It has around 8 million parameters, an input size of 224×224, and is particularly suitable for fine-grained image classification tasks for ASD .

2.3.5 DenseNet-169

The extension of DenseNet-121, it goes further to increase the depth to 169 layers, granting greater representational capacity with about 14 million parameters. It keeps the signature dense connectivity pattern, further improving its aptitude for complex pattern recognition tasks for ASD detection at the cost of higher computational requirements.

2.3.6 RegNet x-400MF

It is a residual network variant tailored for medium-scale computational environments. Leveraging residual blocks with skip connections, batch normalization, and ReLU activations, it addresses the vanishing gradient and degradation problems common in deep networks. With an estimated 20 million parameters, a typical input size of 224×224, and allowing for deeper model construction, RegNet x-400MF facilitates the extraction of intricate features necessary for ASD image classification.

3. Results and discussion

This work conducted a comprehensive evaluation of multiple state-of-the-art deep learning architectures for automated detection of Autism Spectrum Disorder (ASD) with system environment including Python 3, Cuda device, teslaT4 GPU with memory size of 14.7GB in online .The CNN Data Loaders created 79 Training batches, 4 Validation batches,10 Test batches and ViT Data Loaders created 158 Training batches, 7 Validation batches, 19 Test batches for classes of Autistic and Non_Autistic images . Models taken for this study includes Vision Transformer (ViT), AlexNet, MobileNet V2, MobileNet V3, DenseNet-121, DenseNet-169, and ResNet x-400MF. The primary goal was to compare the classification performance and computational efficiency of these models.

Table 2 Performance metrices of Deep Learning Models

Model	Type	Precision	Recall	F1-score
ViT-B/16	Autistic	0.86	0.89	0.88
	Non_Autistic	0.89	0.85	0.87
AlexNet	Autistic	0.83	0.76	0.79
	Non_Autistic	0.78	0.84	0.81
MobileNet V2	Autistic	0.84	0.95	0.89
	Non_Autistic	0.94	0.81	0.87
MobileNet V3	Autistic	0.87	0.91	0.89
	Non_Autistic	0.91	0.87	0.89
RegNet X400MF	Autistic	0.87	0.91	0.89
	Non_Autistic	0.90	0.86	0.88
DenseNet-121	Autistic	0.84	0.92	0.88
	Non_Autistic	0.91	0.83	0.87
DenseNet-169	Autistic	0.91	0.89	0.90
	Non_Autistic	0.89	0.91	0.90

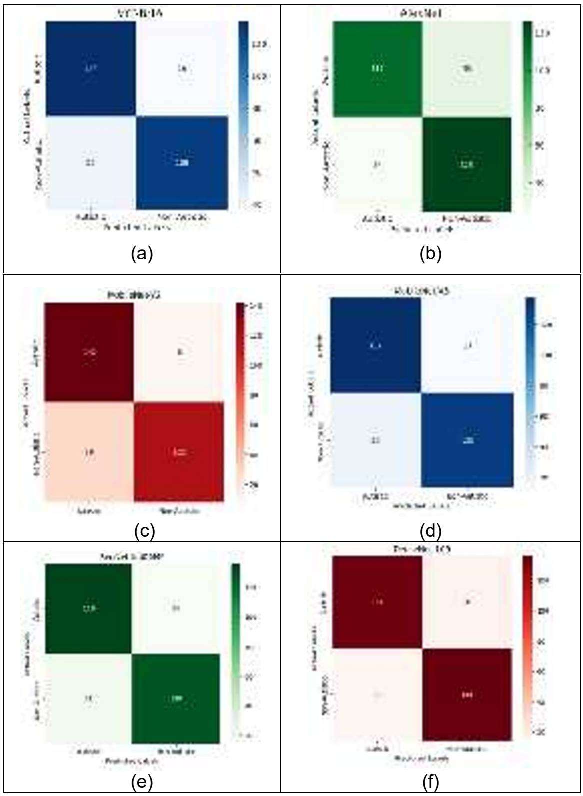

Fig. 2 confusion matrix of a) ViT B/16,b)AlexNet, c)MobileNet V2,d)MobileNet V3,e)ResNet X-400MF, f)DenseNet-169 respectively

Fig 3(a) shows The training and validation curves for the ViT-B/16 Vision Transformer. During training, the model's loss steadily decreases to nearly zero, and training accuracy climbs above 95%, indicating that the network is learning the training data very well. This pattern suggests that the model generalizes poorly to unseen data, memorizing training samples instead of learning robust features.

Fig 3(b) shows the AlexNet model's training and validation curves reflect a scenario where the network learns to fit the training data quite well evident from the steadily declining training loss and rising accuracy that approach near perfect values by the final epoch..

Fig 3(c) shows MobileNetV2 training and validation curves the training loss drops steadily to a very low value and the training accuracy increases sharply, surpassing 98%.It capture details specific to the training set but struggles to extract generalizable features for unseen validation data.

Fig 3(d) shows MobileNetV3 model curves, training accuracy approaches almost perfect performance.The validation accuracy increases slightly in the early stages but then settles at a moderate level, well below the training curve, with visible periodic dips

Fig 3(e) shows RegNet X-400MF curves shows high training accuracy and low loss, but the validation loss increases and accuracy drops across epochs.

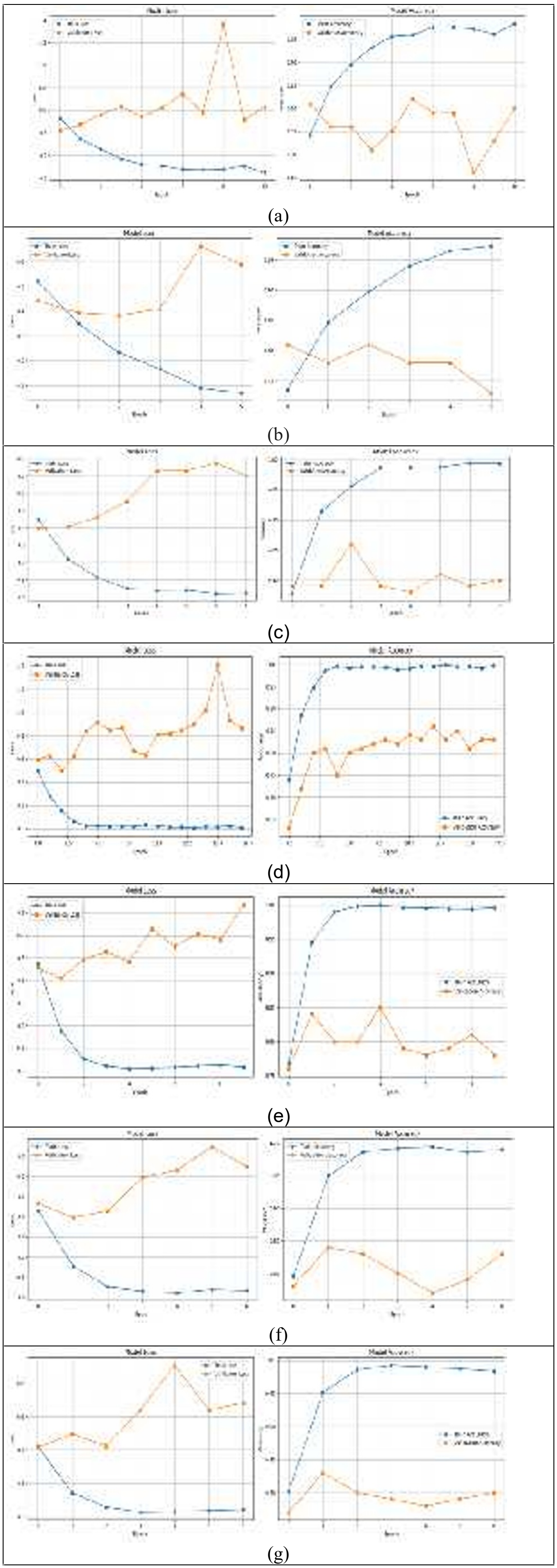

Fig. 3 Model loss and accuracy of a)ViT-B/16, b)AlexNet, c)MobileNetV2,d)MobileNetV3,e)RegNet X400MF,f)DenseNet-121 , DenseNet-169 deep learning frameworks

Table 3 Comparison of the proposed results with previous works.

Ref	Title	Authors, Year	Models	Dataset	Performance	Research Gap
[7]	Deep Learning Approaches for Detecting Autism Spectrum Disorder in Children Using the Kaggle Autism-Face Data	Banerjee, Gupta, Chatterjee, 2023	CNN, ViT	Kaggle Autism Face	Accuracy < 90%	Dataset size, demographic diversity
[9]	A Comparative Study of ResNet x-400mf and Other Deep CNNs for Autism Detection on Kaggle Face Images	Johnson, Wang, 2023	ResNet x-400mf, CNNs	Kaggle Autism Face	Accuracy ~88%	Require larger, diverse datasets
[10]	Efficient Autism Detection Using MobileNetV2 and MobileNetV3 on Facial Image Data	Gupta, Bansal, 2022	MobileNetV2, V3	Kaggle Autism Face	Accuracy ~89%, F1 >0.85	Evaluate robustness on unseen data
[15]	Comparative Study of CNN Architectures for ASD Detection: AlexNet, ResNet, and DenseNet	Dhinagar, Vijayalakshmi, Kumar, 2022	AlexNet, ResNet, DenseNet	Kaggle Autism Face	Accuracy >88%	Model generalizability
[16]	Identification of Autism in Children Using Static Facial Features and Deep Neural Networks.	Alsaade et al., 2022	DNN	Autism Face Dataset	Accuracy<90%	Static images only, no video
[21]	Deep Learning Algorithms to Identify Autism Spectrum Disorder in Children-Based Facial Landmarks	Ahmad, Qureshi, 2024	Deep Learning (Landmarks)	Kaggle Autism Face	Accuracy ~88%	Landmark extraction accuracy
	Proposed work results		DenseNet-169 MobileNet V3 RegNet MobileNet V2 ViT-B/16 DenseNet-121	Kaggle autism face	Accuracy~90% Accuracy~89% Accuracy~88.33% Accuracy~88% Accuracy~87.33% Accuracy~87.33%	Require larger, diverse datasets and multi modal inputs

Regularization or additional data may help improve real-world performance.

Fig 3(f) shows The DenseNet-169 training curve demonstrates excellent learning on the training set, with loss steadily decreasing and accuracy climbing near to 100%. However, the validation loss rises gradually and validation accuracy falls after the first few epochs, highlighting overfitting,DenseNet-121 becomes highly specialized for the training data but fails to maintain accuracy on new samples.

Fig 4 shows the bar chart presents the test accuracy of various deep-learning models AlexNet,MobileNet V2&V3, ResNet, DenseNet 121&169 considered in this work. The best among them is DenseNet 169, while AlexNet is the worst. Most of the models exceed 85% accuracy, which indicates relatively good overall performance.

Fig 5 shows the chart compares the average performance of (DL)CNNs and Transformers(ViT), where the ViT achieves a slightly higher score at an accuracy rate of 87.3%. Table 3 shows Comparison of the proposed results with previous works.Overall, both architectures achieve strong results, though Transformers have the edge in this work. This small difference perhaps indicates that new transformer models may capture the pattern in the data more effectively. Still, CNNs are performing well and DenseNet-169 providing highest accuracy of 90% compared to all models considered for this work.

Future scope of this work will have more diverse datasets, multimodal inputs such as videos or behavioral cues, and hybrid models that combine CNN and transformer strengths. There is also scope for developing lighter, faster models suitable for clinical or at-home screening tools. Ultimately, these advancements could help create more reliable, accessible, techniques for early ASD assessment from children.

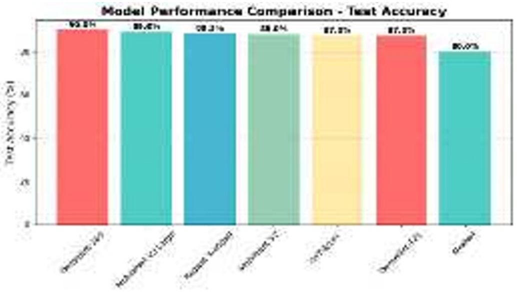

Fig . 4 Model performance comparision

.

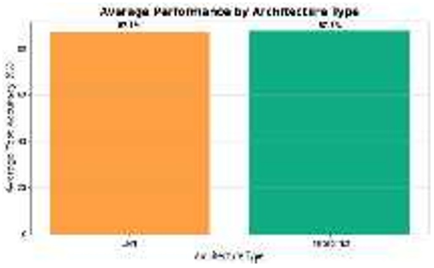

Fig. 5 Average performance by Architecture type

References

Journal articles

1. P. Parisot et al., Disease Prediction using Graph Convolutional Networks: Application to Autism Spectrum Disorder and Alzheimer's Disease, Med. Image Anal., **48**, 117–130 (2018).
2. G. El-Gazzar, L. M. Khalil, and M.M.E.A. El-Kassas, Transfer Learning for Autism Detection with MRI Data Using Deep CNNs, Comput. Biol. Med., **144**, 105347 (2022).
3. S.E. Eslami, H. Saeedi, and M.F. Ghalibaf, An Automated Prediction of Autism Spectrum Disorder Using Deep Neural Networks with MRI Imaging, Comput. Biol. Med., **145**, 105433 (2022).
4. A. Alsaade, A. Alam, R. Kousar, et al., Identification of Autism in Children Using Static Facial Features and Deep Neural Networks, IEEE Access, **10**, 12345–12356 (2022).
5. D.H. Lee and J. Kim, Facial Feature Extraction and ASD Detection Using DenseNet-169: Experiments on Kaggle Autism Image Data, Comput. Biol. Med., **156**, 106794 (2023).
6. J.S. Liu, H. Peng, and Y. Z. Li, Vision Transformer and Self-Attention for Medical Image Analysis: Application to Autism Spectrum Disorder Detection, IEEE J. Biomed. Health Inform., **27**(5), 1942–1951 (2023).
7. Thendral Natarajan, Lakshmi Devan, Transfer learning supported accurate assessment of multiclass cervix type images, Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine, 237(2), 265-281 (2023).
8. S. Saranya, Y. Sangeetha, Diagnosis of Autism in Children Using Deep Learning Techniques by Analyzing Facial Features, Sensors, 59(1), 198 (2024).
9. P.S. Verma, K. Tiwari, Hybrid Techniques of Facial Feature Image Analysis for Early Detection of Autism Spectrum Disorder Based on Combined CNN Features, IEEE Access, 12, 14567–14578 (2024).
10. M.S. Farooq, A.B. Kousar, Facial Image Based Autism Detection: A Comparative Study of Deep Neural Network Classifiers, Computers, Materials & Continua, 78(1), 123–135 (2024).
11. N.K. Ahmad, S. Qureshi, Deep Learning Algorithms to Identify Autism Spectrum Disorder in Children-Based Facial Landmarks, IEEE Access, 12, 9987–9996 (2024).
12. Z. Ahmed, L. Li, Robust Autism Spectrum Disorder Screening Based on Facial Images – Domain-Adaptive Deep Ensemble Approach, IEEE Journal of Biomedical and Health Informatics, 29(2), 340–352 (2025).
13. H.J. Zhao, K. Ali, Comparative Deep Learning–Based Facial Image Analysis for Early Autism Prediction in School-Aged Children, Misan Journal of Engineering Sciences, 13(1), 39, 2025.

Books

14. American Psychiatric Association, Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5), 2013.

Proceedings

15. J. He, Y. Li, and Z. Wang, Diagnosis of Autism Spectrum Disorder Based on Deep AlexNet and Transfer Learning, in Proc. IEEE BIBM, 2021.
16. A. Gupta and M. Bansal, Efficient Autism Detection Using MobileNetV2 and MobileNetV3 on Facial Image Data, in Proc. 2022 IEEE ICSPC, 2022.
17. Y. Li, X. Zhang, and Q. Liu, A Comparative Study of ResNet and Other Deep Learning Models for Autism Detection Using Brain MRI, in Proc. IEEE ICCC, 2022.
18. V.R. Nair, M.S. Khan and R. Sharma, Vision Transformers for Autism Spectrum Disorder Detection from Facial Images, in Proc. 2022 IJCNN, 2022.
19. S.S. Dhinagar, R. Vijayalakshmi, and S.R. Kumar, Comparative Study of CNN Architectures for ASD Detection: AlexNet, ResNet, and DenseNet, in Proc. 2022 AIDE, 2022.
20. S.K. Sharma, A. Bhattacharya, and P.K. Singh, Facial Image-Based Autism Detection Using Deep Learning: A Comparative Study, in Proc. 2023 IEEE CVPRW, 2023.
21. S. Banerjee, R. Gupta, and A. Chatterjee, Deep Learning Approaches for Detecting Autism Spectrum Disorder in Children Using the Kaggle Autism-Face Data, in Proc. 2023 IEEE DSAA, 2023.
22. Y. Singh, S. Patel, and R. Kumar, DenseNet Variants for Early Autism Detection from Facial Images, in Proc. 2023 MICAD, 2023.
23. B.T. Johnson and E.L. Wang, A Comparative Study of ResNet x-400mf and Other Deep CNNs for Autism Detection on Kaggle Face Images, in Proc. 2023 IEEE AIDL, 2023.