Performance Evaluation of Efficientnet-B0, Vision Transformer B/16, Densenet-121, CNN-CBAM And a Soft-Voting Ensemble for Autism Spectrum Disorder Detection from Facial Images

Olalere Mofiyinfoluwa Victoria[1], Emuoyibofarhe Ozichi Nweke[2], Olanloye Odunayo Dauda[3]
[1,2,3]*Computer Science, Bowen University, Iwo, Nigeria*

***Abstract*—autism spectrum disorder (ASD) is a neurodevelopmental condition that presents with difficulties in social interactions ranging unique speech patterns to repetitive behaviours and unconventional responses to sensory processing. Early detection of autism spectrum disorder improves the possibility of the individual being able to adapt to the requirements of society, yet conventional diagnosis heavily relies on time consuming and money consuming behavioural assessments over time by trained specialists, who are scarce in underserved settings such as Nigeria. Deep learning as a subset of machine learning has been proposed as in non-invasive approach to preliminary screening, but published studies have used different datasets over times applying different preprocessing pipelines and this has made the results difficult to compare. This study benchmarks 5 models under a single control preprocessing pipeline the models are namely: EfficientNet-B0, Vision Transformer B/16 (ViT-B/16), DenseNet-121, a custom convolutional neural network with a Convolutional Block Attention Module (CNN-CBAM), and a soft-voting ensemble of EfficientNet-B0 and ViT-B/16. The study was carried out using the pytorch library and it was trained on the publicly available kaggle autism image data set of 5880 facial images which were equally balanced in the database between autistic and non-autistic classes, evaluation was then carried out on a held-out test set of 300 images considering accuracy, precision, recall, F1 score and Area under curve. EfficientNet-B0 performed best, with 89.00% accuracy, 92.70% precision, 84.67% recall, 88.50% F1-score and an AUC of 0.9556. The ensemble followed (87.67% accuracy; AUC 0.9390), then DenseNet-121 (80.33%; 0.8838), CNN-CBAM (78.33%; 0.8628) and ViT-B/16 (74.00%; 0.8173). Results also support Efficient Net – B0 as the best balance of performance and parameter efficiency among the configurations tested.**

***Index Terms*—Autism Spectrum Disorder, Deep Learning, EfficientNet, Facial Image Classification, Screening, Vision Transformer**

I. INTRODUCTION

Autism spectrum disorder [ASD] is there a developmental condition which presents with difficulties in social communication restricted or repetitive patterns of behaviour in individuals and or unusual responses to sensory input. It can vary from one individual to another which sometimes complicates the detection process under search support for autism has to be individualised to each patient [1]. People with autism spectrum disorder often find verbal and nonverbal cues umm difficult to interpret ranging from eye contact to hand signals also experiencing sensory differences compared to non-autistic individuals. They have unique responses to lights, sound, touch, smell all other inputs [2].

Over time the prevalence of autism has varied in different areas and research has pointed out to the variance then hinged on diagnostic abilities our location Stating that high income countries generally report higher cases of autism, which could be reflecting the greater access to diagnostic tools rather than the real status of the occurrence of the spectrum [3]. The Centers for Disease Control and Prevention figure of one in 54 American children contrasts with an estimated 84 in every 10,000 children in Nigeria [4]. Napolitano et al. [5] Shows that in terms of occurrence the ratio remains roughly 4 mil to one female although the study also highlighted that females may be identified later due to the female their ability to mask their difficulties in social cues.

Zwaigenbaum et al. [6] Reported that behavioural observation and parents' interventions are most effective when the individual is younger than the age of 3. Although notable signs begin to occur between 12 and 14 months a study also supports that reliable identification or detection of autism is possible by the second year of the child's growth children are typically diagnosed between the age of three and five as diagnoses typically require structured caregiver interviews behavioural observation over time and clinical judgments from experienced specialists source the window for diagnosis becomes late [7], [8] added that in underserved areas digital mills of detection could reduce the cost of assessment [9], and automated identification using machine learning is faster than manual assessment [10].

Machine learning methods such as artificial neural networks, support vector machines and decision trees have been applied to ASD identification [11]. Deep learning expands this by learning hierarchical representations directly from data, allowing facial images to be analysed for morphological signs associated with the condition [12]. This approach has a biological basis. Tripi et al. [13] and Mujeeb Rahman and Subashini [14] has shown that for individuals with autism there are facial differences including increased upper facial width, increased distance between the eyes, smaller mid Fisher heights and increased width around the mouth. However, Tripi et al. [13] and Mujeeb Rahman and Subashini [14] has not translated into clear guidance on which architecture to use for automated screening. This study is to highlight the best performing model subjects to identical prepossessing conditions using the programme models: EfficientNet-B0, ViT-B/16, DenseNet-121, a custom CNN-CBAM, and a soft-voting ensemble of EfficientNet-B0 and ViT-B/16 to test how they perform in identifying the morphological features presents in facial images. All five use the same 5,880-image dataset, preprocessing and fixed test partition, so observed differences can be attributed to architecture rather than experimental variation.

II. LITERATURE REVIEW

The use of deep learning on facial images is not new study, it has been researched for over a decade with improvements along the line. Liu et al. [15] give an early example with 84% accuracy in detection of autism from children based on behaviour using a convolutional neural network. Arbabshirani et al. [16] utilized neural networks on neural imaging data, advancing the understanding of ASDs biological basis, However, the complexity of the data utilises limited clinical translation. Ahmed et al. [17] trained several CNNs on facial images and identified the mouth outline and eyes as distinguishing regions, with MobileNet reaching 95% accuracy, Xception 94% and InceptionV3 89%.

Krishnamoorthy et al. [18] assessed EfficientNet-B0 to B7 and found B3 and B7 highest at 99% accuracy and B4 lowest at 88%. Shanmathi et al. [19] applied EfficientNet to multimodal neuroimaging, facial expression, EEG and behavioural inputs, reaching 91.46%, and Bhise et al. [20] added an attention mechanism to an EfficientNet backbone, reporting 85% on 2,500 images.

Narala et al. [21] found EfficientNet performed better than ResNet50 on 2,940 facial images, and Thamilselvan et al. [22] reported 54% for DenseNet-121 against 90% for EfficientNet-B0 on a combined Kaggle and social media dataset. Wang and Miao [23] utilised VGG16, VGG19, Xception and EfficientNet-B2 on 2,540 training and 300 test images, with EfficientNet-B2 having the best performance at 90.00% specificity and 93.33% sensitivity. The highest accuracy of 98.8% was observed from the hybrid VGG16-MobileNet and Random Forest pipeline of Awaji et al. [24], although its complexity and cost limit practical deployment.

Vision Transformers processes an image as a sequence of fixed-size patches through stacked self-attention encoders and perform well on large-scale benchmarks, The growth in the dataset from 3500 to 5,880 in 2024 motivated their inclusion here [25] to see if the data available is now large enough for high accuracy results as existing comparisons rely on datasets of 3,500 images or fewer (Bhise et al. [20]: 2,500; Wang and Miao [23]: 2,840; Narala et al. [21]: 2,940), whereas this study uses 5,880.

Its contribution is methodological: the dataset, preprocessing, training environment, partitions and metrics are held constant across all algorithms, so differences in results reflect algorithm performance.

III. METHODOLOGY

Five models in total based on four individual architectures and one soft voting assemble were compared using accuracy, precision, recall, F1-score and AUC-ROC.

All were implemented in PyTorch and run in Google Colab. The workflow is shown in Fig. 1.

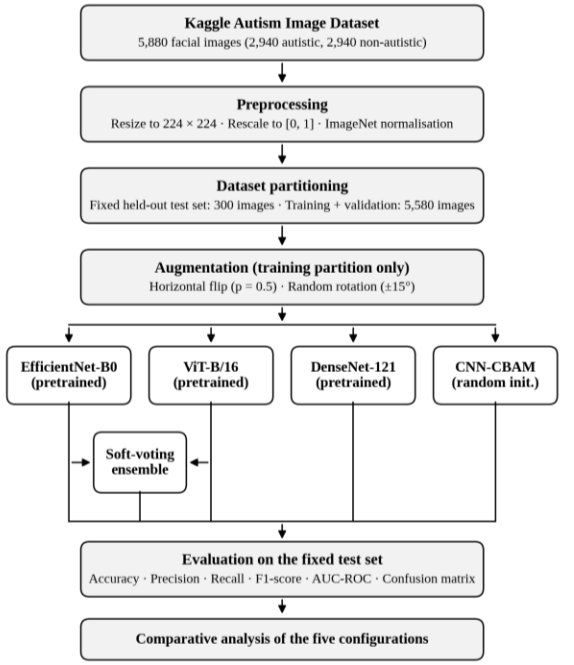

Fig. 1. Flow chart of the study

A. Dataset

This study used the Publicly available autism image datasets which was published in 2024 and contains 5880 facial images divided equally into the autistic and non-autistic classes [26]. The data set was chosen for its size class balance and demographic variation due to the equal division of images into 2940 autistic and 2940 non autistic images accuracy is unbiased.



Fig. 2. Sample autistic and non-autistic facial images from the Autism Image Dataset

B. Data Preprocessing

Preprocessing was applied identically to all five configurations using the Albumentations library [27] in three stages: resizing and normalisation, data augmentation, and dataset partitioning using a PyTorch pipeline.

1) Resizing and Normalisation:

All images were resized to 224 × 224 pixels, the input resolution of the ImageNet-pretrained library which also provided the guidelines to be rescaled to the unit interval and standardised channel-wise using the ImageNet mean (0.485, 0.456, 0.406) and standard deviation (0.229, 0.224, 0.225), as expressed in (1) to (3).

$$x^{(1)} = Resize_{224\times224}(I_i) \in \mathbb{R}^{224\times224\times3} \quad (1)$$

$$x^{(2)} = \frac{x^{(1)}}{255} \in [0,1]^{224\times224\times3} \quad (2)$$

$$x_c^{(3)} = \frac{x_c^{(2)} - \mu_c}{\sigma_c}\ ,\ c \in \{R,G,B\} \quad (3)$$

where I_i is the i-th input image; $x^{(1)}$ is the image resized to 224 × 224 pixels; $x^{(2)}$ is the resized image with its 8-bit pixel values (0–255) rescaled to the interval [0, 1]; $x_c^{(3)}$ is the normalised value of colour channel c; μ_c and σ_c are the ImageNet mean and standard deviation for channel c; and $\mathbb{R}^{224x224x3}$ denotes a 224 × 224 image with three colour channels (R, G, B).

ImageNet preprocessing guidelines were used because EfficientNet-B0, ViT-B/16 and DenseNet-121 are primed with ImageNet weights, and according to He et al. [28] matching the input distribution to those weights is standard transfer learning practice.

2) Augmentation:

Augmentation was applied only to the training partition, so that test performance could not be inflated. It comprised a horizontal flip with probability 0.5 and a random rotation within ±15 degrees, as in (4) and (5).

$$A(x) = R_\varphi \circ F_h(x) \quad (4)$$

$$F_h \sim Bernoulli(0.5)\ ,\ \varphi \sim U(-15°, +15°) \quad (5)$$

where A(x) is the augmented image, F_h the horizontal flip, R_φ a rotation by angle φ, U the uniform distribution and ∘ function composition.

Only geometric transformations that preserve facial proportions were used, since the distinguishing features are in facial proportions such as facial width and distance between the eyes [14].

3) Dataset Partitioning:

The dataset was divided into disjoint training, validation and test partitions, as in (6) and (7). The test partition of 300 images (150 autistic, 150 non-autistic) was held out before training and used to evaluate all five configurations; the remaining 5,580 images formed the training and validation partitions.

$$D = D_{train} \cup D_{val} \cup D_{test}\,,\ D_{train} \cap D_{val} \cap D_{test} = \emptyset \quad (6)$$

$$|D_{test}| = 300\,,\ |D_{train}| + |D_{val}| = 5{,}580 \quad (7)$$

where D is the full dataset; D_train, D_val and D_test are the training, validation and test partitions; $\emptyset$ is the empty set, meaning no image appears in more than one partition; and $|\cdot|$ is the number of images in a partition.

C. Detection Performance

Five algorithms were evaluated: four models and a soft-voting ensemble of two of them. Each providing a result based on the model's estimation of the probability that a facial image belongs to the autistic class, as in (8).

$$f_\theta : \mathbb{R}^{224\times224\times3} \rightarrow [0,1]^2\,,\ f_\theta(x) = (p_0, p_1)\,,\ p_0 + p_1 = 1 \quad (8)$$

where f_θ is the classification model and θ its learnable parameters (network weights); x is a preprocessed 224 × 224 × 3 input image; p_1 is the predicted probability that the image belongs to the autistic class; and $p_0 = 1 - p_1$ is the predicted probability that it belongs to the non-autistic class.

1) EfficientNet-B0:

EfficientNet-B0 uses compound scaling to balance network depth, width and resolution, and is built from mobile inverted bottleneck (MBConv) blocks with squeeze-and-excitation channel recalibration [29], [30]. The ImageNet-pretrained torchvision model was fine-tuned end to end, with its classification head replaced by a dropout layer (rate 0.3) and a two-neuron fully connected output layer. It has about 4.01 million trainable parameters and ran for 21 epochs with a batch size of 32, using the Adam optimiser [31] and cosine annealing with warm restarts [32].

2) Vision Transformer B/16 (ViT-B/16):

The Vision Transformer represents an image as a sequence of fixed-size patch tokens with a prepended classification token, processed by 12 Transformer encoder blocks of multi-head self-attention and feed-forward sublayers [25]. The ImageNet-pretrained torchvision ViT-B/16 was used, with approximately 85.80 million trainable parameters, the largest standalone model in the study. Because it lacks the spatial inductive biases of CNNs, it depends heavily on large-scale data, which was expected to limit its performance on this dataset. It was trained for 11 epochs.

3) DenseNet-121:

DenseNet-121 [33] connects each layer to all subsequent layers within four dense blocks, separated by transition layers of 1 × 1 convolution and 2 × 2 average pooling. This improves feature reuse and gradient flow with fewer parameters, and its combination of multi-scale features suits the detection of subtle facial clues. The ImageNet-pretrained model has about 6.96 million trainable parameters and was trained for 14 epochs.

4) Convolutional Neural Network with Convolutional Block Attention Module (CNN-CBAM):

CBAM [34] is a lightweight attention module that applies channel attention, computed from average-pooled and max-pooled features through a shared multi-layer perceptron, followed by spatial attention to highlight the most relevant locations.

The custom CNN-CBAM consists of convolutional blocks of a 3 × 3 convolution, batch normalisation, ReLU, 2 × 2 max pooling and a CBAM module, followed by two fully connected layers with dropout (rate 0.5). It has 1.2 million parameters, and it was trained from random initialisation rather than ImageNet weights, for 10 epochs.

5) ViT-EfficientNet Soft-Voting Ensemble:

The ensemble averages the class probabilities of the trained EfficientNet-B0 and ViT-B/16 models and assigns the class with the highest average probability, as in (9).

$$\hat{y} = arg\,max_c\ \frac{1}{2}\left(p_c^{(EfficientNet)} + p_c^{(ViT)}\right) \quad (9)$$

where p_c^(EfficientNet) and p_c^(ViT) are the probabilities assigned to class c by EfficientNet-B0 and ViT-B/16 respectively, and $\hat{y}$ is the predicted class.

The pairing combines complementary inductive biases from EfficientNet-B0 and ViT-B/16 and has 89.81 million parameters the largest in this study.

D. Evaluation Metrics

All Models were tested on a fixed test partition data size of 300 using five metrics namely accuracy, precision, recall, F1 score and AUC-ROC. True positive are images which were autistic in the data set and accurately classified as autistic by the model (TP), true negative a non-autistic image that were correctly classified as non-autistic (TN) false positives are non-autistic images that were classified as autistic (FP), while false negatives are autistic images which were wrongly classified as non-autistic (FN).

1) Accuracy:

Accuracy highlights the number of test images in the data sets that we've correctly classified either as autistic or non-autistic.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (10)$$

2) Precision:

Precision represents the number of images correctly classified as autistic; this also highlights the model's ability to avoid false positives.

$$Precision = \frac{TP}{TP + FP} \quad (11)$$

3) Recall (Sensitivity):

Recall highlights the number of autistic images that were correctly identified as autistic, hello recall would mean that missed cases would not proceed for clinical assessment.

$$Recall = \frac{TP}{TP + FN} \quad (12)$$

4) F1-Score:

The F1-score is the harmonized mean of precision and recall, reported alongside both because models with equal F1-scores may distribute their errors differently.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (13)$$

5) Area Under the ROC Curve (AUC-ROC):

AUC-ROC Is a graph plot of true positive against false positive rate where 1.0 indicates perfect discrimination and 0.5 chance. it tests the model's ability to separate autistic from non-autistic classes rather than merely performing well at the 0.5 cut-off.

IV. RESULTS AND DISCUSSION

A. Training Behaviour

EfficientNet-B0 demonstrated fast and steady improvement, with training accuracy moving from 75.6% at the first epoch to 99.6% by epoch 21. This trajectory shows the model had a good adaptation of pretrained ImageNet representations to the ASD domain.

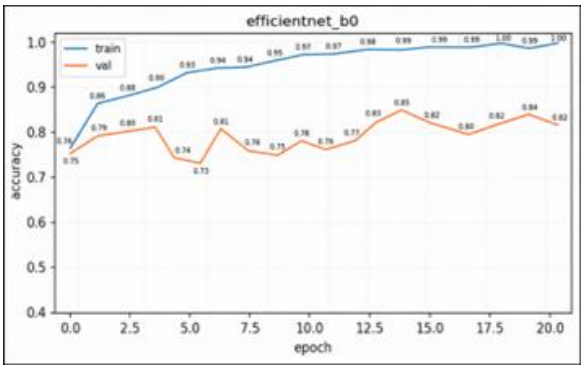

Fig. 3. Training and validation curves for EfficientNet-B0.

DenseNet-121 showed improvement in training accuracy from 76.6% to 96.2%, with validation accuracy peaking at 81.0% at epoch 7.

The steady rise across both curves is consistent with the dense connectivity of the architecture, where the direct connections between layers support feature reuse and allow gradients to move more smoothly through the network during training.

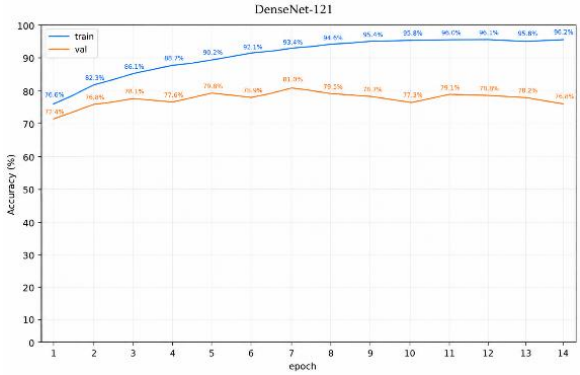

Fig. 4. Training and validation curves for DenseNet-121.

CBAM improved gradually from 65.1% to 79.7% in training accuracy, with validation accuracy reaching 80.0% at epoch 3 before stabilising in the 73–78% range, a pattern consistent with the challenge of learning from random weight initialisation.

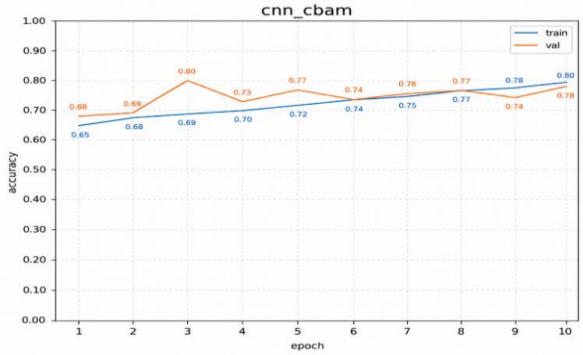

Fig. 5. Training and validation curves for CNN-CBAM.

The ViT-EfficientNet ensemble was not trained as a separate network, since it combines the frozen predictions of two already-trained models.

Its curve therefore tracks the accuracy obtained by averaging the two probability outputs at each epoch of the EfficientNet-B0.

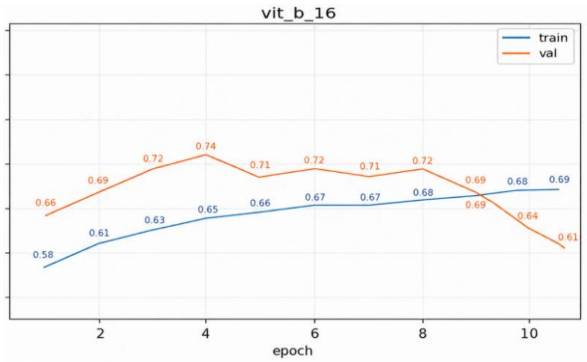

Fig. 6. Training and validation curves for ViT-B/16.

B. Receiver Operating Characteristic Analysis

AUC-ROC ranged from 0.8173 (ViT-B/16) to 0.9556 (EfficientNet-B0), with the ensemble at 0.9390, DenseNet-121 at 0.8838 and CNN-CBAM at 0.8628. All values are well above the chance level of 0.5.

C. Summary of Performance Evaluation

Table I summarises the performance, parameter counts and training epochs of the five configurations.

TABLE I Summary of Performance Evaluation across the Five Configurations

Configuration	Parameters	Epochs	Accuracy	Precision	Recall	F1-Score	AUC
EfficientNet-B0	4.01 M	21	89.00%	92.70%	84.67%	88.50%	0.9556
ViT-EfficientNet Ensemble	89.81 M	21	87.67%	87.92%	87.33%	87.62%	0.9390
DenseNet-121	6.96 M	14	80.33%	80.13%	80.67%	80.40%	0.8838
CNN-CBAM	1.20 M	10	78.33%	76.40%	82.00%	79.10%	0.8628
ViT-B/16	85.80 M	11	74.00%	78.57%	66.00%	71.74%	0.8173

D. Discussion

EfficientNet-B0 performed best (89.00% accuracy, AUC 0.9556), with precision (92.70%) exceeding recall (84.67%), highlighting the model's ability to avoid false positives. DenseNet-121 (80.33%) had nearly balanced precision and recall, CNN-CBAM at an accuracy of 78.33% had the second-highest standalone recall (82.00%), consistent with Bhise et al. [20]. ViT-B/16 performed lowest with accuracy at 74.00% and 66.00% recall, however this was also consistent with the data requirements of Vision Transformers [25]; reinforcing the need to have a large dataset when utilising the vision transformer. The ensemble (87.67%, AUC 0.9390) ranked second, trading precision (87.92%) for the highest recall in the study (87.33%).

Table II compares these results with the similar facial-image autism detection studies. ViT-B/16 had not been evaluated for this task in the literature reviewed; its underperformance shows that benchmark strength does not automatically transfer to specialised medical imaging at moderate data scale.

TABLE II Comparison of the Present Study with Recent Deep Learning Studies on ASD Detection from Facial Images

Study	Dataset	Algorithm	Accuracy
Narala et al. [21]	Facial images (2,940)	EfficientNet, ResNet50	96.00% (ResNet50 lower than EfficientNet)
Ahmed et al. [17]	Kaggle (3,014)	MobileNet, Xception, InceptionV3	95% / 94% / 89%

Study	Dataset	Algorithm	Accuracy
Krishnamoorthy et al. [18]	Kaggle Autism Image Dataset (5,880)	EfficientNet-B3 / B7 / B2 / B5 / B4	99.00% / 99.00% / 98.00% / 97.00% / 88.00%
This study	Kaggle Autism Image Dataset (5,880)	EfficientNet-B0, ViT-EfficientNet Ensemble, DenseNet-121, CNN-CBAM, ViT-B/16	89.00% / 87.67% / 80.33% / 78.33% / 74.00%

Overall, EfficientNet-B0 had the best balance of performance and parameter efficiency, which matters where computing resources are limited.

E. Ethical Considerations

The dataset is publicly available for research and contains no personally identifiable information; no participants were recruited. Its labels were not verified against clinical records, and any model output would require confirmation by a qualified clinician. Thus, this study does not generalise to clinically confirmed populations. Future work should evaluate larger EfficientNet variants, data-efficient Transformers such as DeiT [35] using clinically confirmed cases from Nigeria. The models are intended as a first-stage screening aid, not a diagnostic tool.

REFERENCES

[1] B. Rashid and E. Shaker, "Social communication in autism spectrum disorder: A review of current research," *Current Developmental Disorders Reports*, vol. 10, no. 1, pp. 36–44, 2023.

[2] N. Patil and M. Kaple, "Autism spectrum disorder: Behavioural analysis and deep learning classification," *International Journal of Engineering Research and Technology*, vol. 12, no. 4, pp. 601–607, 2023.

[3] N. Salari, S. Rasoulpoor, S. Rasoulpoor, S. Shohaimi, S. Jafarpour, N. Abdoli, *et al*., "The global prevalence of autism spectrum disorder: A comprehensive systematic review and meta-analysis," *Italian Journal of Pediatrics*, vol. 48, no. 1, Art. no. 112, 2022.

[4] A. Mulay, *Autism Spectrum Disorder: Prevalence and Global Trends*. DataPandas Health Analytics, 2023.

[5] A. Napolitano, S. Schiavi, P. La Rosa, M. C. Rossi-Espagnet, S. Pecci, C. Napolitano, *et al*., "Sex differences in autism spectrum disorder: Diagnostic, neurobiological, and behavioral features," *Frontiers in Psychiatry*, vol. 13, Art. no. 889636, 2022.

[6] L. Zwaigenbaum, M. L. Bauman, D. Fein, K. Pierce, T. Buie, P. A. Davis, *et al*., "Early screening of autism spectrum disorder: Recommendations for practice and research," *Pediatrics*, vol. 136, suppl. 1, pp. S41–S59, 2015.

[7] Y. Yu, G. W. Andersson, S. Lundstrom, C. Gillberg, and C. Miniscalco, "Use of machine learning to identify autism spectrum disorder from speech and language data," *Frontiers in Neuroscience*, vol. 17, Art. no. 1130700, 2023.

[8] C. Lord, T. Charman, A. Havdahl, and P. Carbone, "The Lancet Commission on the future of care and clinical research in autism," *The Lancet*, vol. 399, no. 10321, pp. 271–334, 2022.

[9] G. Dawson, G. Sapiro, and R. Bernier, "Potential for digital behavioral measurement tools to transform the detection and diagnosis of autism spectrum disorder," *JAMA Pediatrics*, vol. 176, no. 7, pp. 712–718, 2022.

[10] M. Uddin, A. Karim, M. Chowdhury, M. M. Alam, and M. Hossain, "Automated detection of autism spectrum disorder from behavioral and neuroimaging data: A comprehensive review," *IEEE Access*, vol. 12, pp. 12478–12502, 2024.

[11] A. Simeoli, M. De Vita, G. Coppola, and S. Buono, "Supervised machine learning for autism spectrum disorder classification: A systematic review," *International Journal of Environmental Research and Public Health*, vol. 21, no. 2, Art. no. 156, 2024.

[12] F. W. Alsaade and M. S. Alzahrani, "Classification and detection of autism spectrum disorder based on deep learning algorithms," *Computational Intelligence and Neuroscience*, vol. 2022, Art. no. 8709145, 2022.

[13] G. Tripi, S. Roux, D. Matranga, L. Maniscalco, P. Glorioso, F. Bonnet-Brilhault, *et al*., "Cranio-facial characteristics in children with autism spectrum disorders (ASD)," *Journal of Clinical Medicine*, vol. 8, no. 5, Art. no. 641, 2019.

[14] K. K. Mujeeb Rahman and M. M. Subashini, "Identification of autism in children using static facial features and deep neural networks," *Brain Sciences*, vol. 12, no. 1, Art. no. 94, 2022.

[15] W. Liu, M. Li, and L. Yi, "Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework," *Autism Research*, vol. 9, no. 8, pp. 888–898, 2016.

[16] M. R. Arbabshirani, S. Plis, J. Sui, and V. D. Calhoun, "Single subject prediction of brain disorders in neuroimaging," *NeuroImage*, vol. 145, pp. 137–165, 2017.

[17] Z. A. T. Ahmed, N. Alkhaldi, H. AlSulaiman, M. AlRubaian, M. AlSahly, *et al*., "Facial features detection system to identify children with autism spectrum disorder: Deep learning models," *Computers and Electrical Engineering*, vol. 101, Art. no. 108098, 2022.

[18] V. Krishnamoorthy, N. Natarajan, and R. Suresh Kumar, "autism spectrum disorder prediction from facial images using fine-tuned EfficientNet B0–B7 architectures," *Intelligent Systems with Applications*, vol. 26, Art. no. 200460, 2025.

[19] B. Shanmathi, V. Sathiyamoorthi, and K. Dhanya, "EfficientNet-powered multi-modal system for non-invasive diagnosis of autism spectrum disorder," *Biomedical Signal Processing and Control*, vol. 99, Art. no. 106851, 2025.

[20] A. Bhise, A. Rao, N. Rao, S. Shetty, and T. Shah, "Facial-based autism spectrum disorder diagnosis using enhanced deep learning models," in *Proc. 2024 Int. Conf. Inventive Computation Technologies (InC)*, 2024, pp. 1–6.

[21] M. S. V. S. K. Narala, S. Pathak, M. Rajeswari, A. Tamilarasi, and C. Madhavi, "A comparative study for autism spectrum disorder using EfficientNet and ResNet50," *Journal of Theoretical and Applied Information Technology*, vol. 101, no. 15, pp. 6170–6181, 2023.

[22] R. Thamilselvan, G. Archana, M. Akash, and D. Baraniya, "autism spectrum disorder diagnosis using deep learning techniques," in *Proc. 2024 Third Int. Conf. Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*, 2024, pp. 1–6.

[23] H. Wang and Y. Miao, "Deep learning-based autism spectrum disorder detection using facial image data," *Journal of Healthcare Engineering*, vol. 2024, Art. no. 3462785, 2024.

[24] B. Awaji, E. Abusham, and M. Hassan, "Hybrid techniques of facial feature image analysis for early detection of autism spectrum disorder based on combined CNN features," *Technologies*, vol. 11, no. 3, Art. no. 62, 2023.

[25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, *et al*., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2021.

[26] harsh0251, "Autism image dataset," *Kaggle*, 2024. [Online]. Available: Autism Image Dataset

[27] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Papadopoulos, A. A. Kalinin, *et al*., "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, Art. no. 125, 2020.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[29] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Machine Learning*, 2019, pp. 6105–6114.

[30] R. Correa, M. Barroso, and A. Ortiz, "EfficientNet-based classification of brain tumor using MRI," *Computers*, vol. 12, no. 4, Art. no. 88, 2023.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2015.

[32] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2017.

[33] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.

[34] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. European Conf. Computer Vision (ECCV)*, 2018, pp. 3–19.

[35] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, *et al*., "Training data-efficient image transformers and distillation through attention," in

Proc. 38th Int. Conf. Machine Learning, 2021, pp. 10347–10357.

[36] M. Alshammeri, N. Tariq, and M. Humayun, "Ensemble deep neural networks for reliable autism diagnosis in children: A performance and interpretability study," *Diagnostics*, vol. 15, no. 4, Art. no. 502, 2025.

[37] M. S. Farooq, R. Tehseen, M. Sabir, and Z. Atal, "Detection of autism spectrum disorder (ASD) in children and adults using machine learning," *Scientific Reports*, vol. 13, no. 1, Art. no. 9605, 2023.

[38] S. Raj and S. Masood, "Analysis and detection of autism spectrum disorder using machine learning techniques," *Procedia Computer Science*, vol. 167, pp. 994–1004, 2020.

[39] F. Thabtah, "An accessible and integrated screening tool for autism spectrum disorder," *Health Informatics Journal*, vol. 25, no. 4, pp. 1739–1755, 2018.