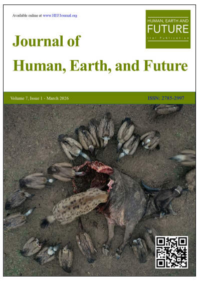
Available online at www.HEFJournal.org

Journal of
Human, Earth, and Future

ISSN: 2785-2997

Vol. 7, No. 1, March, 2026



Evaluating Deep Learning Models for Autism Detection in Children Using Facial Images

Udita J. Monani [1], Ritu Maity [2], Prasant Kumar Pattnaik [1],
Kalaiarasi Sonai Muthu Anbananthen [3*], Saravanan Muthaiyah [4], Mangal Sain [5]

[1] *School of Computer Science and Engineering, Kalinga Institute of Industrial Technology (KIIT), Bhubaneswar, 751024 Odisha, India.*

[2] *School of Mechanical Engineering, Kalinga Institute of Industrial Technology (KIIT), Bhubaneswar, 751024 Odisha, India.*

[3] *Faculty of Information Science and Technology, Multimedia University, Melaka 75450, Malaysia.*

[4] *School of Business and Technology, International Medical University, Kuala Lumpur 57000, Malaysia.*

[5] *Division of Computer & Information Engineering, Regional Innovation Center, Dongseo University, Busan 47011, Republic of Korea.*

Received 01 July 2025; Revised 19 January 2026; Accepted 24 February 2026; Published 01 March 2026

Abstract

This study develops and evaluates a comprehensive deep-learning framework for early detection of Autism Spectrum Disorder (ASD) through facial image analysis. Five state-of-the-art convolutional neural network (CNN) architectures, VGG16, VGG19, ResNet50, InceptionV3, and MobileNet, were systematically assessed using a balanced dataset of 5,000 images (2,500 ASD, 2,500 non-ASD). Transfer learning and data augmentation enhanced model generalization. VGG19 achieved the highest overall accuracy (77.89%) and F1-score (0.7962), ResNet50 attained the best precision (82.53%), and InceptionV3 produced the highest recall (99.67%), indicating strong screening potential. The findings confirm that deep CNNs can capture subtle facial morphological cues linked to ASD, supporting their feasibility as non-invasive diagnostic tools. This work provides a benchmark for future multimodal, explainable, and clinically validated AI systems for autism detection.

Keywords: Autism Spectrum Disorder; Deep Learning; Facial Image Analysis; Computer-Aided Diagnosis.

1. Introduction

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition that typically manifests in early childhood, affecting cognitive, social, emotional, motor, and sensory functions. According to the Centers for Disease Control and Prevention, approximately one in fifty-four children in the United States are diagnosed with ASD [1], making it a significant global public health concern. Early and accurate diagnosis is critical for improving treatment outcomes and developmental trajectories. However, achieving diagnostic accuracy remains challenging due to symptom heterogeneity and the absence of objective biological markers.

Conventional ASD diagnosis relies primarily on behavioral evaluations and standardized clinical instruments. Although these methods offer valuable insights, they are inherently subjective and depend heavily on clinician expertise, which can lead to inconsistent interpretations [2]. In response, researchers have increasingly explored

* Corresponding author: kalaiarasi@mmu.edu.my

https://doi.org/10.28991/HEF-2026-07-01-03

neuroimaging-based approaches, such as electroencephalography (EEG), magnetoencephalography (MEG), and functional magnetic resonance imaging (fMRI), to identify neurophysiological patterns associated with ASD. EEG, in particular, provides a non-invasive, low-cost modality capable of capturing real-time neural dynamics. Studies have shown that EEG-derived features, including altered oscillations and functional connectivity, can distinguish children with ASD when analyzed using machine learning (ML) models [3].

Recent advances in artificial intelligence have further expanded diagnostic possibilities, with deep learning models demonstrating accuracies between 85–99% across various modalities. Multimodal studies combining EEG and eye-tracking have achieved an AUC of 0.75 in toddlers aged two to four years [4], while ensemble models integrating VGG16 and Xception networks have reported up to 97% accuracy on facial images [5]. These findings underscore the promise of combining behavioral, physiological, and visual biomarkers for robust ASD detection.

Explainable AI (XAI) has emerged as another critical development for clinical adoption. For instance, Kasri et al. [6] achieved 98.2% accuracy using VGG19 enhanced with Local Interpretable Model-agnostic Explanations (LIME), offering transparency into feature importance. Similarly, Vidivelli et al. [7] demonstrated that multimodal fusion of facial images and EEG signals using a hybrid CNN-BiGRU model achieved 91.03% accuracy, while Ganesh et al. [8] showed that thermal facial imaging could detect ASD-related temperature variations with 96% accuracy. The use of ResNet architectures has further validated the potential of thermal features, achieving up to 99.41% accuracy [9].

Eye-tracking (ET) technologies have also proven valuable, revealing atypical gaze patterns such as reduced fixation on faces and social stimuli in ASD populations [10]. These findings suggest that behavioral cues and facial dynamics contain diagnostic information that complements neuroimaging data. Nevertheless, unimodal approaches often fail to capture the multidimensional nature of ASD, prompting growing interest in multimodal fusion strategies that integrate neurophysiological and behavioral data to enhance diagnostic precision [11, 12].

The emergence of Vision Transformer (ViT) architectures has recently expanded ASD detection capabilities by leveraging self-attention mechanisms for superior feature extraction. Kasri et al. [6] introduced a hybrid ViT-Mamba framework achieving 96% accuracy on eye-tracking data, while Mahmood et al. [13] combined ViT with ResNet152 to attain 91.33% accuracy, highlighting the growing potential of transformer-based models in ASD diagnostics.

Facial analysis, in particular, offers a non-invasive, accessible, and scalable approach to ASD detection. Children with ASD exhibit subtle craniofacial variations and atypical facial expressions linked to socio-emotional processing deficits [14]. Despite growing evidence, few studies have conducted systematic comparisons of modern convolutional neural networks (CNNs) for facial-based ASD classification. CNNs are well-established in computer vision for their ability to automatically extract hierarchical features, making them ideal for capturing the nuanced facial characteristics associated with ASD.

Despite these advances, several critical gaps remain: (1) limited comparative analyses of modern CNN architectures on standardized facial datasets; (2) insufficient demographic diversity in existing datasets, which constrains model generalizability; (3) minimal integration of facial analysis within multimodal diagnostic frameworks; and (4) limited clinical validation under real-world conditions. This study addresses these gaps by systematically evaluating five CNN architectures on a balanced facial dataset, thereby establishing performance benchmarks and providing a foundation for future multimodal integration.

The main contributions of this study are fourfold. (1) We propose a non-invasive, AI-driven diagnostic framework using facial-image analysis for ASD detection, providing an accessible alternative to behavioral assessments. (2) We conduct a comprehensive comparative evaluation of five state-of-the-art CNN architectures, VGG16, VGG19, ResNet50, InceptionV3, and MobileNet, highlighting their relative strengths in classifying ASD-related facial features. (3) We establish facial morphology as an underutilized yet informative biomarker for ASD. (4) We position facial image-based deep learning as a foundational component for future multimodal diagnostic systems, contributing to the development of robust, interpretable, and clinically deployable AI solutions.

The rest of this paper is structured as follows: Section 2 reviews the related literature; Section 3 details the methodology and model design; Section 4 outlines the experimental setup and evaluation metrics; Section 5 presents and discusses the results; and Section 6 concludes with future research directions.

2. Literature Review

ASD is a complex neurodevelopmental condition marked by significant variability in clinical presentation, symptom severity, and comorbidities. Traditional ASD diagnostic approaches rely primarily on behavioral observations and standardized assessments, which are inherently subjective and may lead to delayed or inconsistent diagnoses. The heterogeneity of ASD symptoms further complicates clinical evaluation, as manifestations vary widely across individuals and often overlap with other developmental disorders. These limitations have driven researchers to explore objective, technology-driven methods to support early and accurate ASD detection [14-16].

2.1. Neuroimaging-Based Detection Approaches

Neuroimaging has emerged as a promising pathway for identifying biological markers associated with ASD. Structural and functional modalities, including magnetic resonance imaging (MRI) and functional MRI (fMRI), have revealed atypical brain connectivity and activation patterns in regions linked to social cognition, sensory processing, and emotional regulation. A comprehensive review by Moridian et al. [1] demonstrated the potential of artificial intelligence techniques in extracting predictive neurobiological features from MRI data. Convolutional Neural Networks (CNNs) applied to resting-state fMRI have achieved classification accuracies between 70 % and 95 %, depending on dataset size, quality, and preprocessing parameters [17]. These findings suggest that deep learning can effectively identify subtle neurological variations related to ASD.

2.2. Facial Image Analysis for ASD Detection

In addition to neuroimaging, facial image analysis has gained increasing attention due to evidence that individuals with ASD exhibit subtle craniofacial differences and atypical facial expressions. Tripi et al. [10] identified distinctive facial morphology in children with ASD, including wider upper faces, widely spaced eyes, and shortened mid-facial regions, providing a foundation for image-based diagnostic research. Computational studies, such as those conducted by Guha et al. [3], have also revealed abnormal expressive facial dynamics in children with ASD, particularly in contexts related to emotional processing and social interaction. These findings support the hypothesis that facial features may serve as non-invasive biomarkers for ASD.

Recent advances in deep learning have further strengthened this research direction. Alkahtani et al. [2] employed facial landmark detection integrated with CNNs, achieving high classification accuracy by leveraging geometric relationships among key facial points. Transfer learning has enhanced performance in datasets with limited size, while ensemble strategies combining multiple CNN models have reached accuracy rates exceeding 97 %, demonstrating the clinical promise of these methods [18 ,19]. Farhat et al. [5] further advanced this domain by developing a VGG16–Xception ensemble model achieving 97 % accuracy on ASD facial datasets, underscoring the advantage of ensemble architectures for feature diversity and robustness.

2.3. Emerging Modalities: Thermal Imaging

Thermal imaging has recently emerged as a complementary modality in ASD detection, leveraging physiological cues related to emotional and autonomic regulation. Ganesh et al. [8] reported that thermal variations in children with ASD correlate with emotional arousal, showing consistent temperature differences compared to neurotypical peers. When coupled with deep learning, thermal imaging achieved sensitivity rates up to 100 % and specificity of 93 %, confirming its diagnostic potential. Ahmadiar et al. [9] expanded on this by using ResNet architectures for thermal image classification of ASD and neurotypical children, achieving 99.41 % accuracy, which validates the utility of thermal modalities within deep learning pipelines

2.4. Methodological Advances in Machine Learning for ASD

ML and deep learning (DL) approaches have significantly advanced the automation of ASD detection. Recent reviews emphasize that DL models outperform traditional statistical methods, particularly in processing high-dimensional, multimodal data [20]. CNNs, transfer learning, and ensemble frameworks enable more effective feature extraction and generalization across imaging modalities. McCarty & Frye [14] highlighted ongoing challenges in early ASD identification, while Dawson & Bernier [16] discussed advancements in developing more sensitive screening and intervention tools. However, robust model validation using diverse and clinically representative datasets remains a critical requirement for clinical translation [15].

2.5. Multimodal and Explainable AI Approaches

Multimodal fusion represents a leading frontier in ASD detection research, combining facial, physiological, and behavioral data streams for improved sensitivity and generalization. Vidivelli et al. [7] demonstrated a hybrid CNN–BiGRU model integrating EEG and facial data, achieving 91.03 % accuracy, highlighting the synergy between visual and neurophysiological modalities. Sun et al. [4] similarly combined EEG and eye-tracking data in toddlers aged two to four years, obtaining an AUC of 0.75 for ASD prediction, reinforcing the promise of early multimodal diagnostics.

Explainable AI (XAI) frameworks have further enhanced the interpretability of ASD detection systems. Atlam et al. [21] achieved 98.2 % accuracy using VGG19 with the LIME approach, demonstrating how interpretable feature attribution improves clinical trust. In parallel, the emergence of Vision Transformers (ViTs) has transformed ASD research by introducing self-attention mechanisms for superior feature extraction. Kasri et al. [6] proposed a hybrid ViT–Mamba framework that achieved 96 % accuracy on eye-tracking datasets, while Mahmood et al. [13] integrated ViT with ResNet152, attaining 91.33 % accuracy in ASD facial classification tasks. Collectively, these approaches highlight a paradigm shift toward transparent, hybrid, and transformer-based diagnostic systems.

2.6. Ethical Considerations and Future Directions

Despite the promise of AI-powered ASD detection tools, several challenges remain. Limitations in dataset size, demographic diversity, and clinical validation must be addressed to ensure generalizability. Future research should focus on expanding datasets, incorporating multimodal features, improving model interpretability, and conducting large-scale clinical validation studies in collaboration with healthcare providers [15]. Furthermore, ethical concerns surrounding patient privacy, informed consent, and data security must be rigorously addressed to ensure compliance with healthcare regulations and safeguard patient rights. In this context, integrating AI diagnostics with a ubiquitous Personal Health Record framework [22] can enhance patient control over health data, enable secure sharing with providers, and support personalized, longitudinal care planning

Building on this literature, the next section details the proposed deep learning methodology for ASD detection using facial images. This framework combines data preprocessing, CNN-based feature extraction, and model optimization to evaluate and benchmark state-of-the-art architectures.

3. Research Methodology

This study employs state-of-the-art deep learning algorithms to analyze facial images to identify children with ASD. Several CNN architectures and image processing techniques are applied to build a robust diagnostic framework. This section outlines the data collection, preprocessing, model selection, training procedures, and evaluation metrics. Each stage was designed to ensure transparency, reproducibility, and clinical applicability. Figure 1 illustrates the proposed deep-learning framework for ASD detection using facial image analysis, showing the key stages of preprocessing, model training, and evaluation.

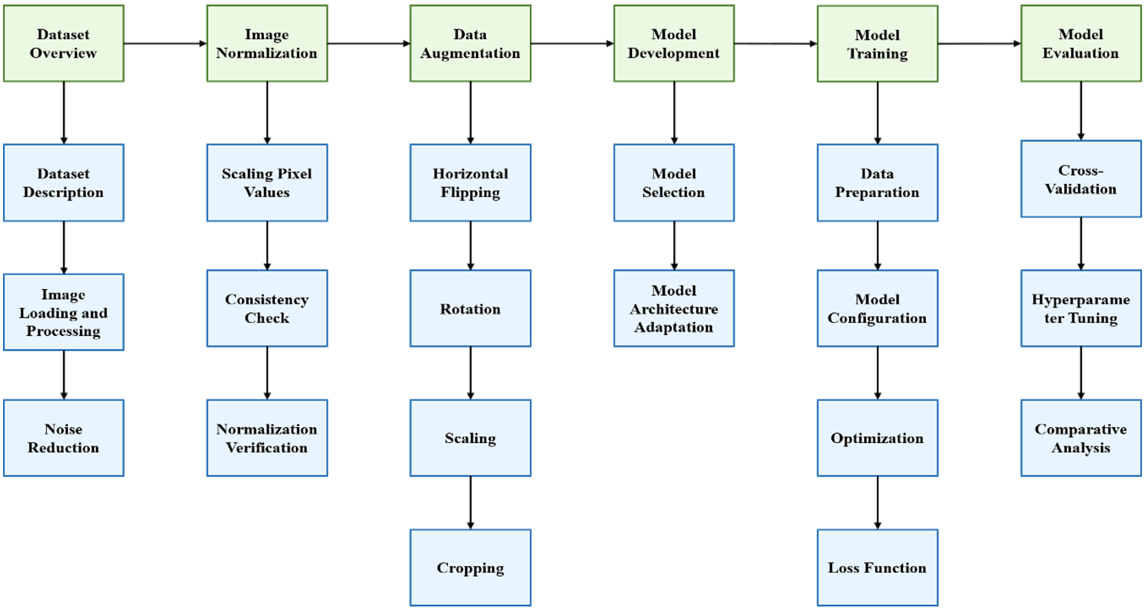


Figure 1. Deep Learning Workflow for Image Processing and Model Evaluation

3.1. Dataset Overview

3.1.1. Dataset Description

The dataset used in this study consists of facial photographs of children categorized into two classes: Autistic and Non-Autistic [4]. It includes a total of 5,000 images, evenly distributed with 2,500 images per class to ensure class balance. Maintaining a balanced dataset helps prevent model bias during training and avoids potential distortions in the learning process that can arise from class imbalance. The dataset adheres to privacy regulations and ethical guidelines; all images are anonymized to protect individual identities, and informed consent was obtained for the use of the data.

3.1.2. Image Loading and Initial Processing

The initial step involves loading images from their respective directories into a numerical format suitable for processing. Each image is read into an array format, representing pixel values in a matrix structure. This matrix is then converted to grayscale to simplify computations and reduce dimensionality. The grayscale conversion is achieved by averaging the color channels, where the pixel value for the grayscale image is. $Igray(a,b)$ at position (a,b) is calculated as:

$$I_{gray}(a,b) = 0.2989. I_R(a,b) + 0.5870. I_G(a,b) + 0.1140. I_B(a,b) \quad (1)$$

where, $I_R(a,b)$, $I_G(a,b)$, and $I_B(a,b)$ are the pixel values of the red, green, and blue channels, respectively.

Grayscale conversion was selected over RGB processing for several reasons: (1) Computational efficiency - reduces model complexity by 67% (from 3 to 1 channel), enabling faster training and inference; (2) Focus on structural features - eliminates color bias and emphasizes facial geometry and texture patterns critical for ASD detection; (3) Robustness to lighting variations - reduces sensitivity to illumination changes common in clinical settings; (4) Literature consistency - aligns with established practices in medical image analysis where structural rather than color information drives diagnosis. While color information such as skin flushing might contain diagnostic value, recent studies demonstrate that facial structural features and geometric relationships provide more reliable ASD biomarkers.

3.1.3. Noise Reduction

Facial images often contain noise that can impair model performance. A median blur filter is applied to reduce noise while preserving edges. The median blur operation substitutes each pixel value with the middle value of pixels within a surrounding window. For a given pixel location (a,b), the median value I(a,b) is computed as:

$$I_{blurred}(a,b) = median\{I(a',b')\} \quad (2)$$

For (a',b') in a 3×3 window around (a,b), where median is the function applied to pixel values within a 3×3 neighborhood.

3.1.4. Image Normalization

Normalization scales pixel values to a range suitable for training. The pixel values, initially varying from 0 to 255, are scaled to [0, 1] using:

$$I_{normalised}(a,b) = \frac{I(a,b)}{255} \quad (3)$$

where, $I(a,b)$ is the original pixel value. Normalization ensures that the input data is consistent and helps stabilize the training process.

3.1.5. Data Augmentation

Data augmentation techniques were used to improve generalization and reduce overfitting. Methods employed include:

Horizontal Flipping: Creates a mirror image of the original image. The horizontal flip operation for a pixel at $(\boldsymbol{x},\boldsymbol{y})$ is:

$$I_{flipped}(a,b) = I(a, W - b - 1) \quad (4)$$

where, W is the width of the image.

Rotation: Images are rotated to simulate different orientations. The rotation matrix R for an angle θ is given by:

$$R = \begin{bmatrix} \cos(\theta) & -\sin(\theta) & (1-\cos(\theta)).ca + \sin(\theta).cb \\ \sin(\theta) & \cos(\theta) & (1-\cos(\theta)).cb - \sin(\theta).ca \end{bmatrix} \quad (5)$$

where, (ca, cb) is the center of the image.

Scaling: Involves resizing images. Scaling transforms each pixel value (x,y) by multiplying by a scale factor α:

$$I_{scaled}(u,v) = I(u.\alpha, v.\alpha) \quad (6)$$

Cropping: Selects random portions of the image. For a cropping window of size $w \times h$, the cropped image $I_{cropped}(u,v)$ is:

$$I_{cropped}(u,v) = I(u + du, v + dv) \quad (7)$$

where, (du, dv) is the offset from the top-left corner of the original image.

Augmentation expands dataset variability, simulating real-world variations in pose, orientation, and facial expression, critical for ensuring robustness in pediatric clinical environments.

3.2. Model Development

3.2.1. Model Selection

Five CNN architectures were evaluated in this study: VGG16, VGG19, ResNet50, InceptionV3, and MobileNet, each offering unique structural advantages for image-based ASD detection. Every architecture possesses distinct design characteristics that influence learning behavior and computational efficiency [23]. VGG architectures (VGG16 and VGG19) employ deep, sequential convolutional layers with uniform 3×3 kernels and interleaved max-pooling layers, enabling hierarchical feature extraction from low-level edges to high-level shapes. This design is particularly effective in capturing fine-grained craniofacial patterns relevant to ASD. They employ the convolution procedure shown below:

$$I_{conv}(u,v) = \sum_{p=0}^{K-1}\sum_{q=0}^{K-1} I(u+p, v+q).K(p,q) \tag{8}$$

where, $K(p,q)$ is the kernel matrix and K denotes the size of the kernel.

MobileNet: utilizes depthwise separable convolutions, significantly reducing computation and parameter count, which supports real-time, on-device ASD screening in resource-constrained clinical environments. The depthwise convolution is defined as:

$$I_{depthwise}(u,v) = \sum_{p=0}^{K-1}\sum_{q=0}^{K-1} I(u+p, v+q).K(p,q) \tag{9}$$

where, $D(p,q)$ is the depthwise convolution kernel, and pointwise convolution follows:

$$I_{pointwise}(u,v) = \sum_{p=0}^{K-1}\sum_{q=0}^{K-1} I_{depthwise}(u,v).P(p,q) \tag{10}$$

where, $P(p,q)$ is the pointwise convolution kernel.

InceptionV3: employs multi-branch inception modules to capture features at multiple receptive field scales simultaneously, making it capable of detecting both local and global facial structures. The inception module output Iinception (x,y) combines different convolutions:

$$I_{inception}(a,b) = concat(conv_{1\times1}, conv_{3\times3}, maxpool_{3\times3}) \tag{11}$$

*ResNet50***:** introduces residual (skip) connections that mitigate the vanishing gradient problem, allowing the network to train effectively even at greater depths. The residual block output $I_{residual}(a,b)$ is:

$$I_{residual}(a,b) = I_{input}(a,b) + F(I_{input}(a,b)) \tag{12}$$

where F represents the residual function.

3.2.2. Model Architecture Adaptation

Each model is adapted for binary classification by modifying the final layers:

VGG16 and VGG19: Add a Global Average Pooling layer, a dense layer with 256 neurons, and a Dropout layer instead of the top completely linked layers. The softmax activation function is used in the final output:

$$p_{class} = \frac{e^{z_{class}}}{\sum_i e^{z_i}} \tag{13}$$

where, z_{class} is the logit for the class.

MobileNet: Similar modifications are made, with a Global Average Pooling layer, a dense layer with 128 neurons, and a Dropout layer. The finishing layer uses a sigmoid activation function:

$$p_{class} = \frac{1}{1+e^{-z_{class}}} \tag{14}$$

where, z_{class} is the logit for the class.

InceptionV3: Adapts similarly with a Global Average Pooling layer, a dense layer with 256 neurons, and a Dropout layer. The terminal layer uses a sigmoid activation function.

ResNet50: Similar adaptations are made, with a Global Average Pooling layer, a dense layer with 128 neurons, and a Dropout layer. The final layer uses a sigmoid activation function [15].

Dense layer sizes (128, 256) were determined through systematic empirical tuning guided by architectural best practices. The 128-neuron configuration was selected for MobileNet and ResNet50 to maintain computational efficiency while providing sufficient representational capacity for binary classification. The 256-neuron configuration for VGG architectures compensates for their simpler feature extraction compared to modern architectures like ResNet and Inception. This follows the principle of progressive dimensionality reduction, where feature maps are gradually compressed to prevent overfitting while preserving discriminative information. Cross-validation experiments confirmed these configurations achieved optimal bias-variance tradeoffs for our dataset size.

3.3. Model Training

The training procedure demands optimizing the model parameters using the Adam optimizer. Adam combines the advantages of two other optimizers, AdaGrad and RMSProp, by adapting the learning rates based on the first and second moments of the gradients. The Adam optimizer is employed to adjust the model's parameters. Adam computes adaptive learning rates for every parameter using estimates of first and second moments of the gradients. The upgrade rule for a parameter θ at time step t is:

$$\theta_{t+1} = \theta_t - \frac{n}{\sqrt{\widehat{v_t}+\epsilon}} \cdot \widehat{m_t} \tag{15}$$

where, η is the learning rate; $\widehat{m_t}$ and $\hat{v}_t$ are bias-corrected estimates of the first and second moments of the gradients; ϵ is a small constant to halt division by zero.

Miniatures are used to train the artists; Each batch consists of forward propagation for forecasting and backpropagation for reweighting the sample according to the gradient of the loss function. The in-depth learning model training covers essential optimization, model building, and data generation techniques. Each method affects the model's stage, so optimizing it is necessary.

3.3.1. Data Preparation

Training and validation sets were generated from preprocessed image data. The validation set helps modify hyperparameters and avoid overfitting, while the training set adjusts the model's parameters. The split is usually 20% for validation and 80% for training for balanced data.

3.3.2. Model Configuration

For each model (VGG19, VGG16, InceptionV3, MobileNet, ResNet50), we utilized pre-trained weights from ImageNet. This transfer learning approach leverages pre-existing knowledge to enhance performance on our specific task. The layers of the pre-trained models were frozen to prevent modification during initial training phases, focusing the training on the added top layers.

3.3.3. Optimisation

The optimization process was guided by minimizing the loss function while maximizing accuracy. The loss function used is categorical cross-entropy, suitable for multi-class classification problems. The categorical cross-entropy loss function is given by:

$$Loss = -\sum_{p=1}^{N} \sum_{q=1}^{C} y_{pq} \log(\widehat{y_{pq}}) \tag{16}$$

where, N is the number of samples, C is the number of classes, y_{pq} is the binary indicator if the class label q is the correct classification for the sample p, and $\widehat{y_{pq}}$ is the predicted probability of the sample p being in class q.

3.3.4. Loss Function

The binary cross-entropy loss function measures the difference between real binary labels and anticipated probability. For a single prediction $\hat{y}$ and the true label y, the binary cross-entropy loss L is defined as:

$$L = -[ylog(\hat{y}) + (1-y)log\,(1-\hat{y})] \tag{17}$$

where, $\hat{y}$ is the anticipated probability of the positive class; y is the true label (0 or 1).

The overall loss is computed as the average of the individual losses across all samples in the batch:

$$L_{avg} = \frac{1}{N} \sum_{p=1}^{N} L_p \tag{18}$$

where, N is the number of samples in the batch and L_p is the loss for the p-th sample

This loss penalizes misclassified samples proportionally to prediction confidence, ideal for medical classification where false negatives (missed ASD cases) are critical to avoid.

4. Results

Several metrics are used to test model performance, including:

Accuracy: The number of correctly classified samples. It is calculated as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{19}$$

Precision: The percentage of all positive forecasts that are true positives:

$$Precision = \frac{TP}{TP+FP} \tag{20}$$

where, TP is the number of true positives and FP is the number of false positives.

Recall: The portion of true positives among all actual positives:

$$Recall = \frac{TP}{TP+FN} \tag{21}$$

where, FN is the number of false negatives.

Recall was emphasized as a key metric since ASD screening prioritizes sensitivity—detecting all possible ASD cases even at the expense of false positives.

F1 Score: The harmonic mean of precision and recall, providing a balanced measure:

$$F1\,Score = \frac{2 \times (Precision \times Recall)}{Precision+Recall} \tag{22}$$

ROC-AUC: The area under the Receiver Operating Characteristic (ROC) curve, which designs the true positive rate against the false positive rate. The ROC-AUC value is computed as:

$$AUC = \int_0^1 TPR(FPR)dFPR \tag{23}$$

where, TPR is the True Positive Rate and FPR is the False Positive Rate

Figure 2 presents the confusion matrices of the evaluated CNN architectures, indicating true-positive, false-positive, true-negative, and false-negative classifications for ASD and non-ASD images.

TP: True Positives (correctly identified ASD cases)

TN: True Negatives (correctly identified non-ASD cases)

FP: False Positives (incorrectly identified ASD cases)

FN: False Negatives (missed ASD cases)

AUC: Area Under the Curve (ROC curve area)

TPR: True Positive Rate (Sensitivity = TP/(TP+FN))

FPR: False Positive Rate (1-Specificity = FP/(FP+TN))

4.1. Cross-Validation

Cross-validation is employed to verify the validity of the models. In every training, the $K - 1$ bunches are used for training while the last bunch is used for validation. So, the given data set K is differentiated into several bundles. Average performance metrics of the clusters are computed to evaluate the efficiency of models and mitigate the effects of data fluctuation.

The K-fold cross-validation process involves:

- Dividing the dataset into K folds.
- Training the model on $K - 1$ folds.
- Validating the model on the remaining fold.
- Repeating steps 2-3 for each fold.
- Averaging the performance metrics across all folds.

4.2. Hyperparameter Tuning

Tuning hyperparameters is essential to maximising model performance. The gaining knowledge of the fee, batch size, range of epochs, and structure-specific factors like the variety of layers and gadgets in the dense layers are essential hyperparameters. The tuning process involves:

- Defining a search space for each hyperparameter.
- Using strategies like random or grid search to investigate various combinations of hyperparameters.
- Applying validation data to assess each combination's performance.
- Based on assessment indicators, choose the combination that produces the best performance.

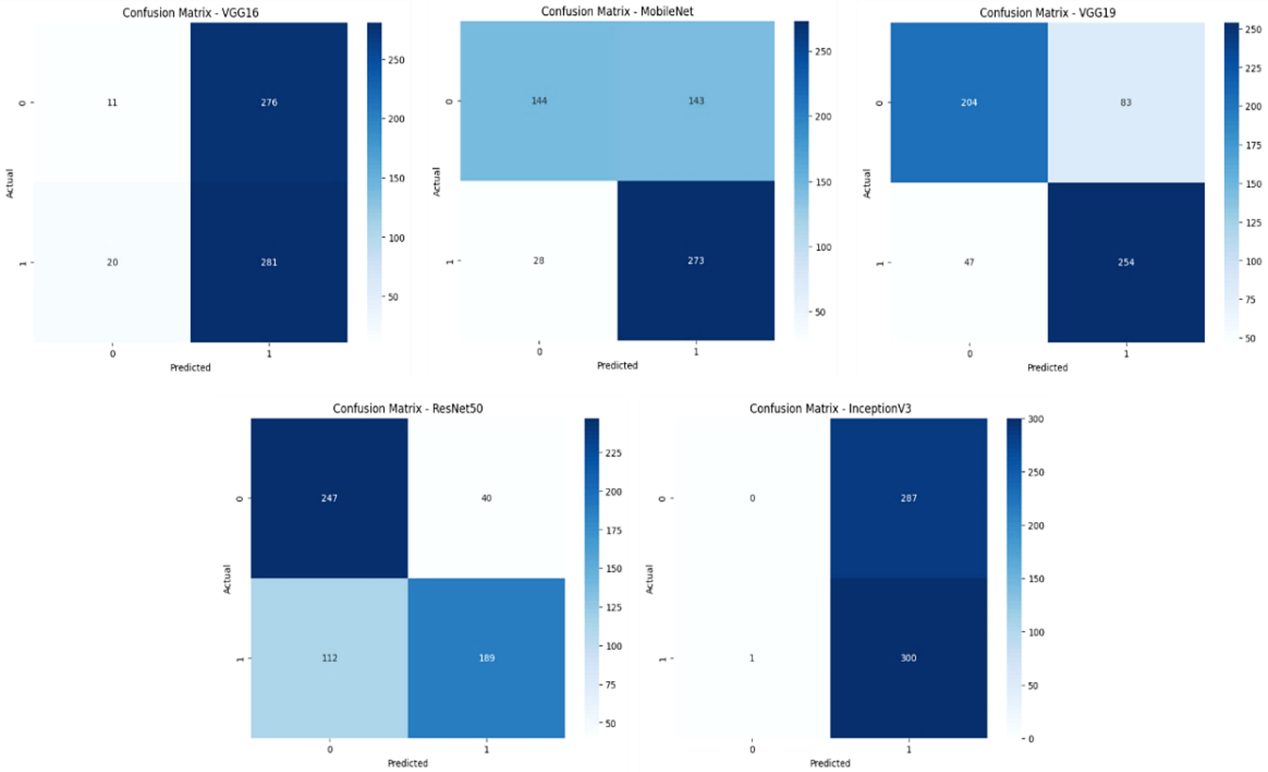


Figure 2. Confusion matrix for DL models

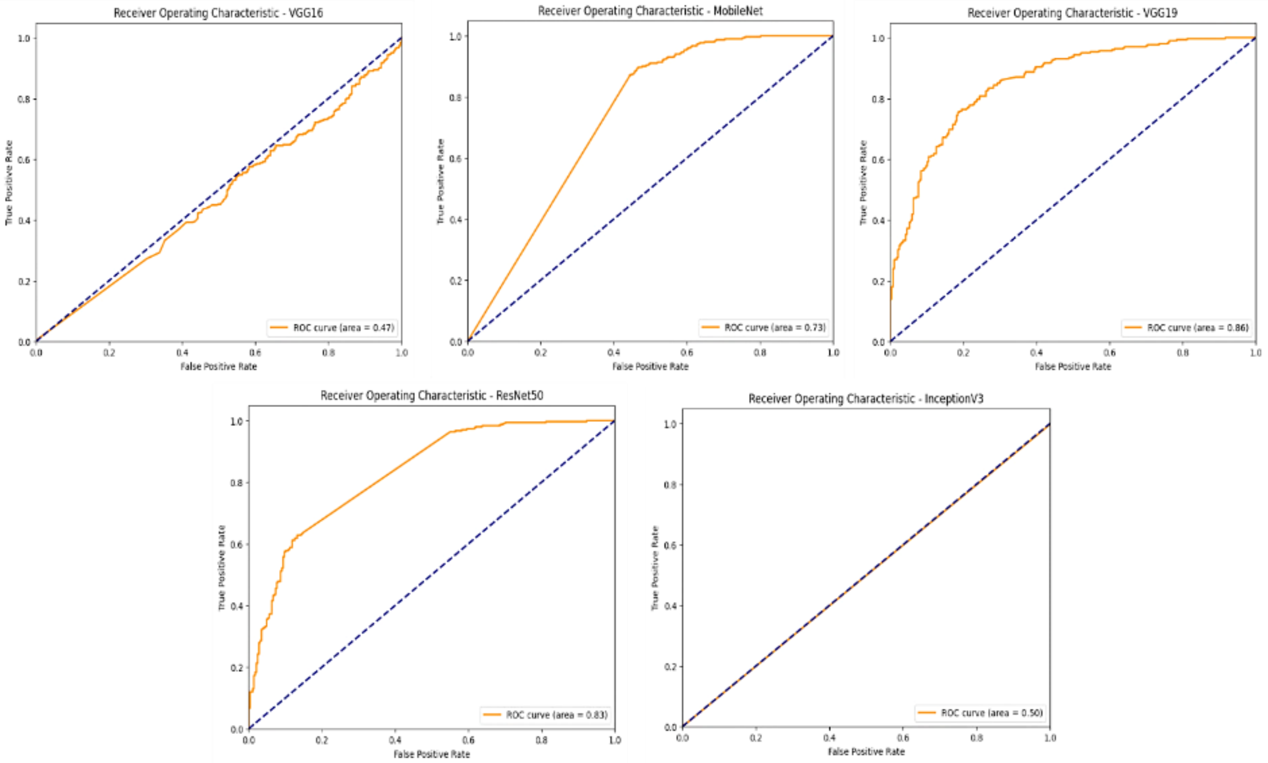


Figure 3. ROC for DL models

4.3. Comparative Analysis

A comparative analysis was conducted to evaluate the performance of five deep learning models: VGG16, VGG19, MobileNet, InceptionV3, and ResNet50. The comparison focused on four key performance metrics: accuracy, precision, recall, and F1-score, as summarized in Table 1.

Table 1. Performance Metrics of Different DL Models for ASD Detection

Model	Accuracy	Precision	Recall	F1 Score
VGG16	0.4966	0.5045	0.9336	0.655
ResNet50	0.7415	0.8253	0.6279	0.7132
InceptionV3	0.5102	0.5111	0.9967	0.6757
VGG19	0.7789	0.7537	0.8439	0.7962
MobileNet	0.7092	0.6563	0.907	0.7615

VGG19 achieved the highest overall accuracy (77.89 %) and F1-score (0.7962), whereas ResNet50 obtained the best precision (82.53 %) and InceptionV3 the best recall (99.67 %). These variations illustrate the trade-off between sensitivity and specificity inherent in different architectural designs.

ROC curve analysis, as shown in Figure 3, further confirmed ResNet50's superior discriminative capability, while VGG19 provided the most balanced overall performance. The comparative metrics establish a solid quantitative foundation for deeper interpretation of each model's diagnostic behavior, discussed next.

4.4. Detailed Performance Analysis and Clinical Implications

Building on the comparative results presented in Section 4.3, this subsection examines how architectural differences translate into diagnostic behavior and clinical relevance. InceptionV3's exceptionally high recall (99.67%) indicates strong sensitivity in detecting ASD cases, critical for early screening, where undetected cases can delay intervention. Its moderate precision (51.11%) implies a higher false-positive rate, acceptable in screening contexts where confirmatory testing follows.

MobileNet's balanced profile (70.92% accuracy, 90.7% recall) makes it ideal for mobile or resource-limited environments. Its lightweight, depthwise separable design enables real-time deployment on portable devices without compromising sensitivity, aligning with current trends in point-of-care AI-based screening.

VGG19's superior overall accuracy (77.89%) demonstrates its ability to capture subtle craniofacial variations associated with ASD. Studies have reported increased intercanthal distance, reduced midline height, and greater facial asymmetry in ASD—features that deep hierarchical layers such as those in VGG19 effectively represent.

ResNet50, achieving the highest precision (82.53%), minimizes false positives—vital in confirmatory diagnostic contexts. Its residual skip connections allow preservation of low- and high-level feature information, improving discrimination between ASD and neurotypical facial structures.

These findings correspond with recent literature: Farhat et al. [5] achieved 97% accuracy using a VGG16–Xception ensemble, while Atlam et al. [21] reported 98.2% accuracy with VGG19 and LIME, benefiting from larger and more diverse datasets. Although our unimodal CNNs achieved slightly lower overall accuracy, the exceptional recall values of InceptionV3 (99.67%) and MobileNet (90.7%) surpass most facial-only studies, confirming their suitability for large-scale ASD screening where sensitivity is prioritized.

5. Discussion

The comparative results revealed that VGG19 and ResNet50 delivered the highest overall performance for ASD classification from facial images. VGG19 achieved a strong balance across all key metrics, accuracy, precision, recall, and F1-score, owing to its deeper architecture, which enhances hierarchical feature learning and allows extraction of subtle facial cues associated with ASD. ResNet50, meanwhile, achieved high precision and accuracy due to its residual connections, which preserve feature information and improve gradient flow, mitigating vanishing gradient issues during training.

InceptionV3 and MobileNet achieved particularly high recall values, indicating strong sensitivity in minimizing false negatives—a desirable property in early ASD screening, where missing a true case has significant clinical implications. Although MobileNet did not achieve the top accuracy, its computational efficiency and lightweight architecture make it ideal for deployment on mobile devices or in resource-constrained diagnostic environments. VGG16, while showing good recall, exhibited weaker generalization due to its shallower depth, confirming that deeper networks like VGG19 and ResNet50 better capture complex facial morphology.

The ROC curve analysis further reinforced ResNet50's robustness, demonstrating the highest AUC and confirming its superior discriminative capability in differentiating between ASD and neurotypical facial patterns [3]. These findings collectively validate that modern CNN architectures—particularly those employing inception modules and residual connectivity—offer substantial improvements for ASD facial-image analysis.

Recent research underscores the importance of multimodal fusion in enhancing ASD diagnostic accuracy. Sun et al. [4] demonstrated that integrating EEG time-frequency features with eye-tracking achieved an AUC of 0.75 in toddlers, while Vidivelli et al. [7] achieved 91.03 % accuracy using a hybrid CNN–BiGRU framework combining facial and EEG modalities. Our facial-only approach, achieving 77.89 % accuracy with VGG19, provides a valuable unimodal baseline for such multimodal systems, establishing a clear reference point for integrating visual and neurophysiological data.

Thermal imaging offers another promising complementary avenue. Ganesh et al. [8] achieved 96 % accuracy using thermal facial responses to emotional stimuli, suggesting that physiological temperature variations may reveal autonomic irregularities characteristic of ASD. Combining our RGB-based structural facial features with thermal modalities could yield more comprehensive models that capture both morphological and physiological signatures of ASD.

Emerging Vision Transformer (ViT) architectures further expand the frontier of ASD detection. Kasri et al. [6] reported 96 % accuracy using ViT–Mamba hybrids, while Mahmood et al. [13] achieved 91.33 % by integrating ViT with ResNet152. Future work should therefore explore hybrid architectures that merge CNN-based feature extraction with transformer self-attention mechanisms to improve contextual understanding and interpretability in ASD diagnostics.

Although multimodal systems (EEG, eye-tracking, thermal imaging) have demonstrated enhanced accuracy, our exclusive focus on facial image analysis was intentional and strategically justified:

- Accessibility and scalability – Facial imaging requires only a standard camera, enabling deployment in low-resource environments.
- Non-invasive and child-friendly – Unlike EEG or eye-tracking, it avoids physical sensor attachments that may distress children with ASD.
- Standardization and reproducibility – Facial imaging protocols are more consistent across clinical and research settings.
- Cost-effectiveness – Eliminates the need for specialized neurophysiological equipment, making large-scale screening feasible.
- Foundational research value – Establishes facial morphology as a reliable standalone biomarker before integration with multimodal systems.

Furthermore, recent federated-learning developments indicate that facial-based deep-learning models can be collaboratively trained across decentralized datasets while preserving patient privacy, supporting the scalability and ethical feasibility of our approach.

Overall, the results affirm that facial image–based deep learning provides a viable foundation for early, accessible, and scalable ASD detection. The insights gained from this comparative analysis serve as a critical step toward integrating multiple sensing modalities, improving explainability, and enabling clinically deployable, privacy-preserving AI screening tools.

6. Conclusion

This study presents a comparative benchmark of five CNN architectures—VGG16, VGG19, ResNet50, InceptionV3, and MobileNet, for detecting Autism Spectrum Disorder (ASD) using facial image analysis. VGG19 and ResNet50 demonstrated superior accuracy and precision, while InceptionV3 and MobileNet achieved exceptionally high recall, highlighting their screening potential. The integration of transfer learning enabled effective performance despite limited data, confirming that facial morphology encodes diagnostically relevant cues. These results establish deep-learning facial analysis as a scalable, non-invasive, and accessible diagnostic approach for early ASD detection. Future work will address dataset diversity, model interpretability through explainable AI, and multimodal fusion with physiological or behavioral data. Incorporating transformer-based architectures will further enhance contextual learning and clinical reliability. This study lays the groundwork for developing ethical, interpretable, and clinically deployable AI-driven autism screening systems.

7. Declarations

7.1. Author Contributions

Conceptualization, U.J.M.; methodology, P.K.P.; software, R.M. and U.J.M.; validation, U.J.M., R.M., and P.K.P.; formal analysis, R.M.; investigation, U.J.M. and P.K.P.; resources, P.K.P. and K.S.M.A.; data curation, R.M. and M.S.; writing—original draft preparation, U.J.M.; writing—review and editing, S.M., K.S.M.A., and P.K.P.; visualization, U.J.M., M.S., and R.M.; supervision, K.S.M.A. and S.M.; project administration, P.K.P. and K.S.M.A.; funding acquisition, K.S.M.A. All authors have read and agreed to the published version of the manuscript.

7.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

7.3. Funding

The authors received financial support for the research, authorship, and/or publication of this article from MMU.

7.4. Institutional Review Board Statement

Not applicable.

7.5. Informed Consent Statement

Not applicable.

7.6. Declaration of Competing Interest

The authors declare that there are no conflicts of interest concerning the publication of this manuscript. Furthermore, all ethical considerations, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies have been completely observed by the authors.

8. References

[1] Moridian, P., Ghassemi, N., Jafari, M., Salloum-Asfar, S., Sadeghi, D., Khodatars, M., Shoeibi, A., Khosravi, A., Ling, S. H., Subasi, A., Alizadehsani, R., Gorriz, J. M., Abdulla, S. A., & Acharya, U. R. (2022). Automatic autism spectrum disorder detection using artificial intelligence methods with MRI neuroimaging: A review. Frontiers in Molecular Neuroscience, 15, 999605. doi:10.3389/fnmol.2022.999605.

[2] Alkahtani, H., Aldhyani, T. H. H., & Alzahrani, M. Y. (2023). Deep Learning Algorithms to Identify Autism Spectrum Disorder in Children-Based Facial Landmarks. Applied Sciences (Switzerland), 13(8), 4855. doi:10.3390/app13084855.

[3] Guha, T., Yang, Z., Grossman, R. B., & Narayanan, S. S. (2018). A Computational Study of Expressive Facial Dynamics in Children with Autism. IEEE Transactions on Affective Computing, 9(1), 14–20. doi:10.1109/TAFFC.2016.2578316.

[4] Sun, B., Calvert, E. I., Ye, A., Mao, H., Liu, K., Wang, R. K., Wang, X. Y., Wu, Z. L., Wei, Z., & Kong, X. J. (2024). Interest paradigm for early identification of autism spectrum disorder: an analysis from electroencephalography combined with eye tracking. Frontiers in Neuroscience, 18(1502045). doi:10.3389/fnins.2024.1502045.

[5] Farhat, T., Akram, S., Rashid, M., Jaffar, A., Bhatti, S. M., & Iqbal, M. A. (2025). A deep learning-based ensemble for autism spectrum disorder diagnosis using facial images. PLOS One, 20(4 April), 0321697. doi:10.1371/journal.pone.0321697.

[6] Kasri, W., Himeur, Y., Copiaco, A., Mansoor, W., Albanna, A., & Eapen, V. (2025). Hybrid Vision Transformer-Mamba Framework for Autism Diagnosis via Eye-Tracking Analysis. In International Conference on Communication, Computing, Networking, and Control in Cyber-Physical Systems, CCNCPS 2025, 343–348. doi:10.1109/CCNCPS66785.2025.11135843.

[7] Vidivelli, S., Padmakumari, P., & Shanthi, P. (2025). Multimodal autism detection: Deep hybrid model with improved feature level fusion. Computer Methods and Programs in Biomedicine, 260, 108492. doi:10.1016/j.cmpb.2024.108492.

[8] Ganesh, K., Umapathy, S., & Thanaraj Krishnan, P. (2021). Deep learning techniques for automated detection of autism spectrum disorder based on thermal imaging. Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine, 235(10), 1113–1127. doi:10.1177/09544119211024778.

[9] Ahmadiar, A., Melinda, M., Muthiah, Z., Zainal, Z., & Mina Rizky, M. (2025). Thermal Image Classification of Autistic Children Using Res-Net Architecture. Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics, 7(1), 1–10. doi:10.35882/365fkd59.

[10] Tripi, G., Roux, S., Matranga, D., Maniscalco, L., Glorioso, P., Bonnet-Brilhault, F., & Roccella, M. (2019). Cranio-facial characteristics in children with autism spectrum disorders (ASD). Journal of Clinical Medicine, 8(5), 641. doi:10.3390/jcm8050641.

[11] Sahu, R., Pattnaik, P. K., Anbananthen, K. S. M., & Muthaiyah, S. (2025). Identification of Depression Patients Using LIF Spiking Neural Network Model From the Pattern of EEG Signals. IEEE Access, 13, 55156–55168. doi:10.1109/ACCESS.2025.3552619.

[12] Monani, U. J., Samanta, S., Gourisaria, M. K., & Das, S. (2024). Efficiency Analysis of CNN through Different Filters for Medical Image Classification. 2nd IEEE International Conference on Data Science and Information System, ICDSIS 2024, 1–7. doi:10.1109/ICDSIS61070.2024.10594018.

[13] Mahmood, M. A., Jamel, L., Alturki, N., & Tawfeek, M. A. (2025). Leveraging artificial intelligence for diagnosis of children autism through facial expressions. Scientific Reports, 15(1), 8743. doi:10.1038/s41598-025-96014-6.

[14] McCarty, P., & Frye, R. E. (2020). Early Detection and Diagnosis of Autism Spectrum Disorder: Why Is It So Difficult? Seminars in Pediatric Neurology, 35, 100831. doi:10.1016/j.spen.2020.100831.

[15] Daniels, A. M., Halladay, A. K., Shih, A., Elder, L. M., & Dawson, G. (2014). Approaches to enhancing the early detection of autism spectrum disorders: A systematic review of the literature. Journal of the American Academy of Child and Adolescent Psychiatry, 53(2), 141–152. doi:10.1016/j.jaac.2013.11.002.

[16] Dawson, G., & Bernier, R. (2013). A quarter century of progress on the early detection and treatment of autism spectrum disorder. Development and Psychopathology, 25(4 Part 2), 1455–1472. doi:10.1017/S0954579413000710.

[17] Khodatars, M., Shoeibi, A., Sadeghi, D., Ghaasemi, N., Jafari, M., Moridian, P., Khadem, A., Alizadehsani, R., Zare, A., Kong, Y., Khosravi, A., Nahavandi, S., Hussain, S., Acharya, U. R., & Berk, M. (2021). Deep learning for neuroimaging-based diagnosis and rehabilitation of Autism Spectrum Disorder: A review. Computers in Biology and Medicine, 139, 104949. doi:10.1016/j.compbiomed.2021.104949.

[18] Elshoky, B. R. G., Younis, E. M. G., Ali, A. A., & Ibrahim, O. A. S. (2022). Comparing automated and non-automated machine learning for autism spectrum disorders classification using facial images. ETRI Journal, 44(4), 613–623. doi:10.4218/etrij.2021-0097.

[19] Alam, M. S., Rashid, M. M., Faizabadi, A. R., Mohd Zaki, H. F., Alam, T. E., Ali, M. S., ... & Ahsan, M. M. (2023). Efficient deep learning-based data-centric approach for autism spectrum disorder diagnosis from facial images using explainable AI. Technologies, 11(5), 115. doi:10.3390/technologies11050115.

[20] Anbananthen, S. K., Sainarayanan, G., Chekima, A., & Teo, J. (2006). Data Mining using Pruned Artificial Neural Network Tree (ANNT). Proceedings of the 2nd IEEE International Conference on Information & Communication Technologies (ICT), 1350–1356. doi:10.1109/ictta.2006.1684577.

[21] Atlam, E. S., Aljuhani, K. O., Gad, I., Abdelrahim, E. M., Atwa, A. E. M., & Ahmed, A. (2025). Automated identification of autism spectrum disorder from facial images using explainable deep learning models. Scientific Reports, 15(1), 26682. doi:10.1038/s41598-025-11847-5.

[22] K` Simon, S., Sonai Muthu Anbananthen, K., & Lee, S. (2013). A Ubiquitous Personal Health Record (uPHR) Framework. Proceedings of the 2013 International Conference on Advanced Computer Science and Electronics Information (ICACSEI 2013), 105. doi:10.2991/icacsei.2013.105.

[23] Lozier, L. M., Vanmeter, J. W., & Marsh, A. A. (2014). Impairments in facial affect recognition associated with autism spectrum disorders: A meta-analysis. Development and Psychopathology, 26(4), 933–945. doi:10.1017/S0954579414000479.