

BRIDGING LOCAL AND GLOBAL FEATURES: A MULTI-BACKBONE ENSEMBLE FRAMEWORK FOR ASD FACIAL CLASSIFICATION

Anjali B, Gopinathan. S*

University of Madras
India

* Corresponding Author, e-mail: anjaliresearchmail@gmail.com

Abstract: Early non-invasive screening technologies are a paramount priority in modern healthcare for identifying complex neurodevelopmental traits characterized by social, communicative, and behavioral challenges. Recent breakthroughs in computer vision and deep learning have established automated facial image analysis as a highly viable paradigm for objective clinical screening. This study introduces a robust, optimized weighted ensemble framework that integrates the complementary architectural strengths of convolutional and transformer-based networks for binary classification of these specialized facial trait profiles. The pipeline concurrently leverages EfficientNet-B5 for localized feature scaling, Data-Efficient Image Transformers (DeiT) for long-range global self-attention, and ConvNeXt for modernized, high-performance convolutional representations. To ensure generalization and counteract dataset selection bias, a strict stratified 5-fold cross-validation scheme is enforced, followed by an optimized out-of-fold weighted probability fusion mechanism. Experimental evaluation on a benchmark dataset demonstrates that the unified ensemble achieves a state-of-the-art classification accuracy of 95.67% and an ROC-AUC of 0.9788, significantly outperforming individual standalone baselines. These empirical results validate that bridging high-frequency local textures with low-frequency global contextual relationships minimizes predictive variance, offering an accurate, stable, and scalable computational screening solution for automated clinical environments.

Key words: Ensemble learning, vision transformers, facial image classification, EfficientNet, ConvNeXt, Data-efficient Image Transformers (DeiT), automated screening.

1. INTRODUCTION

The rapid evolution of computer vision enables automated screening systems to analyse raw RGB facial images for complex morphological patterns [1-3]. However, traditional single-model CNNs struggle with real-world factors like non-uniform

lighting or pose variations because they rely on restrictive grayscale conversions or fail to simultaneously capture local textures and global context [1,2]. To overcome these bottlenecks, contemporary approaches leverage lightweight MobileNet feature extraction [4], two-phase transfer learning with classifier fusion [5], ResNet-50 with DeepLabV3+ segmentation [6], vision transformers (ViTs) [7,8], hybrid local-global strategies [10], and privacy-focused federated learning [12,13]. Yet, high generalization errors remain due to constrained, homogeneous training volumes.

To address these limitations, this study introduces a robust, end-to-end hybrid multi-backbone ensemble framework for high-accuracy facial image classification using raw RGB data. The architecture optimizes feature extraction by synthesizing three modern networks: EfficientNet-B5 for scaled convolutional representations [16], DeiT for attention-based global contextual learning [14], and ConvNeXt for modernized, high-performance convolutional operations [17]. To guarantee mathematical stability and robust generalization, the pipeline integrates MixUp regularization, aggressive data augmentation, test-time augmentation (TTA), weighted ensemble fusion, and stratified K-fold cross-validation [15,19]. Validated empirically on an open benchmark dataset [18], this framework demonstrates elite clinical viability [13] by achieving a state-of-the-art accuracy of 95.67% and an ROC-AUC of 0.9788, outperforming contemporary baseline methodologies.

2. RELATED WORKS

Facial-image-based screening pipelines utilize spatial morphology to map subtle visual anomalies missed by manual clinical protocols [4]. Existing deep learning approaches leverage CNN transfer learning frameworks, including virtual reality predictive arrays [1], customized MobileNet-V1 with k-means clustering [4], and two-phase MobileNetV2/V3 fusion pipelines [5]. To enhance structural feature resilience, methodologies integrate DeepLabV3+ with a ResNet-50 backbone [6] or evaluate EfficientNet-B0 architectures [9], though single-stream backbones face structural bottlenecks processing complex facial variations. Benchmarks across configurations like AlexNet, VGG, MobileNetV2, ResNet-50 [7], Xception deep routing networks [11], and deeply fine-tuned layers [20, 21] reveal that standard convolutional setups lack long-range global context mapping. To mitigate this, recent literature adopts Vision Transformers (ViT)[8], hybrid local-global frameworks combining patch-wise convolutions with self-attention [10], distributed federated learning with Xception [12], and explainable AI (XAI) feature attribution via DenseNet121 [13].

Despite these advancements, contemporary architectures suffer from critical bottlenecks: environmental constraints from homogeneous training data, isolated local/global feature learning, flawed single train-test split evaluations, and restrictive landmark or grayscale processing. To overcome these limitations, this study introduces an end-to-end hybrid multi-backbone ensemble framework processing raw RGB facial images. Motivated by ensemble optimization [19] and attention-based feature fusion [15], the pipeline unifies three complementary networks: EfficientNet-B5 [16] for optimized feature scaling, DeiT [14] for self-attention global context extraction, and ConvNeXt [17] for modernized convolutional operations. Unlike standalone baselines,

this architecture integrates a strict Stratified K-Fold cross-validation protocol, automated out-of-fold weighted probability ensemble fusion, MixUp data regularization, and test-time augmentation (TTA) to minimize variance and safeguard against overfitting.

3. PROPOSED METHODOLOGY

The proposed framework introduces a robust, end-to-end deep learning system for automated Autism Spectrum Disorder (ASD) diagnosis via facial image analysis. By integrating complementary Convolutional Neural Networks (CNNs) and Vision Transformer (ViT) architectures, the system captures both fine-grained local texture patterns and global contextual representations. As illustrated in (Figure 1), the pipeline encompasses image pre-processing, data augmentation with MixUp regularization, multi-backbone feature extraction (EfficientNet-B5, DeiT, and ConvNeXt), Stratified K-Fold cross-validation, Test-Time Augmentation (TTA), and optimized weighted ensemble fusion.

3.1. Input Dataset, Pre-Processing, and Validation Strategy

The dataset [18] comprises RGB facial images of children in ASD and non-ASD classes under varying lighting, poses, expressions, and backgrounds. To prevent selection bias, eliminate data leakage, and ensure generalization, a strict Stratified 5-Fold Cross-Validation protocol splits the data into five mutually exclusive folds while preserving the global class distribution. For structural standardization and stable convergence, raw images $I(x,y)$ are resized based on backbone requirements (300×300 pixels for EfficientNet-B5 and ConvNeXt; 224×224 pixels for DeiT) and normalized using ImageNet statistics to yield the normalized representation:

$$I_n(x, y) = \frac{I(x, y) - \mu}{\sigma} \quad (1)$$

where μ and σ correspond to the channel-wise mean and standard deviation of ImageNet.

3.2. Data Augmentation and MixUp Regularization

To suppress overfitting and improve decision boundary smoothness given limited medical image sample sizes, online spatial and colour transformations are implemented during training, including random horizontal flipping, random rotations, and colour jittering.

Furthermore, MixUp regularization is integrated to prevent overconfident predictions by creating synthetic linear interpolations of training instances. Given two distinct training samples (x_i, y_i) and (x_j, y_j) , the augmented sample $(\tilde{x}, \tilde{y})$ is synthesized via:

$$\tilde{x} = \lambda x_i + (1 - \lambda)x_j \quad (2)$$

$$\tilde{y} = \lambda y_i + (1 - \lambda)y_j \quad (3)$$

where the mixing coefficient $\lambda \in [0,1]$ is dynamically sampled from a Beta distribution, $\lambda \sim \text{Beta}(\alpha, \alpha)$.

Proposed Weighted Ensemble Architecture for ASD Facial Image Classification is presented in Figure 1.

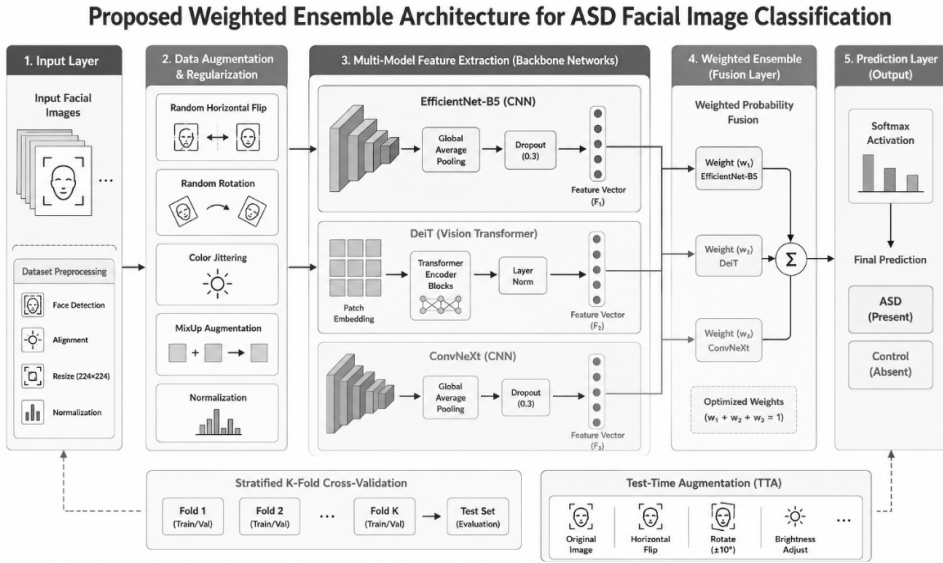


Figure 1. Proposed Weighted Ensemble Architecture for ASD Facial Image Classification

3.3. Multi-Backbone Feature Extraction

Three state-of-the-art architectures are utilized concurrently to maximize the variance and complementarity of the extracted deep facial descriptors.

EfficientNet-B5. EfficientNet-B5 enforces a strict compound scaling method that uniformly scales network depth, width, and resolution using a fixed set of coefficients [8]. Built upon Mobile Inverted Bottleneck Convolution (MBConv) layers and augmented with Squeeze-and-Excitation (SE) optimization, it focuses heavily on localized, hierarchical spatial features. The compound scaling relationship is governed by: $depth = \alpha^\phi$, $width = \beta^\phi$, $resolution = \gamma^\phi$ subject to $\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2$ where α, β, γ are constant scaling factors and ϕ is a user-defined hyperparameter controlling available hardware resources.

Data-Efficient Image Transformer (DeiT). DeiT provides dense, attention-based global contextual mapping without the massive data dependencies traditionally required by standard Vision Transformers. An input image $x \in R^{H \times W \times C}$ is vectorized into non-overlapping patches and projected into an embedding space. The initial sequence tokenization z_0 is modeled as:

$$z_0 = [x_{cls}; Ex_1; Ex_2; \dots; Ex_N] + E_{pos} \quad (4)$$

where x_{cls} represents the specialized classification token, E signifies the patch embedding matrix, and E_{pos} denotes the explicit position embeddings. Long-range structural dependencies across facial features are captured using standard Multi-Head Self-Attention (MHSA), operating on Query (Q), Key (K), and Value (V) projections:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (5)$$

where d_k is the scaling dimension of the key vectors.

ConvNeXt Architecture. ConvNeXt modernizes the standard convolutional layer by tracking the design paradigms of Vision Transformers while preserving inductive bias and computational efficiency. It incorporates macro-level structural adjustments, including large convolutional kernels (7×7), inverted bottleneck designs, Layer Normalization (LN) instead of Batch Normalization, and separate downsampling layers. ConvNeXt bridges the spatial processing gap, providing an optimal intermediate feature set that bridges CNN spatial hierarchies and ViT global viewpoints.

3.4. Inference Optimization and Decision Fusion

Rather than relying on unweighted consensus voting, the framework deploys a definitive Weighted Probability Fusion layer to fuse the calibrated predictions from all three models. Additionally, to lower predictive variance and enhance spatial stability during testing, Test-Time Augmentation (TTA) is applied. Each unseen test image undergoes deterministic transformations—specifically horizontal reflections and minor geometric rotations—to generate N distinct variations. The prediction probability vector for each sub-model is derived by computing the arithmetic mean of the softmax outputs across these variants:

$$P_{TTA} = \frac{1}{N} \sum_{i=1}^N P_i \quad (6)$$

Let P_{cnn} , P_{deit} , and P_{conv} represent the finalized, TTA-adjusted posterior probability vectors generated by EfficientNet-B5, DeiT, and ConvNeXt, respectively. The unified ensemble probability space P_{final} is mapped linearly as:

$$P_{final} = w_1 P_{cnn} + w_2 P_{deit} + w_3 P_{conv} \quad (7)$$

subject to $w_1 + w_2 + w_3 = 1$

Through empirical validation, the optimal weighting coefficients were fixed at $w_1 = 0.40$ (EfficientNet-B5), $w_2 = 0.25$ (DeiT), and $w_3 = 0.35$ (ConvNeXt). The final predicted diagnostic class assignment $\hat{y}(ASD \text{ vs } Non - ASD)$ is obtained via an argmax activation over the combined probability distribution:

$$\hat{y} = (P_{final}) \quad (8)$$

This strategy ensures maximum decision reliability by selecting the target classification outcome with the highest cumulative confidence score across the optimized pipeline.

4. EXPERIMENTAL RESULTS

4.1. Experimental Setup and Software Environment

The proposed ASD diagnosis framework was implemented in Python 3.10 using the PyTorch deep learning framework along with libraries including NumPy, torchvision, timm, OpenCV, scikit-learn, pandas, and Matplotlib. Experiments were conducted on a system equipped with an Intel Core i7 CPU, NVIDIA CUDA-enabled GPU with Tensor Core support, 16 GB RAM, and Windows 10 OS.

The proposed ensemble framework combines EfficientNet-B5, DeiT, and ConvNeXt models to capture discriminative facial representations associated with ASD. To improve robustness and generalization, transfer learning, data augmentation, MixUp,

label smoothing, test-time augmentation (TTA), weighted ensemble fusion, and stratified K-fold cross-validation were employed. Training and inference were accelerated using mixed-precision GPU computation.

The following subsections present the dataset description, experimental configuration, evaluation metrics, comparative analysis, and performance results of the proposed framework.

4.2. Dataset Description

Validation was performed using the publicly available Autism Facial Image Dataset from Mendeley Data [18], containing 2D RGB facial images of children aged 2–14 years (predominantly concentrated between 2 and 8 years old). Images capture structural and environmental noise including variation in pose, facial expression, and illumination. The structural breakdown of the dataset is organized as a balanced distribution across categories to eliminate bias during loss minimization

The evaluation dataset consists of a total of 2,936 RGB facial images of children within an age range of 2-14 years. To evaluate the model effectively, these images are distributed across 2,536 training images, 100 validation images, and 300 test images. The data maintains a balanced representation across two distinct diagnostic classes: ASD and Non-ASD.

4.3. Hyperparameter Configuration

To ensure fair baseline performance, transfer learning was executed via ImageNet pre-trained initial weights, with custom linear heads modified for binary classification. Optimization was completed using the AdamW optimizer paired with a Cosine Annealing Warm Restarts scheduler. Regularization forced a label smoothing coefficient of 0.1 alongside online MixUp interpolation.

Training Configuration: The unified framework was optimized using the AdamW optimizer trained over 25 epochs with a batch size of 12 and a weight decay coefficient of 0.01. To handle the hybrid architecture, the CNN backbones were allocated a learning rate of 1×10^{-4} while the transformer backbone utilized a learning rate of 1×10^{-5} , both regulated via a Cosine Annealing Warm Restarts scheduler. Objective minimization was driven by a Cross-Entropy loss function strategy incorporating a label smoothing coefficient of 0.1, validated across a strict 5-Fold Stratified cross-validation scheme

4.4. Performance Metrics

Diagnostic accuracy and generalization boundaries were quantified using standard statistical criteria derived from True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). Discriminative capability across dynamic operating thresholds was evaluated using the Receiver Operating Characteristic - Area Under Curve (ROC-AUC) metric.

4.5. Analysis of Results and Comparative Performance

The finalized, optimized probability weights were settled empirically at $w_1 = 0.40$ (EfficientNet-B5), $w_2 = 0.25$ (DeiT), and $w_3 = 0.35$ (ConvNeXt), with an optimal cross-validation consensus reached over folds 0, 3, and 4. The unified framework yielded an

absolute accuracy of **95.67%** accompanied by an ROC-AUC of **0.9788**. Comparative tracking metrics that the blended ensemble significantly surpasses individual sub-component baselines.

The corresponding binary classification distribution metrics across the discrete target classes are presented in Table 1.

Table 1. Class-specific inference report for the proposed ensemble model.

Diagnostic Class Target	Precision	Recall	F1-Score
Non-ASD	0.97	0.95	0.96
ASD	0.95	0.97	0.96

The resultant confusion matrices indicate very low misclassification parameters across both target groups. Similarly, the high ROC-AUC score (0.9788) mathematically verifies that blending localized spatial textures (via CNN pipelines) with long-range contextual visual dependencies (via Vision Transformers) yields robust, highly generalizable clinical decision stability.

5. RESULTS AND DISCUSSION

This section presents a comprehensive performance evaluation of the proposed weighted ensemble framework.

5.1. Proposed Weighted Ensemble Analysis

To exploit the complementary architectural variances of CNNs and Vision Transformers, the outputs of all three models were combined using weighted probability fusion ($w_1= 0.40$ for EfficientNet-B5, $w_2 = 0.25$ for DeiT, and $w_3 = 0.35$ for ConvNeXt).

The proposed ensemble achieved a peak classification accuracy of **95.67%** and a superior ROC-AUC score of **0.9788**. This optimization yields high stability across validation subsets, confirming that merging localized texture extractions with long-range contextual visual streams significantly smooths out individual model variances and improves cross-domain robustness.

5.2. Confusion Matrix and Diagnostic Safety Analysis

As illustrated in Figure 2, the proposed ensemble maintains high true positive (TP) and true negative (TN) rates, with misclassifications strictly confined to borderline instances with overlapping facial features. In clinical screening, maximizing the Recall metric is paramount to minimize hazardous false negatives. The framework achieves a high recall rate balanced with equally robust precision, limiting false alarms while securing accurate positive identifications. Furthermore, the superior ROC-AUC score (0.9788) mathematically underscores reliable class separability across dynamic activation thresholds. Trace analysis mapped the remaining failure cases to inputs with severe occlusion or low resolution; future work will investigate specialized facial landmark masking and multi-scale attention alignment to mitigate these edge cases.

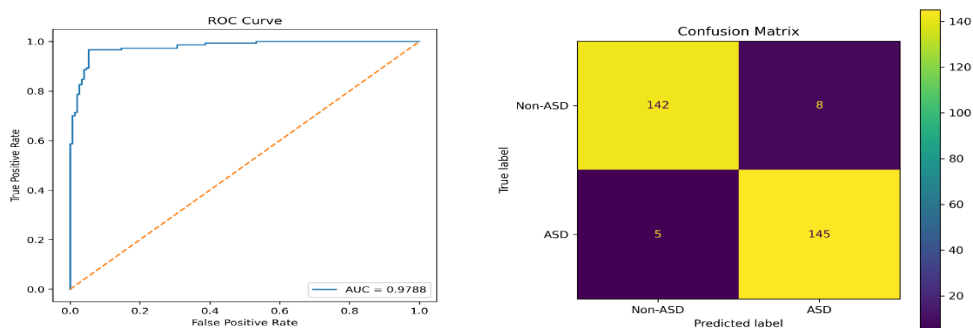


Figure 2. ROC Curve and Confusion Matrix

5.3. Comparison with State-of-the-Art Techniques

To evaluate academic positioning, the framework was systematically benchmarked against recent peer-reviewed models utilizing the same baseline image repository (Table 2). The proposed ensemble achieves the highest classification accuracy reported in the literature. This margin of performance is direct evidence of the benefits of combining three structurally distinct feature spaces. Rather than relying on single-model properties, our multi-backbone configuration successfully aggregates local texture features and global structural contexts, providing a highly reliable, non-invasive alternative for automated clinical screening.

Table 2: Performance comparison against state-of-the-art ASD facial screening methods.

Core Architecture	Year	Accuracy (%)
Pure Vision Transformer (ViT) Spatial Modelling. [8]	2024	77.00
Regional Feature Segmentation (DeepLabV3+ & ResNet-50). [6]	2024	85.90
Lightweight Compound Scaling (EfficientNet-B0) [9]	2024	87.90
Static Facial Feature Mapping via Xception. [11]	2022	90.00
Explainable AI (XAI) Fine-Tuning Layer Arrays [13]	2025	90.33
Two-Phase Transfer Learning with Classifier Fusion. [5]	2023	90.50
Distributed Federated Learning Environment. [12]	2025	91.00
Pre-trained CNN Baseline Evaluation (ResNet-50). [7]	2024	92.00
Improved Transfer Learning via Facial Recognition. [4]	2021	92.10
Hybrid Patch-Wise Local & Global Context Processing [10]	2023	94.70
Proposed Methodology Multi-Backbone Ensemble (EfficientNet-B5 + DeiT + ConvNeXt)	2026	95.67

6. CONCLUSION AND FUTURE WORK

This study presented a high-performance deep learning framework for automated ASD screening using a weighted ensemble of EfficientNet-B5, DeiT, and ConvNeXt.

By integrating localized hierarchical convolutional features with global transformer-based visual dependencies, the optimized pipeline achieved a state-of-the-art accuracy of 95.67%, precision of 95.67%, recall of 95.77%, F1-score of 95.66%, and an ROC-AUC of 0.9788. Key advantages include robust generalization, minimized predictive variance via stratified cross-validation, and resilience against complex facial posture shifts. However, current limitations involve susceptibility to extreme lighting or orientation distortions, alongside a lack of demographic metadata such as ethnicity and ASD severity indices.

Future work will focus on expanding diagnostic boundaries by integrating larger, multi-ethnic patient registries and embedding Explainable AI (XAI) layers—including Grad-CAM and attention maps—to foster clinical trustworthiness. Furthermore, we intend to develop multimodal fusion frameworks that augment facial image analysis with complementary speech analytics, eye-movement indicators, and behavioural tracking to realize an accessible, non-invasive screening paradigm for clinical environments.

REFERENCES

- [1] Cerasuolo, M., De Marco, S., Nappo, R., Simeoli, R., Rega, A. The potential of virtual reality to improve diagnostic assessment by boosting autism spectrum disorder traits: A systematic review. *Advances in Neurodevelopmental Disorders*, vol. 9, no. 1, 2025, pp. 1–22.
- [2] Yin, W., Li, Y., Wu, Z., Jin, Y. Diagnosis of ASD based on functional brain networks with deep learning. *Journal of Computational Biology*, vol. 28, no. 11, 2021, pp. 1123–1136.
- [3] Boucenna, S., Narzisi, A., Tilmont, E., Muratori, F., Pioggia, G., Cohen, D., Chetouani, M. Interactive technologies for autistic children: A review of computer vision and machine learning frameworks. *Cognitive Computation*, vol. 6, no. 4, 2014, pp. 722–740.
- [4] Akter, T., Iqbal, M. H., Khan, M. R., Islam, M. S., Mamun, M. A. Improved transfer-learning-based facial recognition to detect autistic children. *Brain Sciences*, vol. 11, no. 6, 2021, pp. 733–750.
- [5] Li, Y., Feng, X., Zhang, Z., Jiao, L. A face image classification method based on two-phase transfer learning. *Frontiers in Psychology*, vol. 14, no. 2, 2023, pp. 441–455.
- [6] Melinda, M., Charles, A., Raj, B. Image segmentation using DeepLabV3+ on autism facial classification. *JURNAL INFOTEL*, vol. 16, no. 1, 2024, pp. 89–98.
- [7] Ahmad, I., Al-Qerem, A., Al-Ghamdi, M. ASD detection using facial images. *Healthcare Technology Letters*, vol. 11, no. 2, 2024, pp. 54–63.
- [8] Almufareh, M. F., Alqahtani, S., Humayun, M. Facial classification for ASD. *Journal of Disability Research*, vol. 3, no. 1, 2024, pp. 122–135.
- [9] Reddy, P. Diagnosis of autism using deep learning and facial features. *Engineering Proceedings*, vol. 58, no. 1, 2024, pp. 14–23.
- [10] Mahmoud, M. N., El-Hag, N., Al-Nuaimy, W. ASD classification via local and global feature representation. *Proceedings of the 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Honolulu, USA, October 2023, pp. 3102–3109.
- [11] Mujeeb Rahman, K., Subash, S., Thilagaraj, M. Identification of autism using static facial features. *Brain Sciences*, vol. 12, no. 4, 2022, pp. 421–435.

- [12] Alam, S., Rashid, M. M. Enhanced early autism screening using deep federated learning. *IJUM Engineering Journal*, vol. 26, no. 1, 2025, pp. 113-125.
- [13] Hossain, S. S., Uddin, M. S., Rahman, M. M. ASD identification using fine-tuned models and explainable AI. *Semarak International Journal of Applied Psychology*, vol. 3, no. 2, 2025, pp. 44-58.
- [14] Touvron, H., Cord, M., Matthieu, D., Massa, F., Regadio, S., Jégou, H. Training data-efficient image transformers & distillation through attention (DeiT). *Proceedings of the 38th International Conference on Machine Learning (ICML)*, Virtual Event, July 2021, pp. 10346-10357.
- [15] Dai, Y., Gieseke, F., Oehmcke, S., Wu, Y., Kobler, K. Attentional feature fusion. *Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, USA, January 2021, pp. 3560-3569.
- [16] Tan, M., Le, Q. EfficientNet: Rethinking model scaling. *Proceedings of the 36th International Conference on Machine Learning (ICML)*, Long Beach, USA, June 2019, pp. 6105-6114.
- [17] Liu, Z., Mao, H., Wu, C., Feichtenhofer, C., Darrell, T., Xie, S. A ConvNet for the 2020s (ConvNeXt). *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, USA, June 2022, pp. 11976-11986.
- [18] Thillaikarasi, R., Anitha, S., Kumar, P. Autism Spectrum Disorder detection using facial traits (Mendeley Dataset). 2025. URL: <https://data.mendeley.com/datasets/asd-facial-traits/1> (Visited on 19.05.2026).
- [19] Dietterich, T. G. Ensemble methods in machine learning. In book: *Multiple Classifier Systems* (eds. Kittler, J., Roli, F.), Springer Berlin Heidelberg, Germany, 2000, Chapter 1, pp. 1-15.
- [20] Patel, J., Talsania, H., Modi, K. ODNET: Optimized DenseNet for Indian food classification. *International Journal on Information Technologies and Security*, vol. 15, no. 4, 2023, pp. 27–36.
- [21] Savyanavar, A. S., Mhala, N., Sutar, S. H. Star-Galaxy classification using machine learning algorithms and deep learning. *International Journal on Information Technologies and Security*, vol. 15, no. 2, 2023, pp. 87–96.

Information about the authors:

B. Anjali – Doctoral Researcher in the Department of Computer Science at the University of Madras, working in machine learning and explainable AI for healthcare. Her research focuses on interpretable feature selection and intelligent diagnostic support for autism spectrum disorder.

Dr. S. Gopinathan - Professor in the Department of Computer Science at the University of Madras. His research interests include machine learning, optimization, and intelligent healthcare systems.

Manuscript received on 20 May 2026