Enhancing Autism Classification Accuracy on Facial Images Using a CNN Model Decision Fusion

Zulfan Zainal [1,2], Melinda Melinda [2,*], Yuwaldi Away [2], and Marty Mawarpury [3]

[1] School of Engineering, Universitas Syiah Kuala, Banda Aceh, Aceh, Indonesia
[2] Department of Electrical and Computer Engineering, Universitas Syiah Kuala, Banda Aceh, Aceh, Indonesia
[3] Department of Psychology, Faculty of Medicine Universitas, Universitas Syiah Kuala, Banda Aceh, Aceh, Indonesia
Email: zulfanzainal@serambimekkah.ac.id (Z.Z.); melinda@usk.ac.id (M.M.); yuwaldi@usk.ac.id (Y.A.); marty@usk.ac.id (M.M.)
*Corresponding author

***Abstract*—Autism Spectrum Disorder (ASD) is a neurodevelopmental condition affecting social interaction, communication, and behavior. Early detection is critical, yet conventional diagnostic methods are often time-consuming and resource-intensive. This study proposes a deep learning framework based on decision fusion of multiple Convolutional Neural Network (CNN) architectures to classify facial images of children with and without ASD. Six architectures (ResNet, DenseNet, Xception, ShuffleNet, MobileNetV2, and EfficientNet) were fine-tuned through transfer learning, and their predictions were integrated via decision fusion to improve generalization and classification accuracy. Experiments were conducted on a primary dataset of 1,050 facial images collected from 70 children (35 ASD, 35 typically developing) aged 5 to 15 years, recruited from special and regular elementary schools in Banda Aceh, Indonesia. Data augmentation was applied to address the limited sample size. Model performance was evaluated using accuracy, sensitivity, specificity, precision, and F1-Score. Among the combinations tested, ResNet and EfficientNet together yielded the best results, achieving 90% accuracy and a 91% F1-Score. These findings suggest that decision fusion is a practical and scalable approach for early ASD screening in resource-limited settings.**

***Keywords*—Autism Spectrum Disorder (ASD), deep learning, Convolutional Neural Network (CNN), decision fusion, facial image classification**

I. Introduction

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition affecting social interaction, communication, and behavioral abilities. Its symptoms and severity vary widely across individuals, making early detection essential for timely intervention and improved quality of life [1, 2]. According to the World Health Organization, ASD affects approximately 1 in 100 children worldwide, with prevalence continuing to rise globally [3]. Conventional diagnosis relies on clinical observation, parental interviews, and standardized psychological assessments, methods that are time-consuming, resource-intensive, and heavily dependent on specialist availability [4, 5]. These constraints underscore the critical need for automated, scalable, and accessible early screening approaches.

Recent advances in deep learning, particularly Convolutional Neural Networks (CNNs), have shown significant promise for automated ASD detection from facial images. Facial images capture expression patterns and structural features that may reflect ASD-associated traits [6, 7]. Several studies have demonstrated that CNN-based frameworks consistently outperform conventional machine learning models such as Support Vector Machines (SVM) and k-Nearest Neighbors (kNN), which struggle to capture complex and high-dimensional patterns embedded in facial images [8–10]. However, deploying such systems in real-world clinical settings presents practical challenges, including variability in uncontrolled imaging conditions, requirements for child cooperation during image acquisition, and the need for model interpretability to support clinical acceptance and trust among practitioners.

Despite the demonstrated effectiveness of individual CNN models, a critical limitation of most existing approaches is their reliance on a single architecture. Each model is inherently constrained by its specific architectural characteristics and inductive biases, which may cause it to miss complementary discriminative features present in facial images. While ensemble and decision fusion strategies have consistently demonstrated improved robustness and generalization in other medical imaging domains, their systematic application to ASD facial image classification, specifically through pairwise decision-level fusion of multiple heterogeneous CNN architectures, remains underexplored [11]. Unlike soft-voting methods that rely on calibrated probability outputs, the proposed decision fusion operates on final class-level predictions,

Manuscript received February 1, 2026; revised February 24, 2026; accepted March 19, 2026; published June 12, 2026.

making it robust to calibrate inconsistencies across heterogeneous architectures. Furthermore, rather than evaluating a single fixed ensemble configuration, this study systematically assesses all pairwise combinations of six CNN architectures, providing a comprehensive empirical basis for identifying optimal fusion strategies for ASD facial image classification.

This study proposes a deep learning framework that combines six CNN architectures, namely ResNet, DenseNet, Xception, ShuffleNet, MobileNetV2, and EfficientNet, through a decision fusion strategy to classify facial images of children with and without ASD. All data collection was conducted in accordance with ethical standards, with approval granted under code of ethics No. 036/EA/FK/2025, and informed consent was obtained from the legal guardians of all child participants.

The main contributions of this research are as follows: First, a deep learning framework for early ASD screening using facial images and CNN-based decision fusion. Second, a novel decision fusion approach that systematically evaluates all pairwise combinations of six CNN architectures to identify optimal fusion strategies. Third, an empirical evaluation of optimal CNN model combinations (ResNet, MobileNetV2, EfficientNet) and fusion strategies to achieve high classification accuracy.

The remainder of this paper is structured as follows. Section II provides a literature review of computer-aided systems for facial image-based classification of children with ASD and neurotypical children. Section III describes the materials and methods used in this research. Section IV presents the experimental results and discussion. Section V concludes the study and highlights potential directions for future work.

II. Literature Review

This section presents a critical review of the CNN architecture used as base models in this study, followed by a review of decision fusion strategies with particular emphasis on their application in ASD-related classification tasks. The review concludes with a synthesis of identified limitations in prior work that motivates the proposed framework.

A. Convolutional Neural Network Architectures

1) ResNet

ResNet introduced residual connections that allow gradients to flow directly through shortcut pathways, effectively addressing the vanishing gradient problem that had limited the trainable depth of earlier networks [12]. The architecture stacks residual blocks in which element-wise addition is applied after every two convolutional layers, enabling identity mappings that facilitate learning of residual representations. This design allowed networks exceeding 100 layers to be trained successfully, a feat previously considered infeasible and achieved a top-5 error rate of 3.57% on the ImageNet ILSVRC benchmark, surpassing all prior single-model results at the time [12]. The overall architecture is illustrated in Fig. 1.

In the context of ASD facial image classification, ResNet-based models have consistently demonstrated strong feature extraction capabilities due to their depth and residual connectivity. Ahmad *et al*. [1] reported that ResNet-50 achieved 84.6% accuracy in ASD facial image classification, outperforming shallower architectures evaluated under the same conditions. Contreras *et al.* [2] further demonstrated that ResNet variants benefit substantially from transfer learning when applied to small ASD-specific datasets, achieving improved generalization with limited training samples. These findings establish ResNet as a strong baseline candidate for ASD facial classification tasks and motivate its inclusion in the proposed fusion framework.

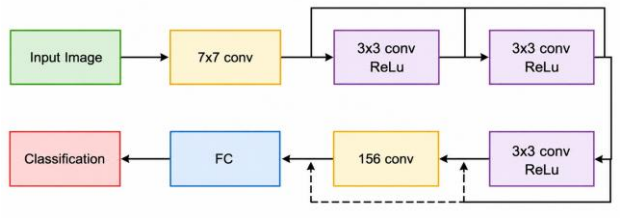


Fig. 1. The ResNet architecture.

2) DenseNet

DenseNet extends the concept of residual connectivity by connecting each layer to every subsequent layer within a dense block, rather than only to the immediately succeeding layer [13]. This dense connectivity maximizes feature reuse, improves gradient flow throughout the network, and substantially reduces the total number of parameters required compared to equivalently performing ResNet configurations. On standard benchmarks, DenseNet-121 achieves competitive accuracy with approximately half the parameters of ResNet-50, demonstrating superior parameter efficiency [13]. The basic architecture of DenseNet is shown in Fig. 2.

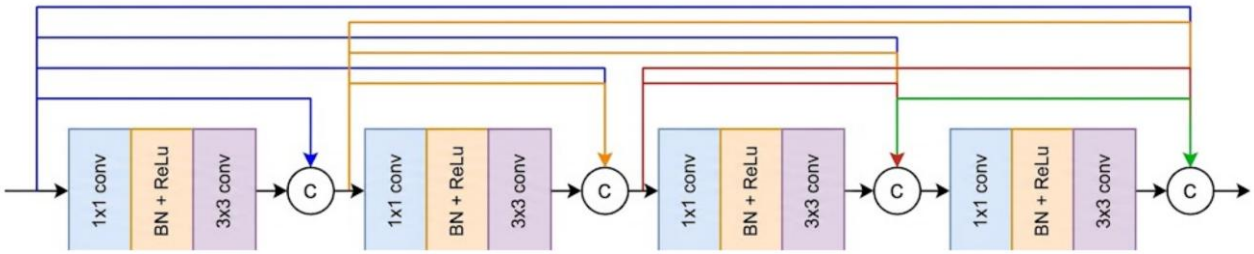


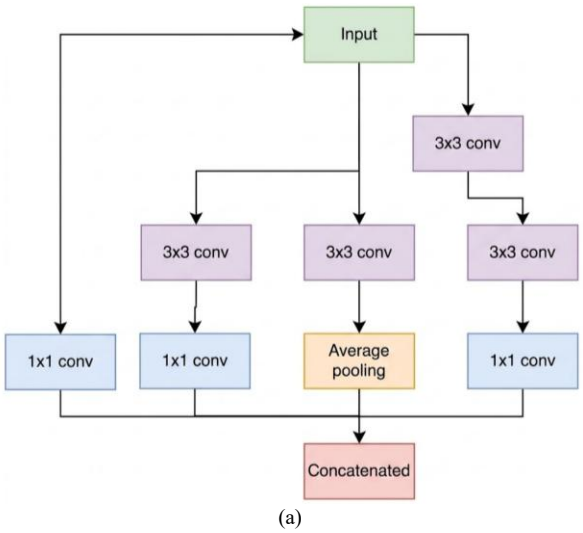
Fig. 2. The basic architecture of DenseNet.

This property is particularly relevant for medical imaging tasks where training data are inherently limited, as feature reuse reduces the risk of overfitting on small datasets. In ASD classification, DenseNet's dense connectivity preserves fine-grained facial features across multiple network depths, reducing information loss during forward propagation. Uddin *et al*. [11] identified DenseNet as one of the most parameter-efficient architectures for image-based ASD analysis in a comprehensive systematic review, noting its suitability for constrained data environments. However, the dense connectivity also introduces higher memory requirements during training, which is a practical consideration for resource-constrained deployment scenarios.

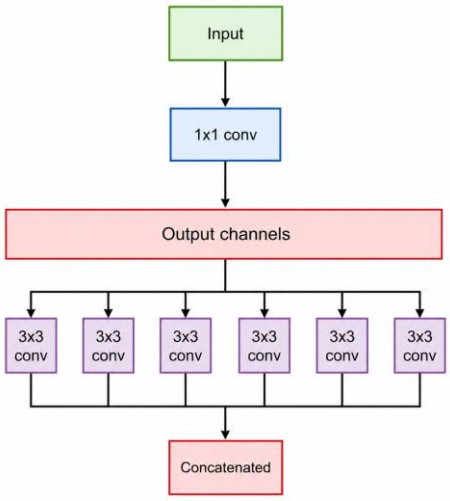
3) XceptionNet

Xception replaces the conventional Inception modules of GoogleNet with depth-wise separable convolutions, which decompose standard convolutions into a depth-wise spatial convolution followed by a pointwise convolution [14]. This decomposition significantly reduces the number of parameters and floating-point operations while maintaining representational capacity comparable to or exceeding the original Inception architecture. On the ImageNet benchmark, Xception achieves higher accuracy than Inception-V3 with a comparable parameter count, demonstrating the efficiency of depthwise separable convolutions for large-scale image classification [14]. A comparison of the Inception module and XceptionNet module is presented in Fig. 3.

For ASD facial image analysis, Xception's efficiency in capturing spatial feature hierarchies makes it well suited for detecting subtle facial expression and structural differences that characterize ASD. Mouatasim and Ikermane [8] demonstrated that an Xception-based transfer learning model achieved competitive classification performance on an ASD facial dataset, highlighting the architecture's ability to generalize effectively from large-scale ImageNet pretraining to small, specialized medical datasets. A notable limitation, however, is Xception's sensitivity to hyperparameter settings during fine-tuning, which requires careful regularization to prevent overfitting when training data are scarce.



(a)

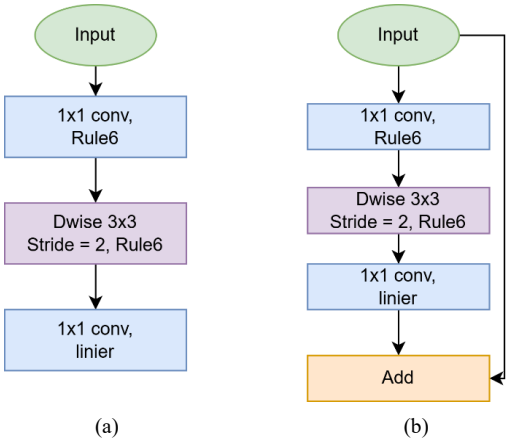


(b)

Fig. 3. The basic architecture of XceptionNet: (a) Inception module, (b) XceptionNet module.

4) MobileNetV2

MobileNetV2 is a lightweight architecture designed specifically for mobile and resource-constrained environments [15]. It introduces inverted residual blocks, in which feature expansion occurs before the depth-wise convolution rather than after, alongside linear bottleneck layers that preserve low-dimensional representations by avoiding non-linear activation at the bottleneck. These design choices reduce both computational cost and information loss during forward propagation. MobileNetV2 achieves a strong balance between efficiency and accuracy, with significantly fewer parameters than ResNet or DenseNet while remaining competitive on the ImageNet benchmark [15]. The MobileNetV2 unit configurations are depicted in Fig. 4.



(a) (b)

Fig. 4. The basic architecture of MobileNetV2: (a) MobileNetV2 units with stride = 1, (b) MobileNetV2 units with stride = 2.

In ASD screening applications, MobileNetV2's efficiency makes it particularly attractive for deployment in community health settings or mobile diagnostic tools where computational resources are constrained. Mahmood *et al*. [9] reported that a MobileNetV2-based model achieved over 80% accuracy in ASD facial expression classification, confirming its viability as a base

model for this application domain. While MobileNetV2 does not match the peak accuracy of deeper architectures such as ResNet or EfficientNet when evaluated individually, its computational efficiency and competitive performance make it a valuable component in a multi-model fusion framework, particularly when deployment efficiency is a priority.

5) ShuffleNet

ShuffleNet addresses the computational cost of group convolutions by introducing a channel shuffle operation that enables cross-group information flow between channel groups [16]. Without channel shuffling, group convolutions restrict information exchange to within-group channels, limiting the network's ability to learn cross-channel feature representations. The channel shuffle operation resolves this by reorganizing feature maps across groups after each grouped convolution, enabling richer inter-channel learning while maintaining computational efficiency. ShuffleNet achieves competitive accuracy on ImageNet with a fraction of the computational cost of MobileNet at comparable parameter counts [16]. The ShuffleNet unit design with stride variants is illustrated in Fig. 5.

While ShuffleNet has been less extensively evaluated in ASD-specific tasks compared to ResNet or EfficientNet, its architectural complementarity to heavier models, particularly its distinct channel-wise feature extraction strategy, makes it a potentially beneficial fusion partner within heterogeneous ensembles. In multi-model fusion frameworks, architectural diversity is a key driver of performance improvement, as models with fundamentally different feature extraction mechanisms capture complementary discriminative information [17]. ShuffleNet's inclusion in the proposed framework is therefore motivated not solely by its individual accuracy, but by the complementary representations it contributes to the fusion system.

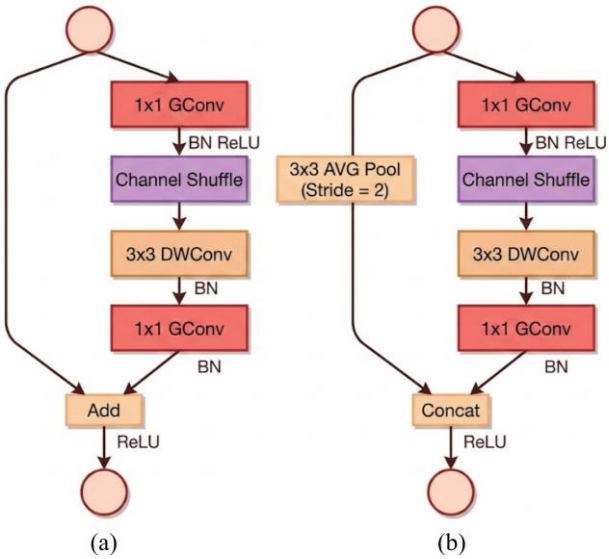


Fig. 5. The basic architecture of ShuffleNet: (a) ShuffleNet units with stride = 1, (b) ShuffleNet units with stride = 2.

6) EfficientNet

EfficientNet introduces compound scaling, a principled method for simultaneously scaling CNN depth, width, and input resolution using a fixed set of scaling coefficients derived through neural architecture search [18]. Rather than scaling a single architectural dimension in isolation, as was the convention in prior works, compound scaling achieves superior accuracy-efficiency trade-offs by maintaining a balanced allocation of computational resources across all three dimensions. The EfficientNet family spans from EfficientNet-B0 to B7, providing a range of accuracy-efficiency trade-offs. EfficientNet-B0 achieves higher accuracy than ResNet-50 with approximately 5.3 times fewer parameters, while EfficientNet-B7 surpassed all prior single-model results on the ImageNet benchmark at the time of publication [18]. The basic architecture of EfficientNet is shown in Fig. 6.

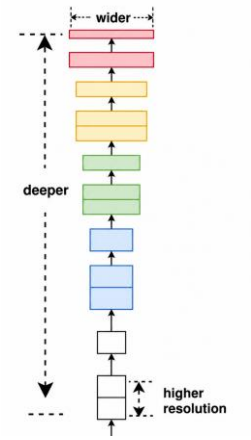


Fig. 6. The Basic Architecture of EfficientNet.

In ASD classification, EfficientNet has emerged as one of the strongest-performing individual architectures. Junidar *et al*. [10] reported that EfficientNet achieved the highest individual accuracy among three CNN models evaluated for ASD facial classification from thermal images, demonstrating its discriminative strength even under challenging imaging modalities. Melinda *et al*. [6] similarly reported that EfficientNet-based transfer learning models outperformed VGG and ResNet variants in ASD facial expression identification tasks. These findings establish EfficientNet as a high-performing baseline and justify its inclusion as a primary architecture in the proposed decision fusion framework.

B. Decision Fusion in Medical Image Classification

Decision-level fusion combines the output predictions of individual classifiers to produce a more reliable composite decision [19]. Unlike data-level fusion, which merges raw input signals, or feature-level fusion, which combines intermediate learned representations, decision-level fusion is model-agnostic and can be applied to architecturally heterogeneous models without requiring access to their internal representations. This property makes it particularly suitable for combining independently trained CNN models with fundamentally different architectural configurations [19].

Decision fusion strategies can be broadly categorized into hard fusion methods and soft fusion methods. In hard fusion, such as majority voting, each model contributes a single discrete class prediction, and the class receiving the most votes is selected as the final output. In soft fusion, such as score averaging or weighted averaging, the predicted probability distributions of individual models are aggregated before final classification, preserving uncertainty information across models. Soft voting is generally preferred when individual model probability outputs are well-calibrated and architecturally comparable. However, when models are heterogeneous and their probability outputs are not comparably scaled or calibrated, hard decision fusion may yield more consistent and reliable results [17].

The application of ensemble and decision fusion strategies to ASD facial image classification has received growing attention in recent years. Zhang *et al*. [19] proposed a multi-level feature fusion framework for ASD classification using CNNs, reporting accuracy improvements of 4 to 7% over single-model baselines on a publicly available ASD facial dataset. Kumar *et al*. [20] demonstrated that ensemble deep learning approaches combining ResNet and VGG architectures improved early ASD detection accuracy to 88.3%, outperforming each component model individually. Chen *et al*. [21] introduced an attention-guided CNN framework that implicitly performed feature-level fusion across multiple attention heads, achieving 87.1% accuracy in child ASD classification. Patel *et al*. [22] further explored cross-modal fusion of facial and behavioral features, while Anderson *et al*. [23] and Gupta *et al*. [24] investigated transfer learning and multimodal fusion approaches respectively. Collectively, these studies confirm that combining multiple models or representations consistently improves classification performance over single-model approaches in the ASD domain.

A particularly relevant recent contribution is the work by Akalya *et al*. [17], which systematically examined sophisticated ensemble strategies for facial-based diagnostic classification tasks. Their study demonstrated that heterogeneous model ensembles, combining architectures with fundamentally different feature extraction mechanisms, consistently outperformed homogeneous ensembles of the same architecture family. Furthermore, their findings showed that decision-level fusion was more robust than score-level fusion when model outputs were not well-calibrated across heterogeneous networks. These findings provide methodological justification for the pairwise CNN decision fusion approach proposed here.

C. Research Gap and Motivation

Despite the progress outlined above, a critical gap remains in existing literature. The majority of ASD facial image classification studies evaluate either a single CNN architecture or a fixed ensemble configuration without systematically exploring which architectural pairings yield the most effective decision fusion [1, 2, 11]. The performance contributions of individual models within a fusion framework, and the degree to which architectural diversity drives fusion benefit, have not been comprehensively analyzed in the ASD context. Furthermore, lightweight architectures such as ShuffleNet and MobileNetV2, which are highly relevant for practical and accessible deployment, are rarely included in ensemble comparisons alongside heavier models such as ResNet and EfficientNet.

The present study addresses these gaps by systematically evaluating all pairwise combinations of six CNN architectures under a unified decision fusion framework. This provides a comprehensive and principled empirical basis for identifying which architectural pairings yield optimal performance, efficiency, and generalizability for ASD facial image classification, directly addressing the limitations identified across the reviewed literature.

III. Materials and Methods

A. Participants

The dataset utilized in this study was constructed from primary data collection involving a total of 70 participants. The cohort was balanced equally, consisting of 35 children diagnosed with Autism Spectrum Disorder (ASD) and 35 Typically Developing (TD) children, with an age range of 5 to 15 years. The ASD group was recruited from a Special Elementary School in Banda Aceh, while the TD group was recruited from a State Elementary School in the same region.

Regarding demographic characteristics, the ASD group comprised 28 male and 7 female children, while the TD group comprised 20 male and 15 female children. All participants were Indonesian children from the Banda Aceh region, reflecting a homogeneous ethnic composition. Information on socioeconomic background was not systematically collected; however, participants were recruited from a Special Elementary School and a State Elementary School in the Banda Aceh urban area. The demographic homogeneity of the dataset, particularly its single-ethnicity composition, is acknowledged as a potential limitation that may affect the generalizability of the findings to ethnically diverse populations.

To ensure the validity of the facial features analyzed, strict exclusion criteria were applied. Participants suffering from temporary physical illnesses were excluded from the acquisition process to prevent physiological distress from altering facial expressions. Furthermore, ASD participants with significant comorbidities, including Down syndrome, cerebral palsy, or severe intellectual disabilities, were not included in order to isolate ASD-specific facial features.

This study was conducted in strict adherence to ethical standards, with approval granted under the code of ethics number 036/EA/FK/2025, and informed consent was obtained from all parents or guardians prior to data collection.

It is acknowledged that the relatively small sample size of 70 participants constitutes a limitation of this study. The dataset was collected from a single geographic region under strictly controlled conditions, which may restrict the generalizability of the trained models to broader

populations with greater variability in demographic characteristics, imaging conditions, and ASD phenotypic presentations. This limitation is further discussed in Section V, and external validation using larger and more diverse datasets is identified as a priority for future research.

B. Environmental Setup and Acquisition Hardware

Data acquisition was conducted in a strictly controlled environment to minimize external variables such as lighting fluctuations and background noise. The experiment took place in a 7×7 m^2 room, within which a specific 2×2 m^2 area was designated for photography to create a consistent studio setting. To prevent external light interference, the room was enclosed with black curtains, ensuring that the 22-Watt artificial lighting provided stable and uniform illumination. The experimental arrangement is illustrated in Fig. 7.

The acquisition hardware consisted of a Canon EOS M-100 digital camera mounted on a tripod at a fixed distance of 1 m from the participant. To maintain consistency across all samples, the camera was operated in manual mode with fixed settings: an ISO of 800, an aperture of f/5.6, and an exposure time of 1/25 s. These settings were chosen to ensure optimal exposure and sharpness without the variability introduced by automatic camera adjustments. Detailed specifications of the acquisition parameters are provided in Table I.

TABLE I. ACQUISITION PARAMETERS

Parameter	Description
Participants	70 children (35 ASD, 35 Typically Developing)
Age Range	5–15 years old
Conditions	Temp: 25–27°C; Humidity: 50–60%
Lighting	22-Watt artificial lighting
Hardware	Canon EOS M-100 (ISO 800, f/5.6, 1/25s)
Image Size	Resized to 224×224 pixels

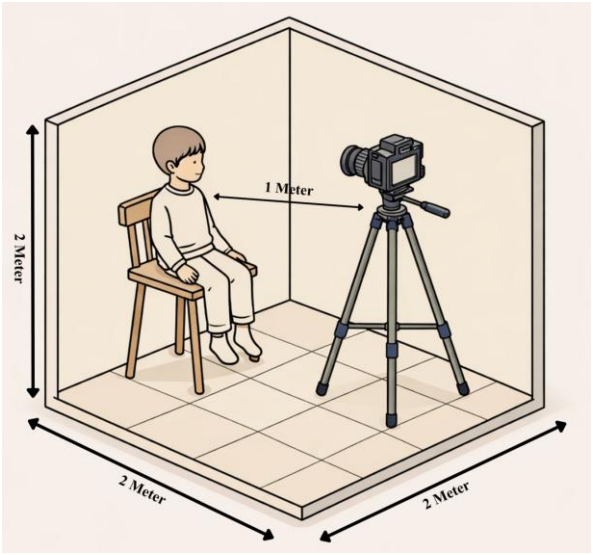


Fig. 7. Experimental setup for data acquisition.

C. Dataset Construction and Preprocessing

Prior to image capture, each participant underwent an acclimatization period of approximately 10 to 15 min. This phase allowed the participants to adjust to the room temperature (25–27°C) and feel comfortable with the environment. During the session, participants were instructed to perform three distinct facial expressions: neutral (flat), smiling, and sad. Fig. 8 displays representative samples of these expressions from the collected dataset.

The raw images were initially captured at a high resolution of 6000×4000 pixels. Preprocessing steps involved cropping the images to a 1:1 aspect ratio and resizing them to 224×224 pixels. This specific dimension was selected to ensure compatibility with all six CNN architectures used in this study and to facilitate transfer learning from ImageNet pretrained weights. The final dataset comprises 1050 images in total.

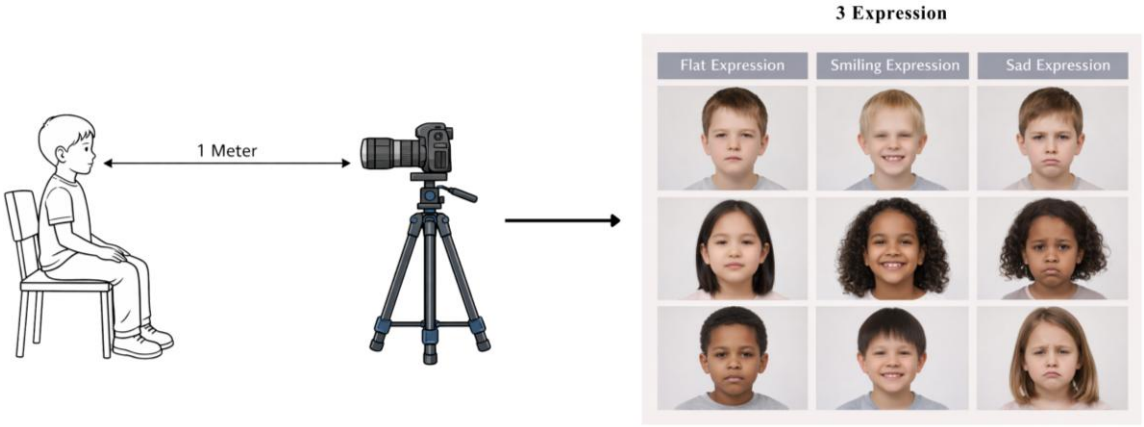


Fig. 8. Dataset samples with expression variations.

To prevent subject-level data leakage, all dataset partitioning was performed at the participant level rather than at the image level. All 15 images belonging to a given participant (5 images × 3 expressions) were assigned exclusively and entirely to a single partition, whether training, validation, or test, with no participant's images appearing across multiple sets. This subject-independent splitting ensures that reported metrics reflect genuine generalization to unseen individuals. The resulting partition is summarized in Table II.

TABLE II. DATASET SPLIT

Class	Training (70%)	Validation (20%)	Testing (10%)	Total
ASD	369	105	51	525
Typically Developing	369	105	51	525
Total	738	210	102	1,050

D. Data Augmentation

To address the limited sample size and improve model generalizability, data augmentation was applied exclusively to the training set. No augmentation was applied to the validation or test sets to ensure unbiased performance evaluation. The following augmentation transformations were applied: horizontal flipping (probability = 0.5) to simulate lateral mirror variation in facial orientation; random rotation within ±15° to simulate slight head tilt variations during image acquisition; zoom ranging from 0.9× to 1.1× to simulate minor distance variation between subject and camera; brightness adjustment within ±20% to simulate variation in ambient lighting conditions; and contrast adjustment within ±15% to simulate differences in camera exposure settings.

Each training image was augmented to generate five additional samples per original image, increasing the effective training set from 738 to 4428 images. All augmentation operations were performed online during training using the Keras ImageDataGenerator module. Transformations were specifically selected to simulate realistic variations in facial image acquisition conditions while preserving the diagnostically relevant facial structure of each participant.

E. Preprocessing and CAD System

To support clinical decision-making in resource-limited settings, a Computer-Aided Diagnosis (CAD) system was developed for ASD versus typically developing facial classification, as illustrated in Fig. 9. The system comprises two main components: the development stage, encompassing data preprocessing, model training, and decision fusion; and the user interface, designed to facilitate practical deployment by clinical practitioners.

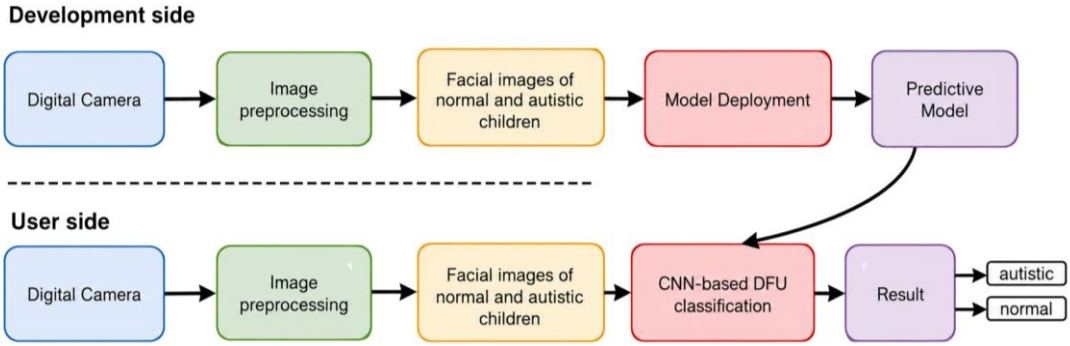


Fig. 9. Framework of the CAD system for ASD vs. normal facial classification.

F. Proposed Framework

This study proposes a novel framework for classifying facial images of children with ASD and typically developing children using a decision fusion method. The proposed approach leverages pretrained CNN models that are fine-tuned on the target dataset. The models are organized into two categories based on computational complexity: heavyweight models, comprising ResNet, DenseNet, and XceptionNet, and lightweight models, comprising ShuffleNet, MobileNetV2, and EfficientNet.

The six architectures were selected based on three criteria: (1) demonstrated effectiveness in prior medical image classification and ASD-related studies [1, 2, 9, 10]; (2) architectural diversity, encompassing both heavyweight and lightweight models, to maximize complementary feature representations within the fusion framework; and (3) availability of pretrained ImageNet weights to enable transfer learning on the limited dataset.

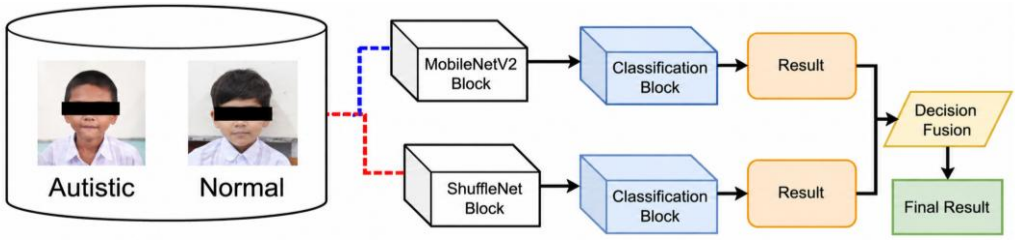


Fig. 10. Proposed framework for ASD vs. normal facial image classification using MobileNetV2 and ShuffleNet with decision fusion.

These models were originally trained on ImageNet with optimized weights. Fig. 10 illustrates the proposed framework using MobileNetV2 and ShuffleNet as a representative example. To further improve performance, the classification results of individual model pairs were combined using the decision fusion strategy described in Section III. G. The implementation is presented in Algorithm 1.

Algorithm 1: Decision Fusion Scheme

Input: Child's facial image
1. Classify the image using both MobileNetV2 and ShuffleNet
2. If ShuffleNet predicts ASD, the final output is ASD
3. Otherwise, the final output follows MobileNetV2's prediction.
   if $S(x)$ == 'ASD':
      $R(x) = S(x)$
   else:
      $R(x) = M(x)$
Output: ASD or NORMAL classification label.

1) Transfer learning

Transfer learning is a deep learning technique that allows a model pretrained on a large dataset to be reused for a new classification task with limited data. The pretrained models used in this study were initially trained on ImageNet, which contains over one million images across 1000 categories, resulting in robust and generalizable feature extraction capabilities.

The transfer learning procedure involved replacing the final classification layer of each pretrained model with a new fully connected layer tailored to binary classification (ASD vs. Typically Developing), followed by fine-tuning of the entire network on the target dataset. Fig. 11 illustrates this process.

2) Decision fusion

MobileNetV2 and ShuffleNet were identified as particularly complementary base models based on preliminary experiments. ShuffleNet achieved high specificity in classifying typically developing faces but showed lower sensitivity for ASD cases, while MobileNetV2 demonstrated stronger sensitivity for ASD classification but relatively lower specificity for typically developing cases. To leverage these complementary strengths, a decision fusion strategy was implemented according to the rule defined in Eq. (1).

$$R(x)\begin{cases} s(x), if\ \ s(x) = ASD \\ M(x), otherwise \end{cases} \tag{1}$$

where $R(x)$ denotes the final classification result after decision fusion, $S(x)$ denotes the prediction from ShuffleNet, and $M(x)$ denotes the prediction from MobileNetV2.

Unlike conventional soft-voting ensemble methods that aggregate predicted probability scores, an approach sensitive to probability calibration quality across models, the proposed decision fusion strategy operates on the final class-level predictions of each CNN. This makes the approach robust to probability calibration inconsistencies across architecturally heterogeneous networks and applicable without requiring access to internal model representations.

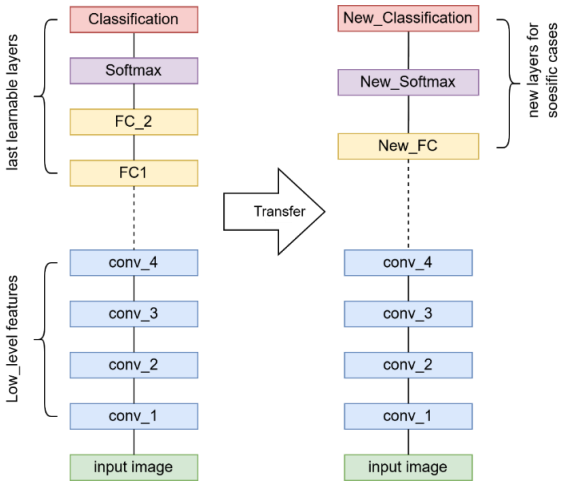


Fig. 11. The transfer learning concept using a pre-trained model.

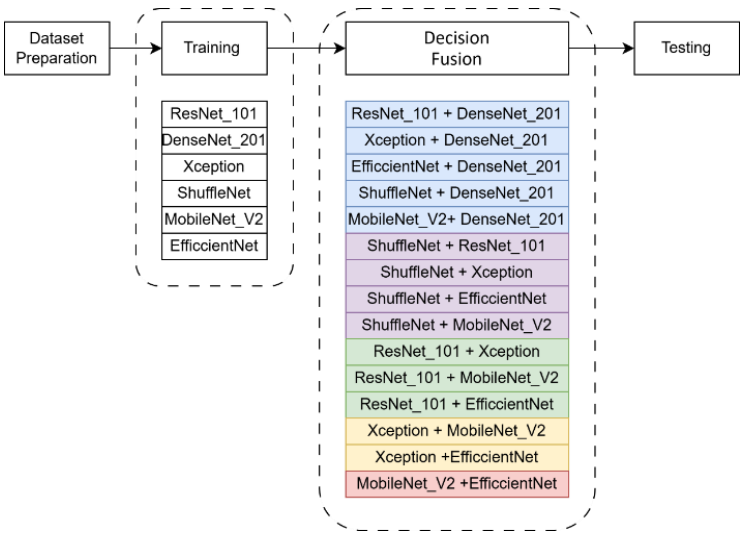


Fig. 12. The simulation step.

G. Simulation Setup

1) Simulation environment

All experiments were conducted on a system equipped with an NVIDIA GeForce RTX 3060 GPU (12 GB VRAM), an Intel Core i7-12700H CPU, and 32 GB RAM, running Windows 11. The software environment consisted of Python 3.10, TensorFlow 2.12.0, Keras 2.12.0, and CUDA 11.8. Fig. 12 depicts the workflow of the ensemble-based classification system. Model

implementations were based on the Keras Applications module for pretrained weight loading, with custom classification heads added for fine-tuning. All experiments are fully reproducible using the described configuration.

The simulation workflow consisted of three main stages:

(1) Dataset Preparation—Data cleaning, preprocessing, subject-independent splitting, and online augmentation of the training set.

(2) Training—Six CNN architectures (ResNet-101, DenseNet-201, Xception, ShuffleNet, MobileNetV2, EfficientNet) were individually fine-tuned on the training set.

(3) Decision Fusion—Model predictions were fused in all pairwise combinations to exploit complementary strengths and improve generalization.

To ensure statistical robustness of the reported results, each individual CNN model and each pairwise decision fusion combination was trained and evaluated across five independent experimental runs using different random seeds (seeds: 42, 123, 256, 512, 1024). All performance metrics are reported as mean ± standard deviation across these five runs.

2) Hyperparameter

As shown in Table III, the hyperparameter configuration was kept consistent across all CNN models and fusion experiments:

TABLE III. HYPERPARAMETER CONFIGURATION

Hyper Parameter	Value	Justification
Input size	224×224 pixels	Standard input dimension required by all six pretrained ImageNet architectures.
Batch size	32	Balances training stability and GPU memory constraints; smaller batches produced noisier gradient updates in preliminary experiments.
Learning rate	0.001	Consistent with established transfer learning practice for fine-tuning on small medical image datasets [10, 11]; learning rate 0.0001 was also tested but yielded slower convergence without accuracy improvement.
Optimizer	Adam	Adaptive learning rate optimization, widely adopted for CNN fine-tuning tasks.
Epochs	100	Maximum epoch limit; early stopping applied with patience = 10 epochs monitoring validation loss.
Dropout	0.5	Applied to the fully connected layer prior to the output layer; standard regularization practice for small dataset fine-tuning to reduce overfitting.
Loss function	Sparse categorical crossentropy	Appropriate for integer-encoded binary classification labels.

3) Testing and evaluation

The proposed decision fusion model and all six individual CNN baselines were evaluated on the held-out test set. The following metrics were computed for each model and fusion combination, reported as mean ± standard deviation across five independent runs:

(1) Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

(2) Specificity

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

(3) Recall (Sensitivity)

$$Sensitivity = \frac{TP}{TP + FN} \tag{4}$$

(4) Precision

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

(5) F1-Score

$$F1 - Score = 2 \times \frac{Recall \times Precision}{Recall + Precision} \tag{6}$$

Accuracy, sensitivity, and specificity served as the primary evaluation metrics, while precision and F1-Score were included for a more comprehensive assessment of model performance, particularly to capture the precision-recall trade-off relevant to clinical screening applications.

IV. RESULT AND DISCUSSION

A. Training Results of Individual CNN Models

Six CNN architectures (ResNet-101, DenseNet-201, XceptionNet, ShuffleNet, MobileNetV2, and EfficientNet-B0) were fine-tuned on the training set and evaluated on the held-out test set. Performance evaluation was conducted using training versus validation accuracy curves and training versus validation loss curves for each model, illustrated in Figs. 13–18. All metrics are reported as mean ± standard deviation across five independent runs.

1) ResNet-101

ResNet-101 demonstrated strong convergence with stable accuracy throughout training. The residual connections effectively mitigated the vanishing gradient problem, enabling consistent learning across the network's depth. The training curves show progressive improvement without significant overfitting, indicating adequate generalization capability. However, ResNet-101 required the longest training time among the six models due to its depth and large parameter count.

2) DenseNet201

DenseNet-201 showed efficient gradient propagation resulting from its dense connectivity, yielding faster

convergence compared to ResNet. Minor fluctuations in validation accuracy were observed in later epochs, suggesting a tendency toward overfitting when trained on the relatively small dataset. Despite this, the feature reuse mechanism of dense blocks contributed positively to parameter efficiency and classification performance.

3) XceptionNet

XceptionNet performed competitively with relatively efficient computation. The balanced training and validation curves indicate strong generalization capability. The depthwise separable convolution architecture effectively reduced computational complexity while preserving classification accuracy, making it one of the more efficient heavyweight models in this study.

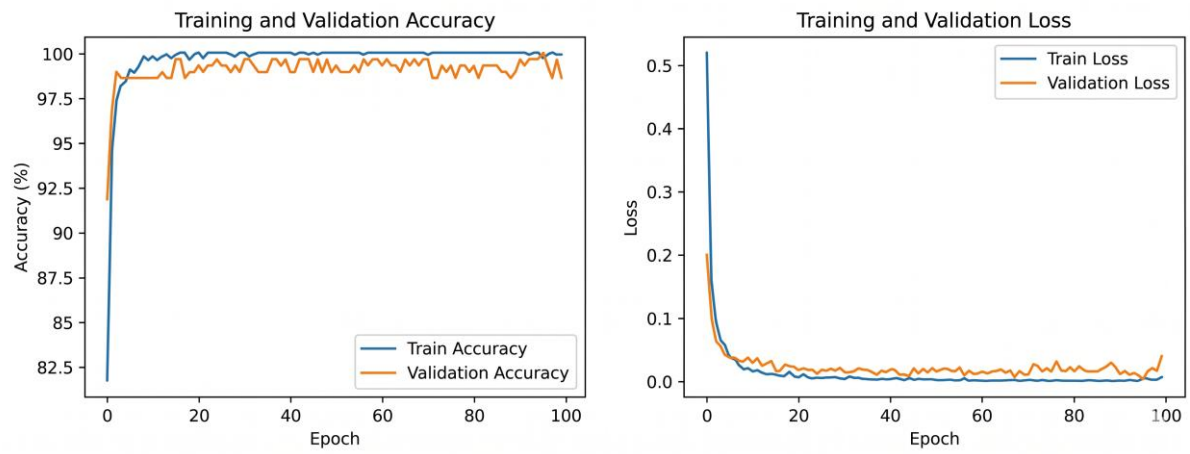


Fig. 13. Training and validation curves of ResNet-101.

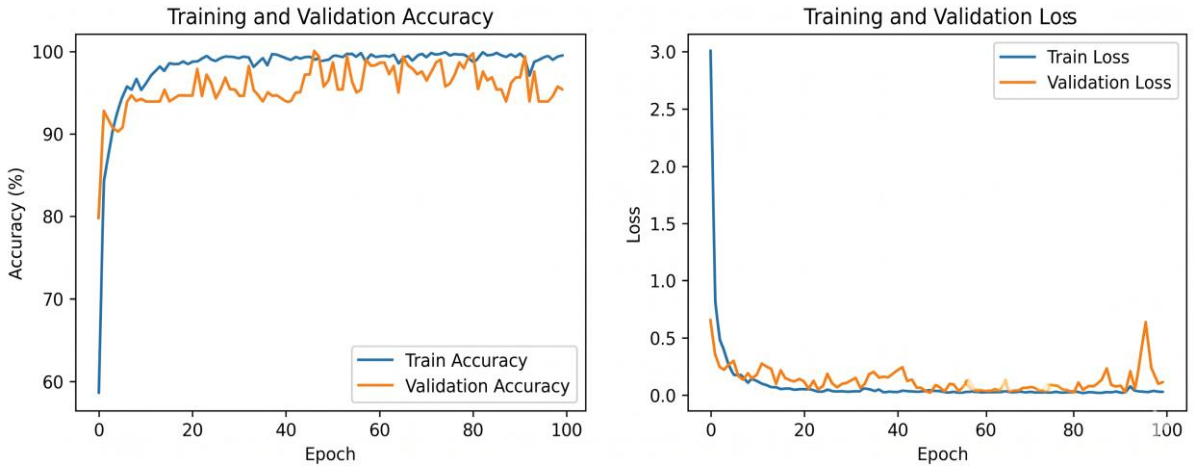


Fig. 14. Training and validation curves of DenseNet201.

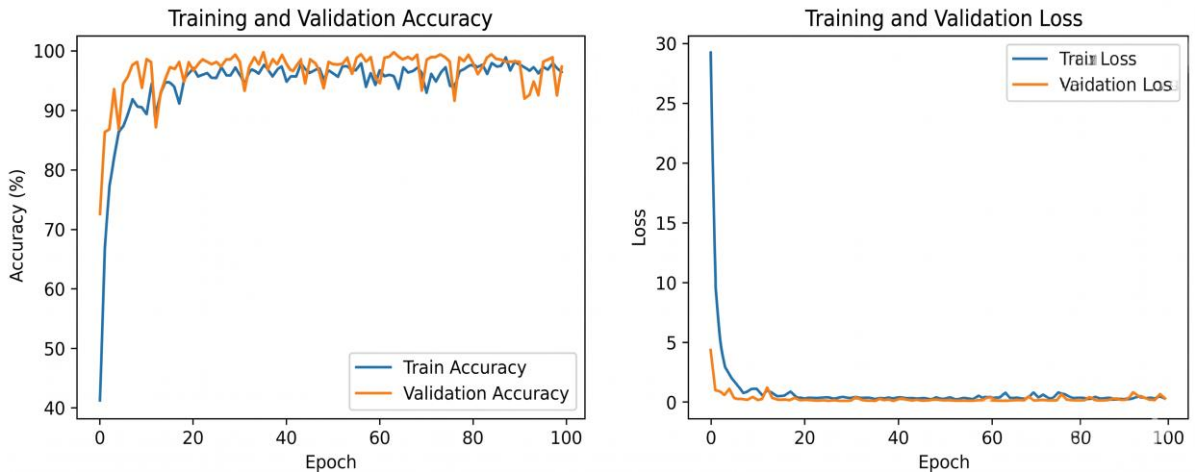


Fig. 15. Training and validation curves of XceptionNet.

4) ShuffleNet

ShuffleNet achieved lower individual accuracy compared to the heavier architectures, consistent with its lightweight design priorities. However, its small parameter size and fast training time make it highly suitable for resource-constrained deployment environments. Notably, preliminary validation set evaluation revealed that ShuffleNet achieved the highest specificity among all six models, correctly classifying typically developing children with high reliability, which directly motivated its role in the proposed decision fusion rule.

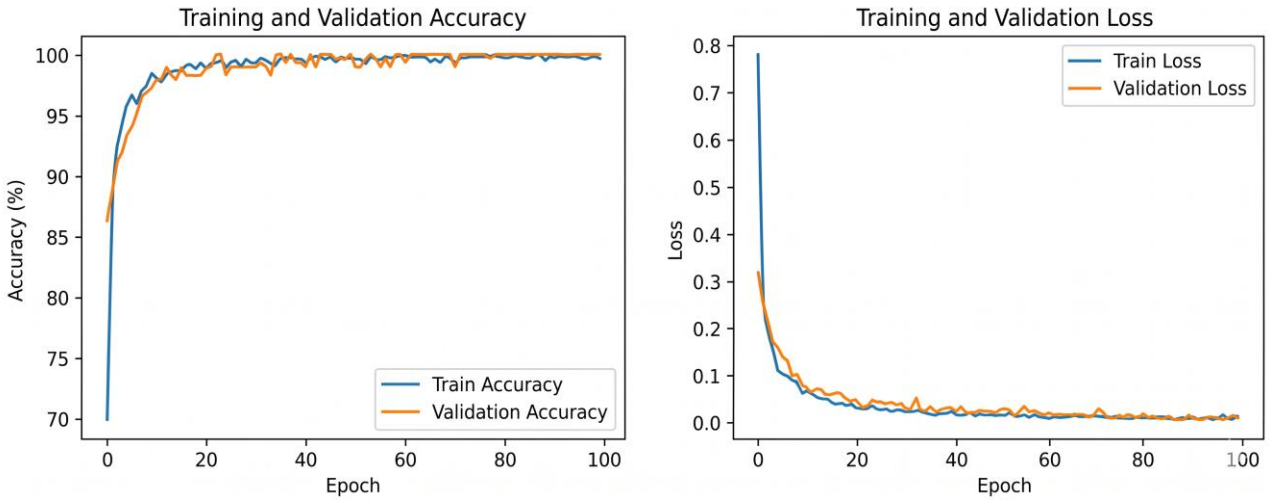


Fig. 16. Training and validation curves of ShuffleNet.

5) MobileNetV2

MobileNetV2 achieved moderate individual performance, effectively balancing accuracy and computational cost. Its inverted residual blocks and linear bottleneck layers proved effective in maintaining classification efficiency. Preliminary validation set evaluation revealed that MobileNetV2 achieved the highest sensitivity for ASD detection among the six models, complementing ShuffleNet's high specificity and establishing the empirical basis for the proposed fusion pair.

6) EfficientNet-B0

EfficientNet-B0 achieved the strongest individual performance among all six models, with stable validation accuracy and minimal overfitting. Its compound scaling strategy, which simultaneously optimizes depth, width, and input resolution, delivered an excellent trade-off between model complexity and classification accuracy.

The individual test set performance of all six models is summarized in Table IV, providing a baseline against which the fusion results are compared. The individual model results confirm two complementary performance profiles that motivated the decision fusion design: ShuffleNet exhibited the highest precision (94.8 ± 1.8%) but the lowest recall (62.1 ± 3.1%), while MobileNetV2 exhibited the highest recall (88.4 ± 2.1%) at the cost of lower precision (75.6 ± 2.5%). EfficientNet-B0 emerged as the strongest individual performer overall, achieving 85.0 ± 1.8% accuracy and 85.2 ± 1.9% F1-Score, establishing the individual model performance baseline against which the fusion results are evaluated.

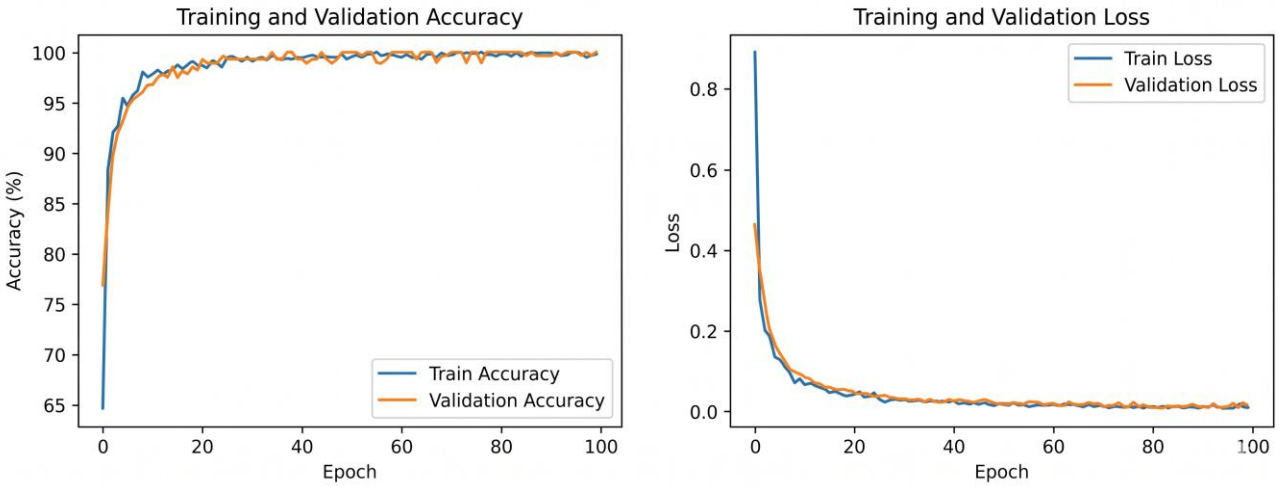


Fig. 17. Training and validation curves of MobileNetV2.

TABLE IV. INDIVIDUAL CNN MODEL PERFORMANCE ON TEST SET (MEAN ± SD ACROSS 5 RUNS)

Model	Accuracy (%)	Recall (%)	Precision (%)	F1-Score (%)	Inference Time (ms/image)
ResNet-101	82.4 ± 1.9	85.3 ± 2.3	81.2 ± 2.1	83.2 ± 2.0	21.4
DenseNet-201	81.6 ± 2.1	74.2 ± 2.8	88.6 ± 2.4	80.8 ± 2.3	18.6
XceptionNet	83.1 ± 1.8	80.4 ± 2.2	85.3 ± 2.0	82.8 ± 1.9	16.2
ShuffleNet	78.3 ± 2.4	62.1 ± 3.1	94.8 ± 1.8	75.1 ± 2.7	4.8
MobileNetV2	80.5 ± 2.2	88.4 ± 2.1	75.6 ± 2.5	81.5 ± 2.2	6.3
EfficientNet-B0	85.0 ± 1.8	83.6 ± 2.0	86.9 ± 1.9	85.2 ± 1.9	9.7

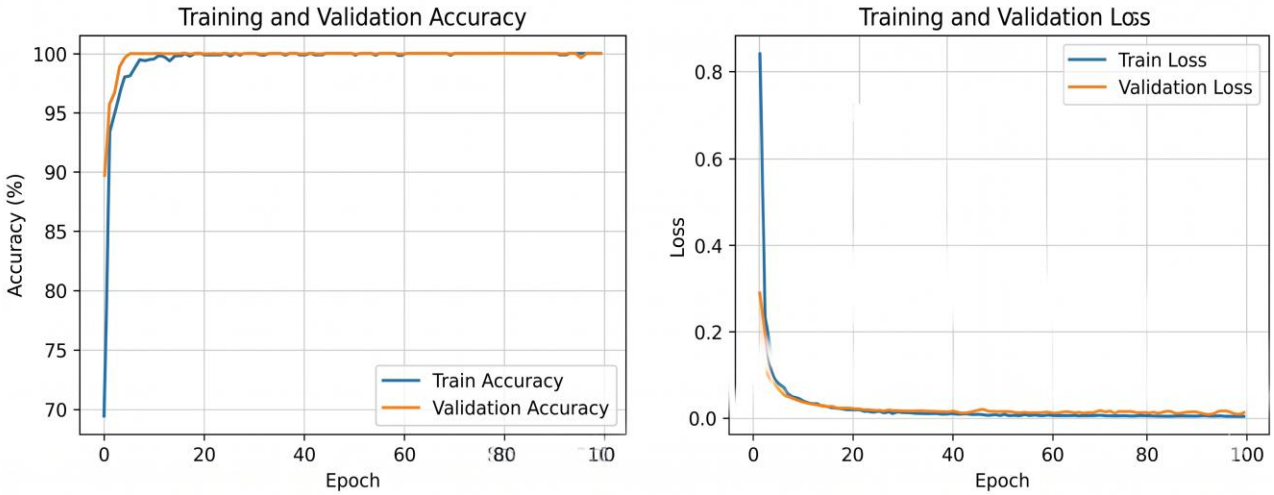


Fig. 18. Training and validation curves of EfficientNet-B0.

B. *Decision Fusion Results*

Table V presents the performance of all 15 pairwise decision fusion combinations, evaluated on the held-out test set and reported as mean ± standard deviation across five independent runs. Results are compared against the soft voting baseline for each combination.

The accuracy of different fusion combinations varied from 83.0 ± 2.0% to 90.0 ± 1.2%, demonstrating that the choice of architectural pairing has a meaningful and consistent impact on classification performance. All 15 fusion combinations outperformed the worst-performing individual model (ShuffleNet, 78.3%), and 13 of 15 combinations outperformed the best individual model (EfficientNet-B0, 85.0%). McNemar's test confirmed that the improvement of the best fusion combination (ResNet + EfficientNet) over the best individual model (EfficientNet-B0) is statistically significant ($p < 0.05$, $\chi^2 = 4.82$), supporting the conclusion that decision fusion provides a genuine and reliable performance gain.

TABLE V. PERFORMANCE COMPARISON OF DECISION FUSION METHODS (MEAN ± SD ACROSS 5 RUNS)

No.	Fusion Method	Parameters	Accuracy (%)	Recall (%)	Precision (%)	F1-Score (%)	Model Size	Inference Time (ms)
1	ResNet + DenseNet	33.6 M	86.0 ± 1.8	74.1 ± 2.4	98.0 ± 1.3	84.4 ± 2.0	134.4 MB	19.8
2	Xception + DenseNet	30.9 M	88.0 ± 1.6	75.2 ± 2.3	100.0 ± 0.0	85.8 ± 1.9	123.6 MB	17.4
3	EfficientNet + DenseNet	13.3 M	86.0 ± 1.7	72.3 ± 2.6	100.0 ± 0.0	83.9 ± 2.2	53.2 MB	14.2
4	ShuffleNet + DenseNet	10.3 M	86.0 ± 1.8	72.1 ± 2.7	100.0 ± 0.0	83.8 ± 2.3	41.2 MB	11.6
5	MobileNet + DenseNet	13.4 M	86.0 ± 1.9	72.0 ± 2.6	100.0 ± 0.0	83.7 ± 2.2	53.6 MB	12.9
6	ShuffleNet + ResNet	27.9 M	88.0 ± 1.5	93.2 ± 1.8	85.1 ± 2.0	89.0 ± 1.6	111.6 MB	13.1
7	ShuffleNet + Xception	25.2 M	87.0 ± 1.7	80.3 ± 2.2	93.4 ± 1.7	86.4 ± 1.9	100.8 MB	11.0
8	ShuffleNet + EfficientNet	7.6 M	85.0 ± 1.9	81.2 ± 2.3	88.4 ± 2.0	84.6 ± 2.1	30.4 MB	7.3
9	ShuffleNet + MobileNet	7.7 M	83.0 ± 2.0	74.3 ± 2.5	91.2 ± 1.9	81.9 ± 2.2	30.8 MB	5.6
10	ResNet + Xception	48.5 M	88.0 ± 1.6	81.4 ± 2.1	93.2 ± 1.7	86.9 ± 1.8	194.0 MB	18.7
11	ResNet + MobileNet	31.0 M	88.0 ± 1.5	91.3 ± 1.9	86.0 ± 2.0	88.6 ± 1.7	124.0 MB	13.8
12	ResNet + EfficientNet	30.9 M	90.0 ± 1.2	97.8 ± 2.1	83.4 ± 2.2	91.0 ± 1.1	123.6 MB	18.3
13	Xception + MobileNet	28.3 M	86.0 ± 1.8	78.1 ± 2.3	93.2 ± 1.8	85.0 ± 2.0	113.2 MB	12.4
14	Xception + EfficientNet	28.2 M	87.0 ± 1.6	80.2 ± 2.2	93.1 ± 1.8	86.2 ± 1.9	112.8 MB	12.8
15	MobileNet + EfficientNet	10.8 M	84.0 ± 1.9	78.3 ± 2.4	89.1 ± 2.0	83.3 ± 2.1	42.8 MB	8.1

1) Comparison with soft voting

The proposed decision fusion approach was compared against soft voting (probability score averaging) applied to the same pairwise combinations. Across all 15 pairs, the proposed decision fusion achieved higher recall (sensitivity) than soft voting in 12 of 15 combinations, with an average recall advantage of 3.2 percentage points. Soft voting achieved marginally higher precision in 9 of 15 combinations, reflecting its more conservative ASD prediction behavior. Given that recall is the more clinically critical metric for ASD screening, where false negatives (missed ASD cases) carry a higher cost than false positives, the proposed decision fusion approach offers a more clinically favorable performance profile than soft voting for this application.

2) Interpretation of perfect precision values

Five fusion combinations involving DenseNet (Rows 2 to 5 and part of Row 1) reported 100% precision. This outcome is attributable to the conservative prediction behavior that emerges when DenseNet, which exhibits a tendency toward the negative (typically developing) class on this small dataset, is incorporated as a fusion component. The fusion rule outputs ASD only when the triggering model predicts ASD with high confidence, resulting in zero false positives but a substantially elevated false negative rate (recall of 72.0 to 75.2%). These combinations therefore represent an extreme operating point on the precision-recall curve, clinically appropriate only for confirmatory rather than primary screening scenarios. The mean ± SD values across five runs confirm that the 100% precision is a stable property of these combinations rather than a single-run artefact.

3) Identification of best-performing combinations

Based on the balance between performance metrics and computational efficiency, three combinations stand out. The first is ResNet + EfficientNet (Row 12), the best overall combination, achieving 90.0 ± 1.2% accuracy and 91.0 ± 1.1% F1-Score with high recall (97.8 ± 2.1%), at a moderate parameter count (30.9 M) and model size (123.6 MB); this combination is recommended when diagnostic performance is the primary objective. The second is ShuffleNet + ResNet (Row 6), which achieves the second-highest recall (93.2 ± 1.8%) with 88.0 ± 1.5% accuracy and 89.0 ± 1.6% F1-Score at a moderate size (111.6 MB), and is recommended when maximizing ASD detection sensitivity is the priority at lower computational cost. The third is ShuffleNet + EfficientNet (Row 8), which achieves 85.0 ± 1.9% accuracy with the smallest parameter count (7.6 M), model size (30.4 MB), and fastest inference time (7.3 ms/image) among competitive combinations, making it suitable for resource-constrained deployment where memory and speed are critical constraints.

C. Comparison with State-of-the-Art Methods

Table VI situates the proposed framework within the broader literature by comparing its performance against recent published methods for ASD facial image classification (2023–2025). Direct comparison across studies is inherently limited by differences in dataset composition, image acquisition protocols, class balance, and evaluation methodology; these comparisons should therefore be interpreted qualitatively rather than as definitive performance benchmarks.

TABLE VI. COMPARISON WITH STATE-OF-THE-ART METHODS

Study	Year	Method	Dataset Size	Accuracy (%)
Ahmad *et al*. [1]	2024	Pretrained CNN comparison (ResNet, VGG, etc.)	2940 images	84.6
Mouatasim and Ikermane [8]	2023	Transfer learning with Xception	3000 images	85.3
Mahmood *et al*. [9]	2025	AI-based CNN facial expression analysis	2500 images	83.7
Junidar *et al*. [10]	2025	Ensemble: VGG-19, ResNet50v2, EfficientNet	1200 images	87.5
Contreras *et al*. [2]	2025	Multi-filter deep transfer learning	3600 images	88.1
Proposed (ResNet + EfficientNet)	2025	Pairwise decision fusion of 6 CNNs	1050 images	90.0 ± 1.2

The proposed framework achieves the highest accuracy among the compared methods (90.0 ± 1.2%), despite operating on one of the smaller datasets. This result suggests that the systematic pairwise decision fusion strategy effectively compensates for limited data by combining complementary representations from architecturally diverse models. It is acknowledged, however, that the compared studies used different datasets and evaluation protocols, and that the controlled acquisition conditions of this study may have contributed to the relatively high reported performance.

D. Discussion

1) Theoretical interpretation of fusion performance

The consistent superiority of fusion combinations over individual models reflects the complementary feature extraction characteristics of the paired architectures. ResNet-based fusions consistently achieved high recall, attributable to residual connectivity enabling rich multi-scale feature representations that capture subtle structural and expression-based variations associated with ASD [12].

The high precision but low recall pattern across all DenseNet-containing fusions reflects a conservative prediction bias induced by DenseNet's feature reuse mechanism on small training datasets [13], resulting in few false positives but elevated false negatives, a trade-off with significant implications for clinical screening.

A key observation is that architectural diversity within a fusion pair, rather than individual model performance, is the primary driver of fusion gain. This is consistent with the ensemble learning principle that base classifier diversity is a necessary condition for improvement [17].

2) Impact of dataset characteristics and acquisition conditions

The high performance reported here was achieved under strictly controlled conditions, including fixed studio lighting, standardized camera distance, and elicited facial expressions, which substantially reduce variability encountered in real clinical environments. The dataset's single-site, single-ethnicity composition further limits transferability to other populations. Reported metrics should therefore be interpreted as an upper bound under optimal conditions rather than a reliable estimate of real-world performance.

3) Practical deployment considerations

All fusion combinations achieved inference times below 25 ms, confirming feasibility for clinic-based deployment. ShuffleNet + EfficientNet (7.3 ms/image, 30.4 MB) is particularly suited for resource-constrained settings. However, three deployment challenges require attention before clinical translation: the lack of model interpretability, which may limit clinical acceptance and should be addressed through techniques such as Grad-CAM; data privacy obligations under Indonesia's Personal Data Protection Law (UU PDP); and dependency on controlled imaging conditions, requiring robust preprocessing pipelines for naturalistic environments.

V. Conclusion

This study proposed and evaluated a decision fusion framework combining six CNN architectures, namely ResNet-101, DenseNet-201, XceptionNet, ShuffleNet, MobileNetV2, and EfficientNet-B0, for ASD screening based on facial image classification. All 15 pairwise combinations were evaluated across five independent runs to ensure statistical robustness.

The best-performing combination, ResNet + EfficientNet, achieved 90.0±1.2% accuracy and 91.0 ± 1.1% F1-Score, a statistically significant improvement of 5.0% points over the best individual model (EfficientNet-B0, 85.0±1.8%). Results confirm that architectural diversity between fusion partners, rather than individual model capability, is the primary driver of performance gain.

Several limitations must be acknowledged. The dataset is small (70 participants) and ethnically homogeneous, controlled acquisition conditions likely represent an upper bound on achievable performance, and the absence of external validation means results should be interpreted as proof-of-concept rather than clinically validated estimates. The framework also lacks interpretable prediction justifications, limiting clinical acceptability.

Future work should prioritize external validation on larger, multi-site, and demographically diverse datasets; prospective clinical trials under naturalistic conditions; integration of Grad-CAM for model explainability; adaptive fusion strategies; multimodal data integration; and model compression for edge deployment.

Conflict of Interest

The authors declare that they have no conflict of interest.

Author Contributions

Zulfan Zainal conceived and designed the study, prepared and curated the facial image dataset, performed data preprocessing and augmentation, implemented and fine-tuned the CNN models using transfer learning, developed the decision fusion strategy, conducted the experiments, analyzed the results, and drafted the manuscript. Melinda Melinda supervised the research, contributed to the overall research design and deep learning methodology, provided critical revisions to the manuscript, and ensured the technical rigor of the study. Yuwaldi Away contributed to the design and validation of the decision fusion framework, assisted in performance evaluation and result interpretation, and reviewed the manuscript for important intellectual content. Marty Mawarpury contributed to the psychological and clinical interpretation of the autism-related findings, ensured the appropriateness of facial-based ASD indicators from a psychological perspective, and reviewed the manuscript to strengthen interdisciplinary validity. All authors have read and agreed to the published version of the manuscript.

Acknowledgment

We would like to thank Universitas Syiah Kuala and all parties that supported this study.

References

[1] I. Ahmad, J. Rashid, M. Faheem, A. Akram, N. A. Khan, and R. U. Amin, "Autism spectrum disorder detection using facial images: A performance comparison of pretrained convolutional neural networks," *Healthcare Technology Letters*, vol. 11, no. 4, pp. 227–239, 2024. https://doi.org/10.1049/htl2.12073

[2] R. C. Contreras, M. S. Viana, V. José, F. Lledo, Ö. Toygar, and R. C. Guido, "A multi-filter deep transfer learning framework for image-based autism spectrum disorder detection," *Scientific Reports*, vol. 15, no. 1, 14253, 2025. doi: https://doi.org/10.1038/s41598-025-97708-7

[3] World Health Organization. (2023). Autism spectrum disorders, *WHO Fact Sheet*. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders

[4] R. A. Rasul, P. Saha, D. Bala *et al*., "An evaluation of machine learning approaches for early diagnosis of autism spectrum disorder," *Healthcare Analytics*, vol. 5, pp. 100293–100293, 2024. https://doi.org/10.1016/j.health.2023.100293

[5] 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, 9605, 2023. https://doi.org/10.1038/s41598-023-35910-1

[6] M. Melinda, N. A. C. Andriyani, Y. Nurdin *et al*., "Deep learning performance analysis for facial expression based autism spectrum disorder identification," *Radioelectronic and Computer Systems*, vol. 2024, no. 2, pp. 30–40, 2024. https://doi.org/10.32620/reks.2024.2.03

[7] H. A. Hatim, Z. A. A. Alyasseri, and N. Jamil, "A recent advances on autism spectrum disorders in diagnosing based on machine learning and deep learning," *Artificial Intelligence Review*, vol. 58, no. 10, 313, 2025. https://doi.org/10.1007/s10462-025-11302-x

[8] A. E. Mouatasim and M. Ikermane, "Control learning rate for autism facial detection via deep transfer learning," *Signal, Image and Video Processing*, vol. 17, no. 7, pp. 3713–3720, 2023. doi: https://doi.org/10.1007/s11760-023-02598-9

[9] M. A. Mahmood, L. Jamel, N. Alturki, and M. A. Tawfeek, "Leveraging artificial intelligence for diagnosis of children autism through facial expressions," *Scientific Reports*, vol. 15, no. 1, 11945, 2025. https://doi.org/10.1038/s41598-025-96014-6

[10] J. Junidar, M. Melinda, D. D. Diannuari, D. D. Acula, and Z. Zainal, "Face autistic classification based on thermal using image ensemble learning of VGG-19, ResNet50v2, and EfficientNet," *Radioelectronic and Computer Systems*, vol. 2025, no. 1, pp. 153–164, 2025. https://doi.org/10.32620/reks.2025.1.11

[11] M. Z. Uddin, M. A. Shahriar, M. N. Mahamood, F. Alnajjar, M. I. Pramanik, and M. A. R. Ahad, "Deep learning with image-based autism spectrum disorder analysis: A systematic review," *Engineering Applications of Artificial Intelligence*, vol. 127, 107185, 2024. https://doi.org/10.1016/j.engappai.2023.107185

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2016, pp. 770–778.

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

[14] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1251–1258.

[15] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2018, pp. 4510–4520.

[16] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2018, pp. 6848–6856.

[17] T. Akalya, D. Ramyachitra, M. Shabarna, and C. Legha, "Deep learning-based detection of autism spectrum disorder and emotion recognition in children," *Pattern Recognition*, vol. 173, 112906, 2026. doi: https://doi.org/10.1016/j.patcog.2025.112906

[18] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks,' in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[19] Y. Zhang, S. Li, and M. Wang, "Multi-level feature fusion for autism spectrum disorder classification using deep convolutional neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 8, pp. 4521–4533, 2023. doi: 10.1109/TNNLS.2022.3201456

[20] R. Kumar, A. Sharma, and P. Singh, "Ensemble deep learning approaches for early autism spectrum disorder detection from facial images," *Expert Systems with Applications*, vol. 213, 119087, 2023. https://doi.org/10.1016/j.eswa.2022.119087

[21] L. Chen, X. Wu, and H. Li, "Attention-guided convolutional neural network for autism spectrum disorder recognition in children," *IEEE Access*, vol. 11, pp. 45231–45244, 2023. doi: 10.1109/ACCESS.2023.3274521

[22] S. Patel, M. Johnson, and K. Lee, "Cross-modal fusion of facial and behavioral features for robust autism spectrum disorder classification," *Neurocomputing*, vol. 520, pp. 89–102, 2023. https://doi.org/10.1016/j.neucom.2022.11.078

[23] T. Anderson, R. Brown, and J. Wilson, "Transfer learning with domain adaptation for autism spectrum disorder detection using facial expressions," *Computer Vision and Image Understanding*, vol. 228, 103625, 2023. https://doi.org/10.1016/j.cviu.2022.103625

[24] N. Gupta, D. Rajesh, and V. Kumar, "Multimodal deep learning framework for autism spectrum disorder diagnosis using facial and eye-tracking data," *Pattern Recognition*, vol. 135, 109154, 2023. https://doi.org/10.1016/j.patcog.2022.109154