Autism Spectrum Disorder Detection from facial images using Deep Convolutional Neural Networks

Marwa Ghrouda
NOCCS Laboratory
ENISo, University of Sousse
Sousse, Tunisia
marwaghrouda@gmail.com

Hend Basly
NOCCS Laboratory
ENISo, University of Sousse
Sousse, Tunisia
basly.hend@gmail.com

Fatma Ezahra Sayadi
NOCCS Laboratory
ENISo, University of Sousse
Sousse, Tunisia
fatmazahra.sayadi@eniso.u-sousse.tn

***Abstract*—Autism Spectrum Disorder (ASD) is a prevalent neurodevelopmental condition for which early screening is essential to support timely intervention. Conventional diagnostic procedures are often time-consuming and resource-intensive, highlighting the need for accessible and cost-effective screening approaches. This study investigates the use of deep learning–based facial image analysis as a supportive tool for ASD screening rather than as a diagnostic solution. A comparative evaluation of three state-of-the-art convolutional neural network architectures EfficientNetB0, MobileNetV2, and ResNet50 is conducted using a unified experimental framework based on transfer learning and fine-tuning. Experiments are performed on a publicly available Autistic Children Facial Image Dataset, with an additional verification step to prevent duplicate images in order to enhance generalization and reduce potential bias. All models are trained and evaluated under identical conditions to ensure a fair comparison. Experimental results indicate that EfficientNetB0 achieves the best overall performance, reaching a test accuracy of 88%, while maintaining a favorable balance between model complexity and generalization ability.**

***Index Terms*—Autism Spectrum Disorder (ASD), Binary Classification, Convolutional Neural Networks (CNN), Facial Images, Deep Learning, Fine-Tuning.**

I. Introduction

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition characterized by persistent challenges in social communication, social interaction, and behavioral patterns. According to the Centers for Disease Control and Prevention (CDC), ASD affects approximately one in 54 children in the United States, highlighting its significant prevalence and public health impact [1]. Early identification is crucial, as it enables timely interventions that can improve developmental outcomes, communication skills, and autonomy. Yet clinical diagnosis requires specialized expertise, standardized assessments, and multidisciplinary evaluation, making diagnosis time-consuming and challenging. In this context, computer-aided screening tools can provide complementary information to clinicians without replacing formal diagnosis. Recently, machine learning has emerged as a promising approach for supporting early detection by analyzing complex behavioral, textual, and image-based data, thereby assisting healthcare professionals in decision-making and optimizing the diagnostic process [2]. Beyond medical diagnosis, deep learning has been successfully applied in various domains such as robotics and intelligent systems [3] and shows high potential in neurodegenerative disease analysis. For instance, hybrid CNN–SVM frameworks for Alzheimer's disease detection leverage deep convolutional features and SVM discriminative power to achieve enhanced classification accuracy and robustness [4], [5].

This versatility underscores deep learning's adaptability to complex biomedical imaging tasks, motivating its application to ASD detection from facial images. DL models analyzing facial features offer a non-invasive, cost-effective, and rapid screening alternative that addresses key limitations of conventional diagnostic protocols [6].

Studies indicate that children with ASD may exhibit distinctive facial features, positioning facial image analysis as a complementary supportive tool for early detection [6], [7]. Despite promising results, ASD screening via facial images remains underexplored. Existing deep learning approaches face challenges such as overfitting, high computational costs, and limited sensitivity to subtle facial cues. Furthermore, reliance on small or imbalanced datasets often compromises generalizability, highlighting the need for more robust, efficient, and scalable models for reliable ASD classification [1].

In this paper, we propose a deep learning framework for ASD screening from children's facial images. We conduct a comparative study of three pre-trained convolutional neural network architectures EfficientNetB0 [6], ResNet50 [8], and MobileNetV2 [1] using transfer learning and fine-tuning for binary ASD classification on the publicly available Autistic Children Facial Image Dataset [9]. Unlike many existing studies that focus primarily on optimizing classification accuracy, our work adopts a more rigorous and realistic methodological approach to ASD screening from facial images. The main contributions of this study are threefold: (i) a systematic and fair comparison of EfficientNetB0, ResNet50, and MobileNetV2 on a public dataset to reduce bias and information leakage; (ii) a joint analysis of classification performance and computational efficiency, crucial for scalable and accessible screening applications; and (iii) an empirical justification for selecting EfficientNetB0, achieving the best trade-off between accuracy, generalization stability, and model complexity.

Overall, this study provides a cautious and representative evaluation of the capabilities of deep learning models for automated ASD screening. It offers a non-invasive, cost-effective, and scalable tool that can complement conventional diagnostic procedures and support earlier detection and intervention, without replacing formal clinical diagnosis.

The remainder of this paper is organized as follows: Section II reviews the related work. Section III describes the dataset used in this study. Section IV presents the proposed methodology, while Section V details the implementation aspects. Section VI reports and discusses the experimental results and performance analysis. Finally, Section VII concludes the paper.

II. Related Work

Research into the early detection of autism has largely focused on identifying early indicators in infants who are at a higher risk, well before the typical age for diagnosis. Pioneering research by Chawarska et al. [10] highlighted that reduced social attention, particularly marked by decreased eye contact and limited interest in moving social environments, can differentiate infants as young as 6 months old who will later receive an ASD diagnosis. Moreover, moving beyond just observable behaviors, researchers like [11] uncovered neural patterns; they found that non-linear electroencephalography (EEG) responses to sounds at 6 and 12 months change as the child grows and may offer predictive insights. The significance of early interactions is underscored by studies focusing on dyadic synchrony [12]. These studies show that infants who develop ASD often display notably different reactivity and engagement patterns with their mothers by the age of 12 months. To apply these insights, specialized screening tools, such as the "Get SET Early" model [13] and the Early Screening for Autism and Communication Disorders (ESAC) [14], have been created and effectively tested for identifying autistic traits in children as early as 12 months of age. This confirms that community-based screening is indeed feasible on a larger scale.

Despite the promise of these infant studies, a major translational barrier exists: the development of robust AI models is hampered by the scarcity and prohibitive cost of acquiring large-scale, high-quality infant data, which is inextricably linked to expensive, gold-standard clinical assessments. Consequently, research has pivoted toward analyzing static facial images of older children (2–14 years) using more accessible public datasets.

Previous studies in this area have mainly applied transfer learning with convolutional neural networks (CNNs), yielding strong performance. For instance, Rahman et al. [6] achieved over 94% accuracy using architectures like Xception. Acknowledging potential dataset biases, Khosla et al. [15] introduced rigorous pre-processing, including image de-duplication, to achieve a more conservative and robust accuracy of 87%. Furthermore, Alam and Rashid [8] recently addressed data heterogeneity and privacy concerns by proposing a federated learning framework, showcasing the field's evolution towards more scalable, secure and ethically responsible methodologies.

While these studies demonstrate the potential of deep learning for ASD screening using facial images, they are limited by small or imbalanced datasets, potential overfitting, and reliance on single metrics like accuracy. These limitations highlight the need for more systematic comparisons, robust validation, and methodological rigor, which is the focus of the present work.

Despite their promise, these studies are limited by small or imbalanced datasets, limited clinical validation, and low model explainability. Our work is positioned within this context. We contribute to this existing body of research by conducting a systematic comparative analysis of three modern, efficient architectures (EfficientNetB0, ResNet50, and MobileNetV2) on a selected public Kaggle dataset. This study aims to provide a clear empirical benchmark of accuracy and loss metrics for these models, offering a clearer understanding of how these models compare in detecting ASD based on facial imagery and establishing a baseline for future development of efficient and accessible screening tools. The following section describes the dataset used in this study.

III. Dataset Description

The dataset comprises 2,926 facial images of children (2–14 years) evenly split between autistic and non-autistic classes. It is pre-partitioned into training (2,526 images), validation (200), and test (200) subsets, all in JPEG format with balanced class distribution (50/50) across splits. This publicly available Kaggle dataset [9], commonly used in prior facial-image-based ASD screening studies, offers a structured and balanced class organization that enables controlled deep learning evaluation. However, like most public datasets, it remains limited in size and diversity limitations that should be acknowledged when interpreting results. Representative facial images from both classes are shown in Figure 1.

IV. Proposed Method

The proposed methodology for ASD screening from facial images comprises four sequential steps: image preprocessing, deep feature extraction via pre-trained CNNs, binary classification, and quantitative performance assessment. This end-to-end framework aims to deliver an accurate, non-invasive screening tool and is illustrated in Figure2.

A. Data Pre-processing

As described in Section III, the dataset is organized into training, validation, and test subsets, facilitating controlled and unbiased model evaluation. The training set is used to update model parameters, the validation set guides hyperparameter tuning, and the test set provides an unbiased assessment of model performance on unseen data.

Prior to training, all images are resized to 224×224 pixels to match the input requirements of the pre-trained CNNs (EfficientNetB0, MobileNetV2, ResNet50) and pixel values are normalized to stabilize and accelerate learning. To improve model generalization and mitigate the limitations of the relatively small dataset, data augmentation was applied to the training set. The applied transformations include rotations of up to $\pm 20°$ to simulate slight head tilts, width and height shifts of up to 10% to account for positional variations, shearing of up to 10% to simulate minor perspective changes, zooming within $\pm 15\%$ to handle distance variations, horizontal flipping to simulate left-right symmetry, and brightness adjustments ranging from 80% to 120%. These augmentations increase the diversity of training data, enabling the models to learn more invariant facial representations and improving classification stability and accuracy. No duplicate images were present in the dataset, so no de-duplication was necessary. Since the training set is balanced, no class weighting was applied.

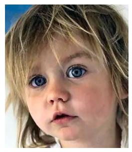
B. Feature Extraction Using Pre-trained CNNs

Extracting features using pre-trained CNNs has proven to be an effective and computationally efficient strategy in medical image analysis, including facial image–based ASD screening. This approach relies on deep architectures pre-trained on large-scale datasets such as ImageNet [1], allowing the reuse of hierarchical visual representations learned from millions of images. These representations capture low-level features (e.g., edges, textures, and color patterns) as well as higher-level abstractions related to facial structure and appearance. Such transferable features provide a solid basis for learning discriminative patterns that may support the classification of autistic versus non-autistic facial images. In this study, the three well-established CNN architectures MobileNetV2, EfficientNetB0, and ResNet50, are employed as feature extractors. Each model is initialized with ImageNet pre-trained weights and adapted by removing its original classification head while retaining the convolutional backbone for feature extraction. The hierarchical feature representations produced by each CNN are then fed into task-specific classification layers for binary ASD screening, as described in section IV-C. During the initial training stage, the convolutional layers are frozen to preserve the generic visual features learned from large-scale natural image data. In a subsequent fine-tuning stage, only the upper layers of the network are unfrozen to adapt the learned representations to the ASD facial image dataset. This two-stage transfer learning strategy reduces training time and helps mitigate overfitting, which is particularly important given the limited size of the dataset.

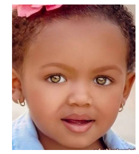
- **MobileNetV2**: Inverted residual architecture with depthwise separable convolutions enables robust classification with minimal computational cost, suitable for lightweight ASD screening [1].
- **EfficientNetB0**: Compound scaling of depth, width, and resolution achieves competitive accuracy with few parameters [6].
- **ResNet50**: Residual skip connections mitigate gradient vanishing and enable robust extraction of discriminative facial features [8].

C. Fine-Tuning and Binary Classification

After feature extraction, the pre-trained CNNs (MobileNetV2, EfficientNetB0, ResNet50) are fine-tuned for binary ASD classification on facial images, providing a screening tool for detecting autistic traits without replacing clinical diagnosis. Initially, convolutional bases are frozen to preserve ImageNet-learned visual representations while training only the newly added classifier. Subsequently, selected upper layers are unfrozen and jointly trained with the classifier to enable domain-specific adaptation to subtle facial patterns relevant for ASD screening. The classification head consists of a Global Average Pooling (GAP) layer, a Dense layer with 128 ReLU units, a Dropout layer (rate = 0.4) to reduce overfitting, and a final sigmoid-activated Dense layer that outputs the probability of each class. Early-stopping and checkpoint strategies are used to retain the best-performing weights. This two-stage fine-tuning framework combines transfer learning with targeted adaptation of higher-level features, ensuring stability while allowing the model to capture subtle ASD-specific facial characteristics, leading to robust and accurate classification. These techniques are particularly suitable for autism-related studies, as facial cues associated with ASD are subtle and datasets are



(a) Autistic



(b) Non-autistic

Fig. 1: Sample facial images from the Autistic Children Facial Image Dataset [9], illustrating the two image classes: (a) image from the autistic class; (b) image from the non-autistic class.

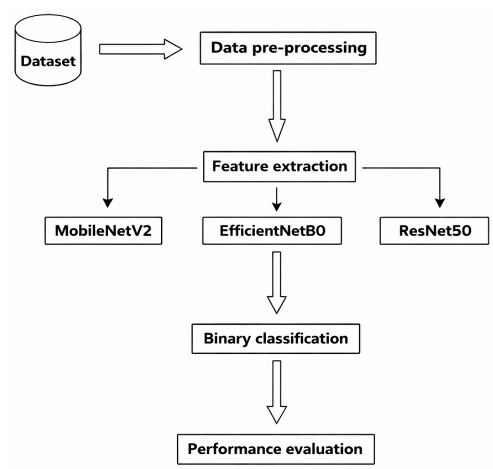


Fig. 2: Flowchart of the proposed methodology.

often small, requiring models that can leverage pre-trained knowledge while adapting to domain-specific features.

V. Implementation details

All experiments were conducted using TensorFlow and Keras on a workstation equipped with an Intel Core i9-14900KF CPU (8 Performance cores + 16 Efficient cores, 32 threads) and an NVIDIA GeForce RTX 3060 GPU with 12 GB of GDDR6 memory. Training and validation were performed independently for each model using the same dataset splits, ensuring a fair and consistent comparison across architectures. Training time per epoch varied depending on the architectural complexity and number of parameters of each model: MobileNetV2 exhibited the shortest training time per epoch, followed by EfficientNetB0, while ResNet50 required the longest time due to its deeper architecture and higher parameter count.

All models were trained using a transfer learning strategy. In the initial phase, only the classification head was trained for 15 epochs while keeping the convolutional base frozen to preserve generic visual features. During fine-tuning, the last 30–40 layers were unfrozen and trained jointly with the classifier for 25 epochs at a reduced learning rate of 1×10^{-4} (initial: 1×10^{-3}). Early-stopping and model checkpoint strategies were applied, monitoring the validation loss, to retain the best-performing weights and reduce overfitting. This progressive training strategy stabilizes optimization while allowing adaptation to domain-specific facial features relevant for ASD screening.

VI. Results and discussion

The performance of the three proposed models EfficientNetB0, MobileNetV2, and ResNet50 was evaluated on an independent test set using multiple quantitative metrics. The accuracy and loss were used as the primary indicators of model generalization performance. To provide a more comprehensive evaluation, confusion matrices and classification reports were generated, including precision, recall, and F1-score for each class. All models were trained and evaluated under the same experimental framework to ensure a fair and consistent comparison, highlighting the comparative strengths of each architecture in the context of ASD screening rather than clinical diagnosis.

Table I summarizes the performance of the three models in terms of training, validation, and testing accuracy after a total of 40 epochs. These results reflect the models' ability to generalize to unseen data within the constraints of the limited, publicly available dataset, and are intended to evaluate the models as screening tools rather than diagnostic systems.

TABLE I: Training, validation, and testing accuracies obtained on using different models.

Model	**Training**	**Validation**	**Testing**
ResNet50	95%	84%	84%
EfficientNetB0	94%	87%	88%
MobileNetV2	91%	84%	84%

From Table I, it is clear that EfficientNetB0 achieved the highest validation accuracy, closely followed by ResNet50 and MobileNetV2, while lightweight and computationally efficient, showed slightly lower training and validation accuracy but remained competitive on the testing set. These differences reflect the ability of each architecture to generalize to unseen data. Importantly, these evaluations are intended for screening purposes and not for clinical diagnosis. Fine-tuning allowed

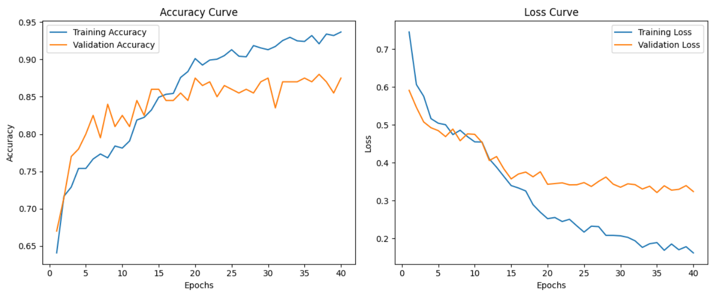


Fig. 3: Accuracy and loss curves for EfficientNetB0.

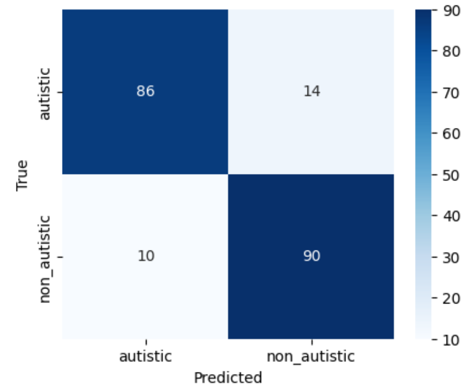


Fig. 4: Confusion matrix of EfficientNetB0 on the test set.

the networks to re-learn dataset-specific patterns, improving performance while avoiding overfitting. Early stopping based on validation loss was applied with a patience of 7 epochs to ensure robust learning and convergence. Table I also highlights the effect of fine-tuning, as models improved substantially from head-only training to full fine-tuning, demonstrating the importance of adapting pre-trained weights to the dataset. This also serves as an ablation insight, showing how dataset specific adaptation impacts performance. For the best-performing model, EfficientNetB0, the final test loss was 0.2969, with a test accuracy of 0.88. Additional evaluation metrics were computed to provide a more detailed assessment of classification performance. The model achieved a precision of 0.90 for autistic and 0.87 for non-autistic, a recall of 0.86 for autistic and 0.90 for non-autistic, and an overall F1-score of 0.88 for both classes. The macro and weighted averages of precision, recall, and F1-score were all 0.88, highlighting the model's balanced performance across classes. These results align with the accuracy and loss curves in Figure 3, which show steadily increasing training and validation accuracy alongside decreasing loss. The minimal gap between training and validation curves indicates strong generalization, while the smooth convergence demonstrates that early stopping and learning rate reduction effectively prevented overfitting confirming EfficientNetB0's reliability for ASD screening. To further evaluate classification performance, the confusion matrix on the test set is presented in Figure 4. This visualization highlights the model's ability to correctly classify ASD and non-ASD instances, complementing the accuracy and loss curves.

All three studies including [1], [16], and our work utilize the same Kaggle ASD facial image dataset [9]. However, as summarized in Table II, prior works [1], [16] are limited to same-family comparisons under non-identical protocols, whereas our strictly identical protocol using the same splits, augmentation, and fine-tuning strategy across all three architectures ensures a fair assessment of the accuracy/efficiency trade-off essential for clinical deployment.

Furthermore, interpretability methods like Grad-CAM pinpoint the facial regions that contribute to the predictions, shedding light on the model's decision-making process and fostering greater confidence among clinicians in AI-assisted screening systems. Finally, while the results are promising, future work should focus on cross-dataset validation or k-fold cross-validation in order to better evaluate the generalizability and robustness of the models across diverse populations. Moreover, the use of face detection methods such as MTCNN [8] could further improve preprocessing consistency and ensure that models focus on the most relevant facial regions, potentially enhancing screening performance across datasets.

To further evaluate the model's practical implications, it is important to note that, from an ethical and clinical perspective, the proposed model is intended as a screening aid for early detection of autism ASD and cannot replace a medical diagnosis provided by qualified healthcare professionals. The use of static facial images from a public dataset also exposes the model to potential biases related to children's age, ethnicity, and image acquisition conditions (e.g., lighting, pose, quality). Consequently, any real-world application would require rigorous clinical validation on diverse and representative data, as well as strict supervision by medical experts, to ensure responsible and ethical use.

VII. Conclusion

This study demonstrates EfficientNetB0's effectiveness for facial image classification, achieving 88% test accuracy on a clean dataset of 2,926 unique images. A key strength was leveraging the naturally duplicate-free dataset structure combined with real-time data augmentation, ensuring reliable and generalizable performance.

TABLE II: Comparison with related works on the common Kaggle ASD dataset [9]

Study	Models compared	Best model (test accuracy)	Protocol
Chandra et al. [1]	VGG16 VGG19	VGG19 (82.5%)	Modified protocol same-family comparison
Li et al. [16]	MobileNetV2 MobileNetV3-Large	MobileNetV3-Large (90.5%)	Sequential protocol same-family comparison
Ours	EfficientNetB0 ResNet50 MobileNetV2	**EfficientNetB0 (88.0%)**	Identical protocol cross-family comparison

However, the primary limitation remains the modest dataset size ($\sim$2,500 training images). Future work will prioritize dataset expansion through multi-source image collection and integration of explainable AI techniques (e.g., Grad-CAM) to enhance interpretability for real-world deployment.

REFERENCES

[1] R. Chandra, S. Tiwari, A. Kumar, S. Agarwal, M. Syafrullah, and K. Adiyarta, "Autism spectrum disorder detection using autistic image dataset," in *2023 10th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*. IEEE, 2023, pp. 54–59.

[2] S. Dodia, V. Meshram, J. Kasle, S. Gomase, H. Amrit, and R. Sarse, "Autism spectrum disorder (asd) detection from facial images using mobilenet," in *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*. IEEE, 2024, pp. 1–7.

[3] H. Basly, M. A. Zayene, and F. E. Sayadi, "Spatiotemporal self-attention mechanism driven by 3d pose to guide rgb cues for daily living human activity recognition," *Journal of Intelligent & Robotic Systems*, vol. 109, no. 1, p. 2, 2023.

[4] M. A. Zayene, H. Basly, and F. E. Sayadi, "Alzheimer's disease detection based on features level selection for cnn-svm combination," in *2023 IEEE International Conference on Artificial Intelligence & Green Energy (ICAIGE)*. IEEE, 2023, pp. 1–6.

[5] S. B. H. Salah, M. Chouchene, M. A. Zayene, and F. E. Sayadi, "Classification of alzheimer's diseases from pet images using a convolutional neural network," in *2024 International Conference on Control, Automation and Diagnosis (ICCAD)*. IEEE, 2024, pp. 1–5.

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

[7] Z. A. Ahmed, T. H. Aldhyani, M. E. Jadhav, M. Y. Alzahrani, M. E. Alzahrani, M. M. Althobaiti, F. Alassery, A. Alshaflut, N. M. Alzahrani, and A. M. Al-Madani, "[retracted] facial features detection system to identify children with autism spectrum disorder: Deep learning models," *Computational and Mathematical Methods in Medicine*, vol. 2022, no. 1, p. 3941049, 2022.

[8] S. Alam and M. M. Rashid, "Enhanced early autism screening: Assessing domain adaptation with distributed facial image datasets and deep federated learning," *IIUM Engineering Journal*, vol. 26, no. 1, pp. 113–128, 2025.

[9] I. Khan, "Autistic children facial image dataset," 2022, 2926 images; balanced classes autistic/non-autistic children aged approx. 2–14 yrs.

[10] K. Chawarska, S. Macari, and F. Shic, "Decreased spontaneous attention to social scenes in 6-month-old infants later diagnosed with autism spectrum disorders," *Biological psychiatry*, vol. 74, no. 3, pp. 195–203, 2013.

[11] F. C. Peck, L. J. Gabard-Durnam, C. L. Wilkinson, W. Bosl, H. Tager-Flusberg, and C. A. Nelson, "Prediction of autism spectrum disorder diagnosis using nonlinear measures of language-related eeg at 6 and 12 months," *Journal of neurodevelopmental disorders*, vol. 13, no. 1, p. 57, 2021.

[12] A. M. Kellerman, A. Schwichtenberg, R. Abu-Zhaya, M. Miller, G. S. Young, and S. Ozonoff, "Dyadic synchrony and responsiveness in the first year: Associations with autism risk," *Autism Research*, vol. 13, no. 12, pp. 2190–2201, 2020.

[13] P. Karen, G. Vahid, B. Elizabeth, C. Eric, C. Amanda, C. B. Cynthia, N. Srinivasa, C. Debra, A. Steven, L. Linda, P. Christie, G. Kim, G. Gohar, C.-C. Terri, and K. Kathy, "Get set early to identify and treatment refer autism spectrum disorder at 1 year and discover factors that influence early diagnosis," *The Journal of pediatrics*, vol. 236, pp. 179–188, 2021.

[14] M. W. Amy, G. Whitney, L. H. Jessica, D. Abigail, N. D. Taylor, W. Juliann, P. Karen, S. M. Stacy, T. Audrey, O. Sally, P. Eva, and L. Catherine, "The early screening for autism and communication disorders: field-testing an autism-specific screening tool for children 12 to 36 months of age," *Autism*, vol. 25, no. 7, pp. 2112–2123, 2021.

[15] Y. Khosla, P. Ramachandra, and N. Chaitra, "Detection of autistic individuals using facial images and deep learning," in *2021 IEEE international conference on computation system and information technology for sustainable solutions (CSITSS)*. IEEE, 2021, pp. 1–5.

[16] Y. Li, W.-C. Huang, and P.-H. Song, "A face image classification method of autistic children based on the two-phase transfer learning," *Frontiers in Psychology*, vol. 14, no. 1226470, 2023.