Exploring Autism Spectrum Disorders Classification Using Facial Images and Deep Learning

*Note: Sub-titles are not captured in Xplore and should not be used

Maybel Herrera
School of Information Studies
Syracuse University
Syracuse, NY, USA
maherrer@syr.edu

Abstract - **Autism Spectrum Disorder (ASD) is diagnosed primarily through behavioral assessment, and there is limited understanding of whether facial image data can support automated analysis in a reliable and interpretable way. This study examines whether deep learning models can classify facial images of autistic and non-autistic children using a controlled image dataset. A publicly available Kaggle dataset was utilized, with the raw version selected to allow full control over preprocessing steps, including image extraction, resizing, augmentation, normalization, and dataset splitting. A custom convolutional neural network and a MobileNetV2 transfer learning model were trained and evaluated using accuracy, confusion matrices, and receiver operating characteristic area under the curve (ROC–AUC) metrics. While both models were able to learn visual patterns within the dataset, the results reveal limitations in generalization and sensitivity to dataset characteristics such as image quality and facial variability. These findings indicate that facial image-based classification remains exploratory in nature rather than diagnostic.**

Keywords—autism spectrum disorder, deep learning, facial image classification, convolutional neural networks, transfer learning

I. Introduction

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition that affects how individuals communicate, interact socially, and process sensory information. Diagnosis is typically based on behavioral assessments conducted by trained professionals rather than biological tests or visible physical markers. Although these diagnostic approaches are well established, they can be time-consuming and may delay access to support and intervention services for some individuals.

ASD presents differently across individuals, and not all children on the autism spectrum exhibit highly noticeable behavioral or sensory challenges early in life. Some experience more subtle difficulties, particularly in areas such as social communication, forming friendships, or maintaining interpersonal relationships. Because these challenges may be less apparent, some individuals remain undiagnosed during childhood and seek evaluation later in life, often after recognizing autistic traits in themselves during adolescence or adulthood. This variability highlights both the complexity of ASD and the limitations of relying solely on observable behavioral indicators.

In recent years, researchers have explored whether computational approaches, particularly computer vision and deep learning techniques, can identify statistical patterns associated with ASD in facial images. Prior work has applied convolutional neural networks to static facial images, demonstrating that deep learning models can learn subtle visual patterns that differ between autistic and non-autistic populations within controlled datasets [1]. In addition, reviews of craniofacial research suggest that facial morphology may vary statistically in individuals with ASD, providing further motivation for exploratory image-based analysis [2]. Earlier foundational studies also reported small but statistically significant group-level differences in facial structure between children with ASD and typically developing peers [3]. Public-facing clinical discussions have further contributed to interest in potential craniofacial characteristics associated with autism, helping to contextualize ongoing research in this area [4].

Despite this growing interest, facial features are not used clinically to diagnose autism, and facial appearance can vary widely due to factors such as age, ethnicity, lighting conditions, pose, and image quality. As a result, facial image analysis should be viewed as an exploratory research direction rather than a diagnostic approach.

In this study, deep learning models are used to examine whether facial images can be classified as autistic or non-autistic within a controlled dataset. A publicly available Kaggle dataset containing facial images of autistic and non-autistic children is utilized, with the raw version selected to allow full control over preprocessing. By comparing a custom convolutional neural network with a MobileNetV2 transfer learning model, this work aims to evaluate both the potential and the limitations of facial image-based ASD classification.

II. Literature Review

Interest in the relationship between facial characteristics and Autism Spectrum Disorder (ASD) predates the use of machine learning. Early research examined whether facial structure differed in measurable ways between autistic and neurotypical populations, focusing on group-level patterns rather than individual diagnosis. One of the earliest studies in this area was conducted by Aldridge et al. [3], who analyzed craniofacial characteristics in prepubertal boys with ASD. Rather than identifying a single or consistent facial profile, the study reported small but statistically significant differences in facial proportions across subgroups of children with ASD when compared to typically developing peers. These findings highlighted substantial variability within the autism spectrum and suggested that any facial differences occur at a population level rather than appearing uniformly across individuals.

This line of inquiry is supported by developmental research indicating that the brain and facial structures develop through closely linked biological mechanisms. Within this context, some studies have reported population-level trends in facial morphology among individuals with ASD, such as a relatively broader upper face, shorter midface region, wider-set eyes, a wider mouth, and differences in the philtrum. However, these characteristics are subtle, variable, and influenced by multiple factors, reinforcing the complexity of interpreting facial features in relation to neurodevelopment.

As additional studies emerged, researchers began to examine craniofacial characteristics more broadly. Quatrosi et al. [2] reviewed a wide range of studies focused on facial morphology in individuals with ASD and found that reported differences were inconsistent across populations. Their review emphasized that results varied substantially depending on age, demographic composition, image quality, and study design. Rather than pointing to clear or universal facial patterns, this body of work underscored the heterogeneity of ASD and the challenges of drawing consistent conclusions from facial analysis alone.

More recently, advances in computer vision and deep learning have shifted attention toward automated analysis of facial images. Instead of relying on manually defined measurements, convolutional neural networks have been used to learn visual patterns directly from image data. Rahman et al. [1] applied deep learning models to static facial images of autistic and non-autistic children and demonstrated that these models could learn features that supported classification within a controlled dataset. At the same time, their findings highlighted important challenges related to dataset size, image variability, and generalization to unseen data, illustrating both the potential and the limitations of image-based approaches.

Beyond academic research, public-facing clinical discussions have also contributed to interest in facial characteristics and autism. Organizations such as Advanced Autism Services provide summaries of existing research and clinical observations for general audiences [4]. While these sources are not peer-reviewed, they reflect broader curiosity surrounding the topic and help contextualize why facial features continue to be explored when combined with rigorous scientific methods.

Taken together, prior work suggests that facial image analysis may capture subtle statistical patterns associated with ASD at a population level, but these patterns are highly variable and sensitive to methodological choices. This body of research provides motivation for continued exploration using modern deep learning techniques, while also emphasizing the importance of careful interpretation. Building on this foundation, the present study investigates facial image-based classification using deep learning models within a controlled dataset, with the goal of understanding both model behavior and practical limitations.

III. Data Acquisition

The dataset used in this study was obtained from a publicly available Kaggle repository focused on autism spectrum detection using facial images. The dataset includes facial photographs of children labeled as autistic and non-autistic and is widely used in exploratory research related to image-based ASD classification.

Rather than manually uploading compressed files, the dataset was downloaded directly within the Google Colab environment using the kagglehub library. This approach ensured a reproducible workflow and eliminated the need for manual file handling. The raw version of the dataset was intentionally selected to allow full control over data organization, inspection, and preprocessing.

Within the downloaded dataset, an old dataset directory was identified containing two subfolders: Autistic and Non-Autistic. Image file paths from both classes were programmatically collected, and a panda DataFrame was constructed to store file locations and corresponding labels. This structure allowed for efficient inspection and validation of the dataset before model training.

An initial review of the dataset showed a total of 3,620 images, with 1,887 images labeled as autistic and 1,733 labeled as non-autistic, indicating a relatively balanced class distribution. No missing labels or file paths were detected. A small sample of images was visually inspected to confirm label integrity and to assess variability in image resolution, lighting conditions, and facial positioning. Image dimensions were found to vary across samples, reinforcing the need for resizing and normalization during preprocessing.

These preliminary checks helped ensure that the dataset was correctly loaded, labeled, and suitable for downstream modeling, while also highlighting characteristics that informed later preprocessing and modeling decisions.

IV. Methodology

Preprocessing and Data Set Preparation

Prior to model training, exploratory data analysis (EDA) was conducted to understand the structure and quality of the dataset. Image file paths and labels were examined to assess class distribution, which showed a relatively balanced dataset between autistic and non-autistic images. A bar chart was used to visualize label counts, and no missing file paths or labels were detected.

To assess image quality and variability, a small random sample of images from both classes was visually inspected. This inspection verified label consistency and revealed variation in lighting conditions, facial alignment, image resolution, background, and facial pose, reflecting the unconstrained nature of the dataset.

Image metadata inspection confirmed that all sampled images were already stored in RGB format, eliminating the need for color space conversion. However, image dimensions varied substantially across samples. To ensure consistent input to the deep learning models, all images were resized to

224 × 224 pixels during dataset loading. Pixel intensity normalization was applied within the model architecture using a rescaling layer that mapped pixel values from the range [0, 255]to [0, 1], helping stabilize training while keeping preprocessing integrated into the model pipeline.

After EDA and preprocessing decisions were finalized, the dataset was split into training, validation, and test sets using TensorFlow's image_dataset_from_directory utility. A 70% training split was created, with the remaining 30% divided evenly to form validation and test sets. Shuffling and a fixed random seed were used to support reproducibility and balanced class representation. Images were loaded in batches using TensorFlow datasets, which served as the input pipeline during training and evaluation.

Data Augmentation

To reduce overfitting and improve model robustness, data augmentation was applied during training. Augmentation was implemented using a Keras Sequential pipeline and included random horizontal flipping, small rotations, and random zoom. These transformations were intentionally kept mild to preserve facial structure while introducing natural variation. Augmentation was applied only to the training set, while validation and test data were left unchanged to ensure fair performance evaluation.

CNN Model Architecture (From Scratch)

A custom convolutional neural network was developed from scratch to serve as a baseline deep learning model for facial image classification. The model accepts RGB images resized to 224 × 224 pixels as input. Data augmentation was applied within the model pipeline and was active only during training.

The architecture consists of a series of convolutional blocks with increasing filter sizes (32, 64, 128, and 256), each using 3 × 3 kernels with ReLU activation and same padding. Max-pooling layers were used after each convolutional block to progressively reduce spatial dimensions. Dropout layers were included after selected convolutional blocks to help mitigate overfitting.

Following the convolutional layers, the feature maps were flattened and passed to a fully connected dense layer with 128 units and ReLU activation. A high dropout rate was applied before the final classification layer to further reduce overfitting. The output layer consists of a single neuron with a sigmoid activation function, enabling binary classification between autistic and non-autistic images.

The model was trained using the Adam optimizer and binary cross-entropy loss. Training was performed for up to 40 epochs with early stopping based on validation loss to prevent overfitting. Early stopping was configured with a patience of five epochs and restored the best model weights. In addition, a learning rate reduction strategy was applied using a ReduceLROnPlateau callback, which lowered the learning rate when validation loss stopped improving. These strategies helped stabilize training and improve generalization performance.

Transfer Learning Model (MobileNetV2)

To compare performance against pretrained architecture, a transfer learning model based on MobileNetV2 was implemented. MobileNetV2 was selected due to its efficiency, lightweight design, and strong performance on image classification tasks, making it well-suited for relatively small datasets.

The pretrained MobileNetV2 model, initialized with ImageNet weights, was used as a feature extractor by excluding the original classification head. During the first training phase, all convolutional layers in the base model were frozen to preserve learned low-level and mid-level visual features. A custom classification head was added, consisting of global average pooling, a dense layer with 128 units and ReLU activation, dropout for regularization, and a final sigmoid output layer for binary classification.

Data augmentation was applied at the input level during training, and images were preprocessed using the MobileNetV2-specific preprocessing function. The model was initially trained using the Adam optimizer with a learning rate of 1e−3, focusing only on the newly added top layers.

In the second training phase, fine-tuning was performed by unfreezing a small subset of 30 layers from the final layers of the MobileNetV2 base network. The model was recompiled with a lower learning rate (5e−5) to allow gradual adaptation of higher-level features without disrupting pretrained weights. Early stopping, learning rate reduction on plateau, and model checkpointing were used to stabilize training and prevent overfitting.

This two-phase transfer learning strategy allowed the model to leverage pretrained representations while adapting selectively to the facial image classification task.

V. Results

CNN Model Performance

The custom convolutional neural network trained from scratch was evaluated using the held-out test set to assess its classification performance on unseen facial images. Training progressed steadily, with training accuracy reaching 87.41%, while test accuracy achieved 83.64%, indicating reasonable generalization with limited overfitting.

A line chart shows the training and validation accuracy and loss curves across epochs. Training accuracy increased consistently, while validation accuracy followed a similar trend with mild fluctuations. Validation loss decreased overall but stabilized earlier than training loss, supporting the use of early stopping to prevent overfitting. (Figure 1.)

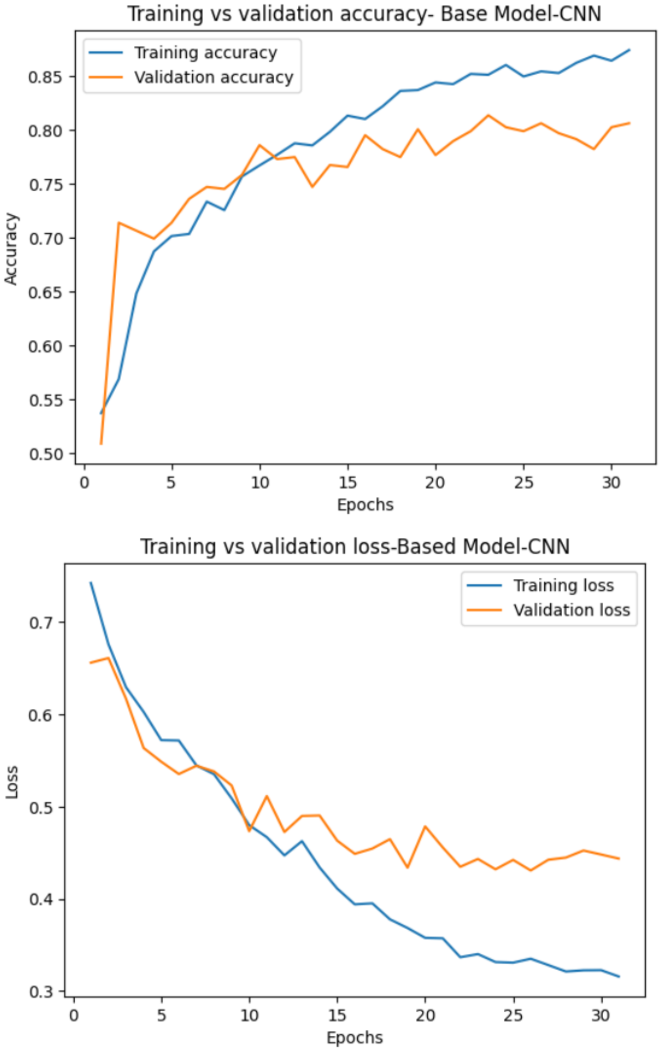
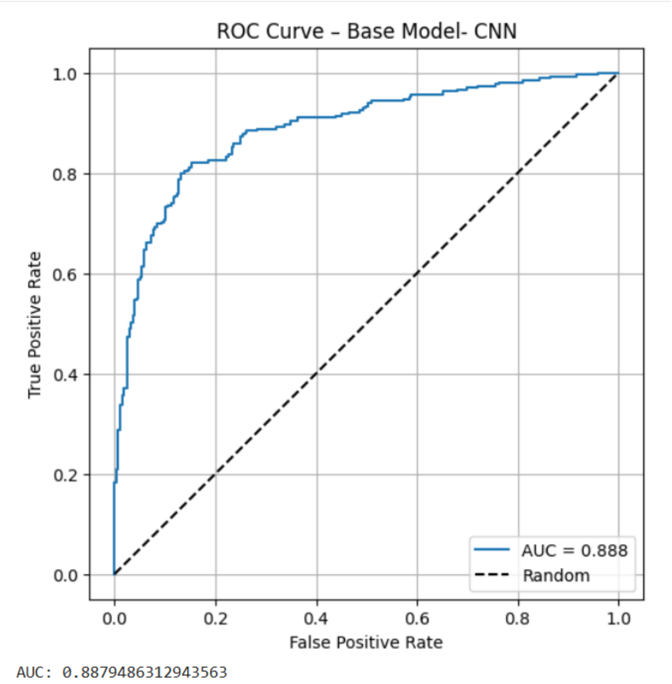


Figure 1. Training vs. validation accuracy & loss curves- scratch CNN model

To further evaluate classification performance, a receiver operating characteristic (ROC) curve was generated. The base model CNN achieved an area under the ROC curve (AUC) of 0.888, indicating good discriminative ability between autistic and non-autistic classes beyond simple accuracy. (Figure 2)



```
AUC: 0.8879486312943563
```

Figure 2. ROC Curve

Classification performance was further evaluated using a confusion matrix and precision–recall metrics. The model correctly classified 231 non-autistic images and 221 autistic images, with 44 false positives and 48 false negatives. While precision and recall were reasonably balanced across classes, recall for autistic images was lower, contributing to an overall weighted F1-score of 0.831 (Figure 3).

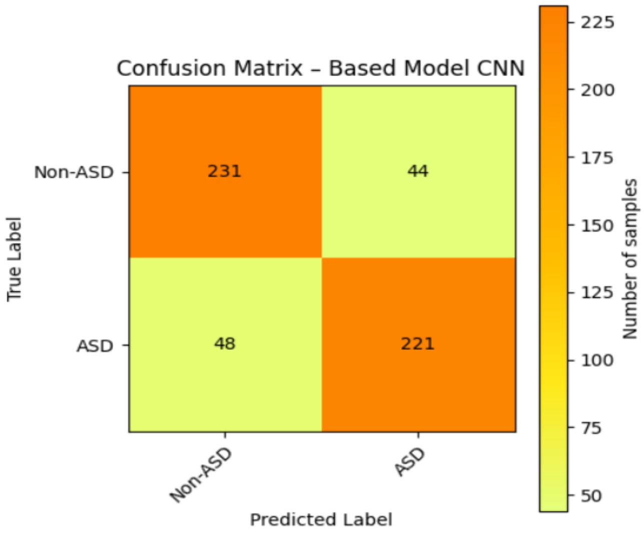


Figure 3. Confusion Matrix for the CNN

These results demonstrate that the base model CNN was able to learn meaningful visual patterns from the dataset, though some misclassifications remained. Performance suggests sensitivity to dataset characteristics such as image variability and background noise, reinforcing the exploratory nature of facial image-based ASD classification.

Transfer Learning Model Results

The MobileNetV2 transfer learning model was evaluated using the same training and test set to enable direct comparison with the scratch CNN. Training was performed in two phases: an initial feature extraction phase with the pretrained base frozen, followed by a fine-tuning phase in which the final 30 layers of the base network were unfrozen. Early stopping, learning rate reduction, and model checkpointing were used to stabilize training and prevent overfitting.

During training, the model achieved high training accuracy after fine-tuning, reaching a peak of 93.17%. However, final test accuracy reached 81.43%, which was comparable to but slightly lower than the base CNN model's test accuracy of 83.64%, suggesting that the pretrained features did not provide a meaningful performance advantage for this task. Figure 4 shows the training and validation accuracy and loss curves across epochs. While training loss continued to decrease during fine-tuning, validation loss exhibited noticeable instability, indicating potential overfitting after unfreezing the pretrained layers.

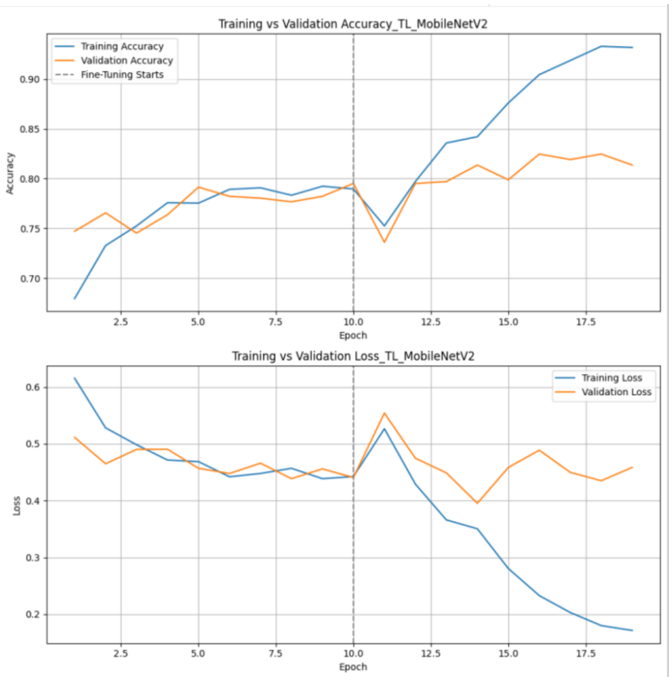
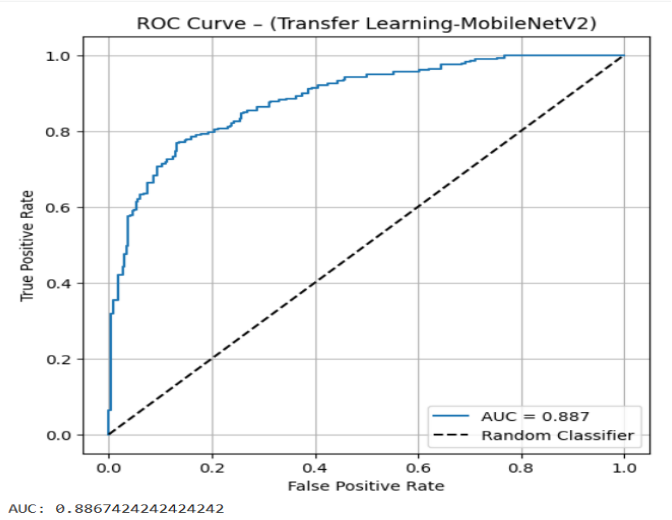


Figure 4. Training vs validation accuracy and loss

To further evaluate classification performance, a receiver operating characteristic (ROC) curve was generated (Figure 5). The MobileNetV2 model achieved an AUC of 0.887, indicating good discriminative ability between autistic and non-autistic classes. This AUC value was comparable to that of the scratch CNN, suggesting that both models learned meaningful visual representations beyond simple accuracy metrics.



```
AUC: 0.8867424242424242
```

Figure 5. ROC

Classification performance was also assessed using confusion matrix, precision, and recall metrics. The confusion matrix shows that the model correctly classified 215 non-autistic and 222 autistic images, with 49 false positives and 58 false negatives. Precision and recall values were relatively balanced across classes, with an overall weighted F1-score of 0.803. (Figure 6)

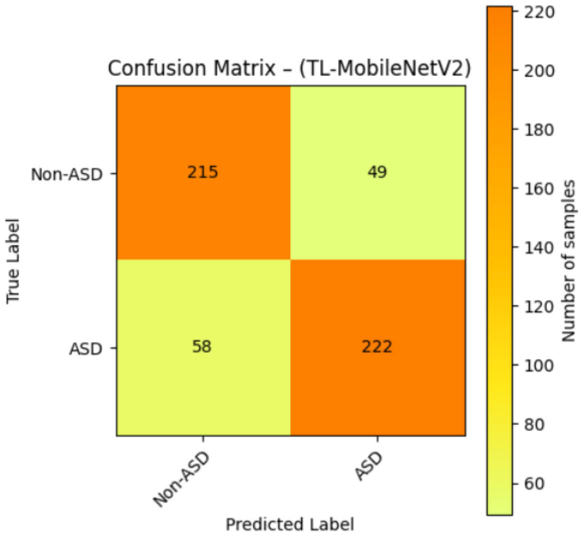


Figure 6. Confusion Matrix for Transfer Learning

Overall, while the pretrained MobileNetV2 model demonstrated strong feature learning capabilities with a higher training accuracy score and achieved competitive AUC performance, it did not outperform the base model CNN in terms of test accuracy or stability. These results suggest that transfer learning did not provide a clear advantage for this dataset and further highlight the sensitivity of facial image classification models to dataset size, image variability, and domain mismatch between pretrained features and task-specific data.

VI. Discussion

This study examined whether deep learning models could distinguish between autistic and non-autistic facial images using a controlled dataset. Two approaches were evaluated: a custom convolutional neural network trained from scratch and a pretrained MobileNetV2 transfer learning model. Although both models were able to learn meaningful visual patterns from the data, their performance differed in ways that highlight how model design and dataset characteristics influence generalization.

The custom CNN demonstrated slightly stronger and more consistent performance overall. Training and validation curves followed similar trends, suggesting that the model learned features that generalized reasonably well rather than memorizing the training data. While misclassifications were still present, as shown in the confusion matrix, the model maintained balanced performance across classes, indicating that it captured useful dataset-specific patterns without severe overfitting.

The MobileNetV2 model learned more quickly during training and achieved high training accuracy, particularly after fine-tuning the final layers. However, this rapid learning was accompanied by less stable validation behavior. The widening gap between training and validation loss suggests that the model became more sensitive during fine-tuning, likely due to differences between the natural images used for pretraining and the smaller, more variable facial image dataset used in this study. As a result, the higher training performance did not translate into improved test accuracy.

Comparing the two models highlights an important takeaway: increased model complexity and pretrained representations do not necessarily lead to better generalization. In this case, the simpler CNN proved more reliable within the constraints of the dataset, while the pretrained model exhibited faster learning but greater instability. These results suggest that alignment between model architecture and dataset characteristics may be more important than architectural complexity alone.

Overall, the findings demonstrate both the potential and the limitations of applying deep learning to facial image-based ASD classification. While the models were able to identify useful visual patterns within a controlled setting, performance remained sensitive to image variability and dataset scale. Continued evaluation using larger and more diverse datasets may help clarify how consistently these approaches perform and their practical applicability in future research.

VII . References

[1] K. K. M. Rahman, M. U. Ahmed, and M. S. Kaiser, “Identification of autism in children using static facial images,” *Sensors*, vol. 22, no. 3, 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8773918/

[2] A. Quatrosi *et al.*, “Cranio-facial characteristics in autism spectrum disorder: A review,” *Brain Sciences*, vol. 14, no. 1, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10856091/

[3] K. Aldridge *et al.*, “Facial phenotypes in subgroups of prepubertal boys with autism spectrum disorders,” *Molecular Autism*, vol. 2, p. 15, 2011. [Online]. Available: https://link.springer.com/article/10.1186/2040-2392-2-15

[4] Advanced Autism Services, “Facial features & physical characteristics of autism,” 2023. [Online]. Available: https://www.advancedautism.com/post/facial-features-physical-characteristics-of-autism